Министерство науки и высшего образования
Российской Федерации

Федеральное государственное учреждение
ФЕДЕРАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЦЕНТР
«ИНФОРМАТИКА И УПРАВЛЕНИЕ»
РОССИЙСКОЙ АКАДЕМИИ НАУК
(ФИЦ ИУ РАН)

На правах рукописи

ГОРШЕНИН Андрей Константинович

ПОЛУПАРАМЕТРИЧЕСКИЕ МЕТОДЫ АНАЛИЗА НЕОДНОРОДНЫХ ДАННЫХ И ИХ ПРИМЕНЕНИЕ В ЗАДАЧАХ МАТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ

Специальность 05.13.18 — математическое моделирование,
численные методы и комплексы программ

ДИССЕРТАЦИЯ
на соискание ученой степени
доктора физико–математических наук

Научный консультант:
д. ф.-м. н., профессор
В. Ю. Королев

Москва
2020

Содержание

Введение

Актуальность

Получение новых результатов во многих современных научных областях неразрывно связано со всесторонним анализом огромных накопленных неоднородных массивов данных с привлечением самых современных инфраструктурных ресурсов и задействованием передовых вычислительных средств – высокопроизводительных кластеров и дата-центров – в рамках комплексных междисциплинарных исследований. Поэтому чрезвычайно важным становится развитие соответствующих методов, которые в последние годы рассматривают в рамках отдельной дисциплины – науки о данных (`data science`) [167, 188, 268, 283, 337, 341, 375, 399]. Данная исследовательская область находится на стыке математического моделирования, математической статистики, машинного обучения, интеллектуального анализа и вычислительно-интенсивных алгоритмов, которые позволяют эффективно обрабатывать даже неструктурированные данные больших объемов [170, 209, 361].

Создание методов и алгоритмов анализа данных для эффективного использования в прикладных задачах с задействованием современных высокопроизводительных вычислительных ресурсов зачастую невозможно без развития математических моделей, описывающих функционирование сложных систем и эволюцию различных процессов в них. В рамках математического моделирования можно выявлять новые знания об объекте на основе используемой модели либо осуществить выбор модели (оценивание неизвестных параметров) на основании известных данных. Первую задачу принято называть прямой, и ее решение ориентировано на выявление или прогнозирование, например, экстремальных характеристик описываемого объекта или явления. Вторая задача является обратной, и ее решение позволяет выбрать модель из некоторой совокупности (семейства), например, с помощью аппарата теории вероятностей и математической статистики. Необходимо отметить высокую актуальность статистических методов в том числе и при решении за-

дач, связанных с данными больших объемов. В частности, они могут применяться при анализе неоднородных наблюдений, разработке аналитических процедур выбора моделей высокой размерности и оценивания их параметров, проверки сложных гипотез [209].

Процесс накопления данных зачастую протекает в условиях неопределенности, обусловленной:

– стохастическим характером интенсивностей потоков информативных событий и взаимодействием большого числа не поддающихся исчерпывающему прогнозированию факторов, которые можно считать случайными;

– неоднородностью или нестационарностью изучаемых закономерностей;

– неполнотой получаемой информации, в частности, из-за стохастического характера поведения внешней среды.

Указанные обстоятельства ведут к необходимости изучения вероятностно-статистических характеристик данных, прежде всего, с использованием смешанных вероятностных моделей наблюдаемых процессов и явлений, что делает вполне естественным применение байесовских статистических методов анализа данных [213]. При этом параметры смешивающего (в байесовской терминологии – априорного) распределения определяются в результате анализа данных о поведении внешних факторов (окружающей среды).

Методы исследования, развиваемые в диссертации, опираются на вероятностно-статистические подходы к описанию объектов и явлений. Основой для построения моделей являются выборки случайного объема и результаты в области предельных теорем для сумм и максимумов случайных величин, а также различных возникающих при этом смешанных распределений. Значительный вклад в развитие указанных направлений теории вероятностей и математической статистики внесли российские математики, среди которых следует упомянуть А. Н. Колмогорова [83], Б. В. Гнеденко [6], И. А. Ибрагимова, Ю. В. Линника [82], Ю. В. Прохорова [107], А. Н. Ширяева [119–122], В. М. Золотарева [447], В. В. Калашникова [279], В. В. Петрова [105], В. М. Круглова [98, 99], В. Ю. Королева [88,99]. В указанных теоремах со случайным объемом выборки в качестве предельных законов для распределений сумм и максимумов или для неоднородных и нестационарных случайных блужданий выступают смеси распределений, предельные в случае выборок неслучайного объема, в том числе сдвиг-масштабные нормальные смеси. При этом удобными

аппроксимациями для них как с аналитической, так и с вычислительной точек зрения являются конечные смеси [88, 89, 334]. Известны многочисленные применения смешанных вероятностных моделей в различных прикладных задачах, например, для описания процессов в турбулентной плазме [384], при анализе финансовых данных [123, 420], в процессе обработки изображений в медицине [173, 405], в ряде социологических исследований [435].

Для оценивания параметров смешанных распределений требуется развитие нетривиальных статистических методов. В рамках разрабатываемых в диссертации подходов на основе предельных теорем теоретически обоснован выбор нормального распределения в качестве смешиваемого. Однако его параметры являются случайными, распределение которых выступает в качестве смешивающего. При этом часто аналитический вид смешивающего распределения неизвестен, поэтому его оценка представляет собой непараметрическую задачу. Таким образом, необходимо развитие методов полупараметрического статистического оценивания [155, 179, 187, 265, 421, 436] для построения смешанных вероятностных моделей объектов и явлений.

Одним из наиболее эффективных методов параметрического оценивания смешанных моделей является итерационная процедура, называемая EM-алгоритмом, которая под таким наименованием была детально описана и исследована А. Демпстером, Н. Лейрдом и Д. Рубиным [192] в 1977 году. Данный метод получения оценок максимального правдоподобия применялся еще в 1958 году Х. Хартли [266] при работе с неполными данными, но и по настоящий момент с учетом многочисленных модификаций остается одним из важных инструментов статистического, в том числе байесовского, и интеллектуального анализа данных [88, 423].

Различные разновидности базового метода разрабатывались в разное время исследователями по всему миру с целью преодоления известных недостатков классического EM-алгоритма. Построенные на его основе процедуры используются в задачах кластеризации [171, 278, 318, 429], регрессии [401, 409], обработки цензурированных и усеченных данных [312], оценивания параметров различных распределений и процессов [313, 424, 432], в том числе с организацией параллельных вычислительных алгоритмов и обучением нейронных сетей [204, 342]. Однако в процессе модификации обычно сохраняется общий принцип наличия E- (от `expectation`) и M-шагов (от `maximization`). Например, в стохастическом (`SEM`) варианте алгоритма [88, 156, 164, 172, 191, 343] вводится допол-

нительный S-этап (от `stochastic`). Он предназначен, в частности, для противодействия свойству жадности классического алгоритма – а именно, выбору методом в качестве оценки локального максимума, который расположен наиболее близко к начальному приближению, но может не являться глобальным. Именно данная модификация использована для оценивания параметров в слоях глубокой смешанной гауссовской модели, предложенной в статье [417] Дж. МакЛахлана, одного из ведущих мировых специалистов по конечным смесям и задачам классификации. Можно также отметить, что классический метод обучения нейронных сетей на основе обратного распространения ошибки является специальным случаем обобщенного EM-алгоритма [136].

Ряд модификаций направлен на повышения скорости сходимости. Так, в статье [346] предложено введение дополнительного «зашумляющего» этапа, улучающего эффективность метода примерно на 10–15%. Идея введения подобной модификации основана на явлении стохастического резонанса [210, 301], которое хорошо известно в области статистической обработки сигналов. Однако определение параметров зашумляющих данных основывается на специальных множествах и теоремах для условных математических ожиданий, которые весьма трудно использовать на практике – прежде всего, с точки зрения автоматизации и программной реализации этапа зашумления. Однако сам подход может рассматриваться в качестве перспективного для повышения эффективности методов анализа данных.

EM-алгоритм может быть использован для исследования и оценивания нестационарности наблюдаемого процесса и выделения тренда, а также его декомпозиции на локальные составляющие. Для решения указанных задач используется метод скользящего разделения смесей (`CPC`) [88]. Он основан на смешанных вероятностных моделях конечномерных распределений наблюдаемого процесса и представляет собой обобщение метода дисперсионного анализа (в рамках модели со случайными факторами) на временные ряды. С помощью CPC-метода возможно осуществить естественную декомпозицию волатильности (изменчивости) анализируемого процесса на диффузионную (случайную) и динамическую (трендовую) компоненты. Таким образом, возникает естественное разложение суммарного тренда процесса на локальные компоненты, наличие которых обусловлено разными факторами. Кроме того, возможно отследить эволюцию данных факторов во времени. Для этого процедуры типа EM-алгоритма используются в режиме скользящего окна для

оценивания неизвестных параметров конечномерных распределений наблюдаемого процесса. С помощью СРС-метода впервые удалось определить число характерных процессов (в среднем от 3 до 5), формирующих ионно-звуковую турбулентность в плазме [1]. Также получены значимые результаты в области анализа волатильности финансовых индексов [239].

Возможны и другие подходы к построению стохастических моделей различных явлений, в том числе на основе стохастических дифференциальных уравнений (СДУ) [305, 370, 371, 379–381]. Методы статистического анализа нормальных смесей могут быть использованы для исследования процессов, задаваемых СДУ вида $dX(\omega,\, t,\,) = a(\omega,\, t)dt + b(\omega,\, t)dW(\omega,\, t)$, которые в физике традиционно называются уравнениями Ланжевена. В них коэффициенты $a(\omega,\, t)$ и $b(\omega,\, t)$ являются случайными функциями, а $W(\omega,\, t)$ представляет собой винеровский процесс. Данные СДУ и различные их обобщения успешно используются в задачах моделирования финансовых рынков [121, 157], ассимиляции данных при анализе разномасштабной изменчивости геофизических переменных [149], взаимодействия частиц в плазме [126, 362, 376]. Кроме того, эти уравнения позволяют расширить традиционный подход на основе многомерных уравнений Фоккера-Планка и самосогласованно моделировать взаимодействие частиц в плазме в стохастических электромагнитных полях [198]. Важной в данных условиях становится задача статистического оценивания коэффициентов в подобных СДУ.

Из вида уравнения Ланжевена следует, что в каждый момент времени распределение приращений случайного процесса, удовлетворяющего этому уравнению, является смесью нормальных законов, что ведет к необходимости развития методов их исследования и оценивания параметров. При этом необходимо учитывать, что статистические закономерности поведения рассматриваемых процессов $X(\omega,\, t)$, $a(\omega,\, t)$, $b(\omega,\, t)$ изменяются во времени нерегулярным образом, результатом чего является отсутствие универсального смешивающего закона. Однако информация об их эволюции может быть использована для нетривиального (за счет характеристик математической модели, а не функционального преобразования исходных наблюдений) расширения признакового пространства [174] для повышения эффективности алгоритмов интеллектуального анализа. Указанная задача оценивания распределений параметров рассмотрена в диссертации с точки зрения разработки соответствующих полупараметрических статистических методов.

С развитием вычислительных мощностей методы машинного обу-

чения и нейронные сети, особенно глубокие, стали одним из наиболее востребованных и эффективных инструментов всестороннего анализа данных [276]. Существенный вклад в их развитие внесли, в частности, В. Н. Вапник и А. Я. Червоненкис [3, 4, 412], Я. Лекун, И. Бенджио и Дж. Хинтон [310, 374]. Подобные процедуры успешно применяются для обработки наблюдений в самом широком спектре областей, включая метеорологию [129, 200], финансы [175, 425], медицину [367, 410] и многие другие. При этом получение прорывных результатов обеспечивается не только выбором подходящих типов архитектур и настройкой гиперпараметров [152], то есть величин, которые не изменяются в процессе обучения: методов оптимизации, количества скрытых слоев и нейронов в них и др. Весьма эффективным является комплексный подход на основе развития сложных математических моделей, применения ансамблей гибридных инструментов обработки данных [161, 185] и различных способов нетривиального расширения признакового пространства, не требующих увеличения объема тренировочных данных, но существенным образом повышающих качество обучения.

Реализация подобных высоко-интенсивных алгоритмов для решения научных задач требует значительных высокопроизводительных вычислительных ресурсов [273, 311, 426]. В частности, достигнуты существенные успехи за счет задействования для проведения расчетов, помимо центрального процессора, графических карт (GPU), прежде всего на основе программно-аппаратной архитектуры NVIDIA CUDA [176, 212]. Применение гетерогенных вычислений [163, 349] на основе подхода GPGPU (от англ. General-Purpose Computing for Graphics Processing Units) для быстрой параллельной обработки данных в научных исследования весьма привлекательно в силу их относительно низкой стоимости, сочетающейся со значительной производительностью, возможностью реализации достаточно точных численных методов, а также с повышением эффективности обучения нейронных сетей. Указанные подходы используются в широком спектре исследовательских областей, в частности могут быть упомянуты:

– гидрологическое моделирование [166], геопространственный анализ и исследование наблюдений за Землей [178](облачные вычисления);

– медицинская диагностика на основе эластографии в режиме реального времени [428], картирование распространения лавы при потенциальном вулканическом извержении [133], гидродинамическое моделирование наводнений в горных водоразделах [269] (технология GPGPU);

– симуляция процесса длинноволнового излучения [323], квантово-химические вычисления [327] (гибридное решение на базе центральных и графических процессоров).

Цель и задачи диссертационной работы

Основной целью диссертации является разработка комплекса новых методов анализа неоднородных данных на основе развития универсальных смешанных вероятностных моделей с аналитическим исследованием их свойств, созданием эффективных вычислительных алгоритмов оценивания и прогнозирования характеристик таких моделей. Для достижения указанной цели в диссертации решены следующие задачи:

– определение вида смешанных законов, являющихся предельными в схемах взятия максимума и суммирования для выборок случайного объема, и аналитическое исследование свойств смешанных распределений;

– создание, развитие и исследование свойств полупараметрических методов анализа неоднородных данных и построение на их основе универсальных вероятностных моделей;

– разработка программных комплексов, реализующих методы оценивания параметров предложенных математических моделей и их прогнозирования с помощью с использованием алгоритмов машинного обучения и нейронных сетей, и их тестирование в высокопроизводительных вычислительных средах;

– применение разработанных подходов для построения математических моделей в различных прикладных областях.

Методы исследования

В работе использованы оригинальные подходы и процедуры, предложенные и развиваемые в диссертации, в том числе:

– полупараметрические методы статистического моделирования, включая CPC-метод, процедуру статистического оценивания распределений случайных параметров стохастических дифференциальных уравнений Ланжевена, а также алгоритм определения связности компонент для выявления числа структурных процессов в данных;

– метод расширения признакового пространства для повышения точности обучения нейронных сетей за счет использования параметров смешанных вероятностных моделей;

– вариации бутстреп-процедур для имитационного моделирования;

– модифицированный метод превышения порогового значения.

Также в работе применяются и классические методы исследования, в том числе:

– современные аналитические методы теории вероятностей и математической статистики для смешанных распределений и выборок случайного объема;

– методы параметрического и непараметрического статистического оценивания;

– аппарат проверки статистических гипотез;

– методы функционального анализа, линейной алгебры и оптимизации;

– методы вычислительной статистики, алгоритмы машинного обучения и нейронные сети.

Для создания комплекса программных решений, предназначенных для автоматизации моделирования, проведения анализа данных и возможности обработки значительных объемов массивов наблюдений, использованы языки программирования `MATLAB` и `Python`, а также современные высокопроизводительные вычислительные ресурсы.

Научная новизна и основные результаты

В диссертации разработаныэффективные полупараметрические подходы к построению математических моделей процессов и явлений на основе анализа динамически формируемых массивов неоднородных данных, объединяющие в себе:

– строгие теоретические обоснования вида используемых в универсальных вероятностных моделях смешиваемых и смешивающих распределений на базе предельных теорем теории вероятностей;

– развитие методологии статистического (байесовского) оценивания этих семейств с использованием дискретных аппроксимаций смешивающих распределений и метода скользящего разделения смесей;

– возможность естественного использования параметров получаемых вероятностных моделей для нетривиального расширения признакового пространства в методах машинного обучения и нейронных сетях с целью повышения точности их работы;

– развитие методов исследования тонкой стохастической структуры процессов и явлений в различных прикладных областях с помощью разложения волатильности (изменчивости) на трендовые и диффузионные компоненты.

Таким образом, основные результаты диссертации являются новыми и состоят в следующем:

1. Развит подход к математическому моделированию процессов и яв-

лений на основе нового варианта центральной предельной теоремы для сумм со случайным числом независимых и необязательно одинаково распределенных слагаемых, в которой в качестве предельных распределений выступают нормальные смеси произвольного вида.

2. Развит подход к математическому моделированию процессов и явлений на основе схемы максимума для выборок, объем которых описывается важным для прикладных задач семейством обобщенных отрицательных биномиальных распределений: получен вид предельного закона и аналитически исследованы некоторые его свойства.

3. Развит подход к математическому моделированию редких событий на основе обобщения классической теоремы Реньи: установлен вид предельного распределения случайных сумм с обобщенным отрицательным биномиальным распределением в законе больших чисел без предположений о независимости и одинаковой распределенности слагаемых.

4. Аналитически показано наличие устойчивости дисперсионно-сдвиговых и конечных сдвиговых смесей нормальных распределений относительно возмущений параметров смешивающего распределения в терминах расстояния Леви, которая обосновывает корректность вычислительных процедур разделений смесей этих семейств распределений.

5. Развиты полупараметрические методы анализа неоднородных данных и аналитически исследованы некоторые их свойства в моделях аддитивного зашумления конечными смесями и округления наблюдений.

6. Разработан полупараметрический подход к статистическому оцениванию распределений случайных коэффициентов стохастического дифференциального уравнения Ланжевена.

7. Развита статистическая методология построения моделей сгруппированных неизвестных наблюдений при заданных характерных точках их эмпирической функции распределения.

8. Разработаны методы и алгоритмы статистической идентификации и классификации экстремальных наблюдений в массивах неоднородных данных на основе обобщенных отрицательных биномиальных распределений числа наблюдений и обобщенных гамма-моделей для данных.

9. Созданы комплексы программных решений для автоматизации обработки данных значительных объемов с использованием высокопроизводительных вычислительных ресурсов, реализующие разработанные полупараметрические методы, и продемонстрировано их применение к решению некоторых задач математического моделирования в физике плазмы, метеорологии, океанологии, селенологии.

Теоретическая и практическая значимость

Результаты диссертации являются одновременно фундаментальными и прикладными, а проведенные исследования – комплексными и имеющими ярко выраженный междисциплинарный характер. Разработанные методы анализа данных и вычислительные процедуры основываются на развитых в диссертации математических результатах, включая предельные теоремы теории вероятностей и математической статистики. При этом они ориентированы на эффективное применение в различных прикладных областях, что продемонстрировано в диссертации на примерах анализа данных в различных предметных областях.

Апробация работы

Результаты работы представлялись на международных и российских научных конференциях и семинарах по тематике исследований:

– International Seminar on Stability Problems for Stochastic Models (ISSPSM) and Workshop «Applied Problems in Theory of Probabilities and Mathematical Statistics related to modeling of information systems»: 2012–2014, 2018, 2020 гг. [221, 227, 249, 291, 328];

– European Conference on Modelling and Simulation (ECMS): 2013–2015, 2017 гг. [229, 234, 237, 240];

– International Conference of Numerical Analysis and Applied Mathematics (ICNAAM): 2013–2016 гг. [222, 228, 235, 238, 243, 295];

– International Conference on Modern Techniques of Plasma Diagnostics and their Application: 2014 г. [101, 383];

– International Congress on Ultra Modern Telecommunications and Control Systems (ICUMT): 2015, 2018 гг. [205, 225, 242];

– International Scientific Conference on Information Technologies and Mathematical Modelling (ITMM): 2015, 2016 гг. [230, 294];

– International Conference on Distributed Computer and Communication Networks: Control, Computation, Communications (DCCN): 2016, 2018, 2019 гг. [231, 232, 248];

– International Conference of Artificial Intelligence, Medical Engineering, Education (AIMEE): 2018, 2020 гг. [413];

– International Symposium «Intelligent Systems» (INTELS): 2018 г. [433];

– International Symposium on Computer Science, Digital Economy and Intelligent Systems (CSDEIS): 2019, 2020 гг. [247];

– Международная Звенигородская конференция по физике плазмы и управляемому термоядерному синтезу: 2013, 2015 гг. [102, 109];

– Международная научно-методическая конференция «Информатизация инженерного образования» (ИНФОРИНО): 2014, 2016 гг. [15, 24];

– Всероссийская конференция (с международным участием) «Информационно-телекоммуникационные технологии и математическое моделирование высокотехнологичных систем»: 2016, 2018 гг. [57, 223];

– Всероссийская научная конференция «Ломоносовские чтения»: 2018–2020 гг. [60];

– Всероссийский Симпозиум по прикладной и промышленной математике: 2014, 2015, 2019 гг. [45, 54, 55];

– Всероссийская научно-практическая конференция с международным участием «Актуальные проблемы глобальных исследований: Россия в глобализирующемся мире»: 2019 г. [53];

– научная конференция «Тихоновские чтения»: 2015 г. [94];

– научный семинар кафедры математической статистики факультета ВМК МГУ имени М. В. Ломоносова «Теория риска и смежные вопросы»: 2012–2020 гг.

Результаты диссертации использованы в Институте общей физики им. А. М. Прохорова Российской академии наук при вероятностно-статистическом моделировании процессов в экспериментах с турбулентной плазмой в стеллараторе Л-2М, в Институте океанологии им. П. П. Ширшова Российской академии наук при анализе статистических закономерностей в метеорологических и океанологических данных, а также апробированы в рамках отдельных тем учебного курса «Прикладной многомерный статистический анализ» Центра компетенций Национальной технологической инициативы по технологиям хранения и анализа больших данных на базе Московского государственного университета имени М. В. Ломоносова.

Публикации

Материалы диссертации опубликованы в **82** печатных работах [11, 14–16, 23, 24, 29, 30, 34, 35, 45, 51–55, 57, 59–61, 64, 66–68, 71–73, 75, 77, 80, 81, 92–96, 101, 102, 109, 147, 205, 221–238, 240–250, 291–297, 328, 329, 383, 413, 433], из них:

– **31** статья в журналах, включенных в перечень ВАК [11, 14, 16, 23, 29, 30, 34, 35, 51, 52, 59, 61, 64, 66–68, 71–73, 75, 77, 81, 92, 93, 95, 224, 226, 244–246, 250];

– **51** статья в изданиях, индексируемых базами Web of Science Core Collection и/или Scopus [23, 29, 30, 34, 35, 59, 61, 66, 72, 73, 75, 77, 93, 147, 205, 222–226, 228–238, 240–248, 250, 292–297, 329, 383, 413, 433], включая журналы

первого и второго квартилей [93, 147, 241, 248, 292, 293].

Получены **35** свидетельств о государственной регистрации программ для ЭВМ [12, 13, 17–22, 25–28, 31–33, 36–44, 46–50, 56, 58, 69, 70, 74, 76], зарегистрированных в Федеральной службе по интеллектуальной собственности (Роспатент).

Личный вклад автора

Основные результаты диссертации получены лично автором. В работах [51–55, 67, 68, 71–73, 75, 77, 80, 81, 228, 229, 237, 242–250, 413, 433] А. К. Горшениным выполнены постановка исследовательских задач, определение ключевых концепций и методов решения, а также проведен всесторонний анализ полученных результатов. В работах [57, 59–61, 64, 66, 92–96, 101, 102, 109, 147, 205, 227, 230–236, 238, 240, 241, 291–297, 328, 329, 383] А. К. Горшениным развиты математические модели, методы и вычислительные алгоритмы анализа реальных данных с реализацией в виде программных решений и их приложениями к обработке наблюдений из прикладных областей. В программах [56, 58, 69, 70, 74, 76] А. К. Горшениным реализованы алгоритмы анализа данных в виде значимых компонентов зарегистрированных инструментов.

Содержание работы

Во **Введении** обоснована актуальность темы диссертации, сформулированы цели, задачи, методы исследования и основные полученные результаты.

В главе 1 доказаны предельные теоремы для схемы максимизации и суммирования элементов выборок, объем которых является случайной величиной с обобщенным отрицательным биномиальным распределением. Первая из схем ориентирована на поиск асимптотического распределения при неограниченном росте объема выборки максимального элемента, а у второго – суммы всех наблюдений. Указанные схемы являются корректными и удобными вероятностно-статистическими моделями в рамках анализа реальных различных типов данных.

В §1.1 вводится понятие смешанного распределения вероятностей и описываются его базовые свойства. В §1.2 описано обобщение отрицательного биномиального распределения как смешанного пуассоновского со смешивающим обобщенным гамма-распределением. Получены рекуррентные представления для данного распределения и формулы для математического ожидания и дисперсии (утверждения 1.1 и 1.2). Результаты опубликованы в статье [232].

В §1.3 доказана теорема об асимптотическом распределении максимальной порядковой статистики в выборке, объем которой является обобщенной отрицательной биномиальной случайной величиной (теорема 1.1). Получены эквивалентные представления данного распределения в виде смесей известных распределений: Фреше, Снедекора-Фишера, строго устойчивого и др. (см. теоремы 1.1 и 1.3) и выражение для моментов произвольных порядков (теорема 1.5). Установлено, что при некоторых ограничениях на параметр данное распределение является безгранично делимым (теорема 1.4 и следствие 1.1). В важном частном случае, когда элементы выборки имеют распределение Парето, получена оценка скорости сходимости к предельному распределению (теорема 1.6). При условии, что объем выборки является классической случайной биномиальной величиной, получен простой аналитический вид предельного распределения (теорема 1.2), а также выписана оценка скорости сходимости для элементов выборки с распределение Парето (замечание 1.5).

Для отрицательных биномиальных объемов выборок получено явное аналитическое представление асимптотических распределений порядковых статистик (теорема 1.7) и выборочных квантилей: в этом случае оно является распределением Стьюдента (см. теорему 1.8). Доказан закон больших чисел для сумм с обобщенным отрицательным биномиальным распределением – обобщение теоремы Реньи, – в котором для слагаемых не предполагается независимость и одинаковая распределенность (теорема 1.9). Указанные результаты опубликованы в статьях [93, 233, 292, 293]. Далее они используются в главе 6 для построения вероятностно-статистических моделей реальных метеорологических и океанологических процессов и определения экстремальных наблюдений в них.

В §1.4 доказан новый вариант центральной предельной теоремы (теорема 1.10) для сумм со случайным числом независимых и необязательно одинаково распределенных слагаемых в схеме серий, в которой в качестве предельных распределений возникают произвольные нормальные смеси. Данный результат используется в главе 4 для обоснования вида моделей астрономических данных (размеров частиц лунного реголита). Результаты опубликованы в статье [241].

Глава 2 посвящена исследованию аналитических свойств моделей на основе конечных нормальных и гамма-распределений. В §2.1 описан метод скользящего разделения смесей и предложено его использование в качестве базовой процедуры статистического оценивания распределений

случайных коэффициентов стохастического дифференциального уравнения Ланжевена. Данный подход позволяет содержательно расширять признаковое пространство в методах машинного обучения за счет использования характеристик адекватных математических моделей. В §5.2 будет продемонстрировано, как использование подобных величин позволяет повысить эффективность прогнозирования нестационарных данных с помощью нейронных сетей. Результаты опубликованы в статье [66].

В §2.1 приведены сведения о важных модификациях EM-алгоритма – медианных, которые ведут к робастным оценкам, а также стохастических, позволяющих эффективнее выбирать в качестве решений глобальные, а не локальные максимумы, а также сформулирована теорема о свойствах стохастического EM-алгоритма. Продемонстрирован вывод формул для итерационных шагов метода скользящего разделения конечных гамма-смесей (утверждение 2.2), а также пример их применения для анализа данных биржевой книги заявок. Результаты опубликованы в статьях [66, 227, 229]. Получено свидетельство о государственной регистрации программы для ЭВМ [38].

В §2.2 и §2.3 рассмотрены две важные модели возмущений параметров смеси – добавления и расщепления компоненты – и приведены результаты относительно асимптотически оптимальных критериев проверки гипотез о числе компонент смеси (теоремы 2.2 и 2.3) и устойчивости конечных масштабных смесей нормальных законов относительно смешивающего распределения в них (теоремы 2.4–2.7) В §2.4 и §2.5 они развиваются для задач устойчивости конечных сдвиговых и дисперсионно-сдвиговых смесей нормальных законов относительно изменений параметров смешивающего распределения. В §2.4 получены оценки устойчивости конечных сдвиговых смесей нормальных законов по отношению к изменениям смешивающего параметра (теоремы 2.8–2.11). Для каждой из моделей выписаны в явном виде двусторонние оценки, связывающие расстояния Леви между смесями и смешивающими законами. Они обосновывают корректность аппроксимации произвольных сдвиговых нормальных смесей, которые в общем случае не являются идентифицируемыми, конечными аналогами в задаче их статистического разделения. Результаты опубликованы в статье [11]. В §2.5 доказана теорема 2.12 об устойчивости дисперсионно-сдвиговых смесей нормальных законов. Показано, что близость смешивающих распределений в смысле расстояния Леви необходимо влечет и близость соответствующих смесей. Полученные соотношения могут быть использованы для обоснования вычисли-

тельных процедур разделения дисперсионно-сдвиговых смесей нормальных законов. Результаты опубликованы в статье [96].

В §2.6 разработаны теоретические подходы к устранению ошибок в смешанной модели округления данных. Получены оценки для математического ожидания наблюдений в предположении зашумления конечными смесями нормальных (теорема 2.13) и гамма-распределений (теорема 2.15). Построены доверительные интервалы для неизвестного математического ожидания в этих случаях с использованием уточненной оценки (2.60) (теоремы 2.14 и 2.16). Соответствующие соотношения зависят только от «экстремальных» значений параметров смесей, но не от числа компонент и весов в распределении зашумляющих наблюдений. Данный подход позволяет учесть большее число случайных факторов, влияющих на величину дополнительной ошибки, связанной с особенностями практической регистрации наблюдений. Результаты опубликованы в статье [34].

В главе 3 разработаны алгоритмы анализа данных, в основу которых положен метод скользящего разделения смесей. В §3.1 получены явные линейные и матричные выражения для моментных характеристик конечных нормальных смесей в CPC-методе (теоремы 3.1 и 3.2). Они существенным образом используются при анализе процессов в физике турбулентной плазмы в §5.2 и §5.3, а также в океанологии в §6.5. Формулировки и доказательства теорем 3.1 и 3.2 опубликованы в статьях [23, 246]. Для метода вычисления моментных характеристик и их визуализации получено свидетельство о государственной регистрации программы для ЭВМ [22].

В §3.2 предложен адаптивный алгоритм выделения полезного сигнала на фоне шума в смешанных нормальных моделях, получен аналитический вид оценок параметров в линейной и матричной формах (см. теоремы 3.3 и 3.4). На примере рассмотрения 24 тестовых выборок с различными комбинациями сигнала и шума продемонстрировано, что предложенный адаптивный алгоритм позволяет эффективно решать задачу определения параметров полезного сигнала. Для использованных тестовых выборок ошибка `RMSE` в большинстве случаев не превышает 1 вне зависимости от соотношений между параметрами сигнала и шума, при этом нормализация данных не производилась. Полученные результаты могут быть полезны в задачах обработки различных экспериментальных данных, например, в физике турбулентной плазмы. Результаты опубликованы в статьях [45, 226, 250].

В §3.3 разработан алгоритм последовательной идентификации (определения локальной связности) компонент смесей вероятностных распределений. В его основу положена комбинация жадного алгоритма поиска числа компонент и одного из методов кластеризации, например, k-средних. Предложенная процедура может быть естественным образом расширена на случай многомерных смешанных распределений. Она будет использована для статистического определения числа формирующих процессов в турбулентной плазме в разделе 5.2.2, а также для статистического оценивания распределений случайных коэффициентов стохастических дифференциальных уравнений типа Ланжевена для потоков тепла между океаном и атмосферой в разделе 6.5.3. Результаты опубликованы в статье [66].

В §3.4 предложен двухэтапный метод детектирования событий в потоке данных на основе анализа динамической компоненты дисперсии (волатильности) изучаемого процесса. На примере прикладной задачи неинвазивного определения областей активности в головном мозге продемонстрирована высокая эффективность данного метода. Соответствующие результаты опубликованы в статьях [236, 240], также получено свидетельство о государственной регистрации программы для ЭВМ [56].

В §3.5 предложен метод повышения точности CPC-аппроксимации с помощью конечных нормальных смесей на основе дополнительного зашумления наблюдений для повышения качества структурного анализа неизвестных процессов в реальных информационных системах. Для этого использовано искусственное зашумление исходных данных с помощью введения дополнительной компоненты, имеющей нормальное распределение с заданными параметрами. Метод позволяет проанализировать закономерности изменения параметров и выявлять краткосрочную изменчивость стохастического процесса в случае сложной внутренней структуры данных. Для модельных примеров из нескольких предметных областей (метеорология, разработка программного обеспечения) продемонстрировано улучшение возможности интерпретации результатов CPC-анализа. Результаты опубликованы в статьях [34, 54, 228, 231, 244].

В главе 4 рассмотрена задача моделирования распределений размеров пылевых частиц лунного реголита, возникающих в результате различных воздействий, при которых развиваются как взрывные процессы разлета частиц с их дроблением, так и спекание в экзотермических плазмохимических реакциях синтеза. В данной главе существенно используются теоретические результаты §1.4 для обоснования корректно-

сти использования логнормальных моделей (см. §4.1) в разработанных статистических процедурах (на основе бутстреп-подхода в §4.2 и минимизации статистики χ^2 в §4.3) обработки всех 317 проб лунного реголита, представленных в каталоге NASA, доставленных миссиями «Аполлон-11, 12, 14–17» и «Луна 24». Продемонстрировано высокое согласие аппроксимационных логнормальных смешанных моделей с данными просеивания реальных образцов лунного реголита. В §4.4 показано, что кластерный анализ параметров, предложенных моделей может оказаться перспективным инструментом выявления структуры подобных реальных данных с учетом физико-химической интерпретации результатов.

Подобные методы могут быть успешно использованы и при решении задач из других предметных областей, в которых неизвестные наблюдения сгруппированы, но для них заданы лишь некоторые характерные точки эмпирической функции распределения. Результаты данного раздела опубликованы в статьях [61, 241], получены свидетельства о государственной регистрации программ для ЭВМ [48, 49].

В главе 5 описываются разработка и применение различных методов интеллектуального анализа данных на основе конечных смесей вероятностных распределений и их скользящего разделения в комбинации с нейросетевыми подходами для моделирования и изучения тонкой структуры процессов, наблюдаемых в экспериментах с турбулентной плазмой.

В §5.1 исследован подход к анализу данных плазменной турбулентности на основе аппроксимации спектров с помощью конечных сдвиг-масштабных смесей вероятностных распределений. Для нескольких серий спектров, полученных для разных режимов низкочастотной плазменной турбулентности, продемонстрирована эффективность использования предложенного метода. С его помощью удалось решить важные для прикладной области задачи, а именно: осуществить идентификацию амплитудного спектра с определением формы гармоник в нем и разделением на компоненты, выявить повторяемость стохастических процессов с характерными средними частотами полуширины спектра, а также определить величину таких физических показателей функционирования плазмы, как величина радиального электрического поля и фазовые скорости флуктуаций. Предложенный подход ориентирован на выявление новых закономерностей в физике турбулентной плазмы с использованием информационных технологий (ИТ). Результаты опубликованы в статьях [14, 55, 64, 101, 102, 109, 238, 328, 329, 383]. Получены свидетельства о государственной регистрации программы для ЭВМ [12, 13, 17–19, 21, 38].

В §5.2 развивается вероятностно-статистический подход к анализу эволюции характеристик микротурбулентности в переходном процессе электронно-циклотронного резонансного (ЭЦР) нагрева плазмы. С помощью процедуры выявления локальной связности, предложенной в §3.3, и CPC-метода проведено определение числа формирующих компонент (и их изменения во времени) для нескольких ансамблей экспериментальных данных. Продемонстрированы возможность получения содержательных физических результатов при исследовании переходного процесса, возбуждаемого в плазме стелларатора Л-2М при включении импульса дополнительного ЭЦР нагрева, на основе анализа моментных характеристик смешанной вероятностной модели для приращений наблюдений исходного процесса и повышения точности прогнозирования значений экспериментальных данных с помощью нейронных сетей за счет расширения признакового пространства указанными моментными характеристиками. Результаты опубликованы в статьях [147, 235]. Получены свидетельства о государственной регистрации программ для ЭВМ [31, 46].

В §5.3 представлены методы прогнозирования значений моментных характеристик, полученных в процессе анализа экспериментальных рядов турбулентной плазмы. Рассматриваются нейросетевые архитектуры для решения задач классификации и регрессии, причем как для сетей прямого распространения, так и для рекуррентных модификаций. Продемонстрировано построение совместных (векторных) прогнозов для всех рассматриваемых моментных характеристик – математического ожидания, дисперсии, коэффициентов асимметрии и эксцесса. Полученные результаты важны для развития вероятностно-статистического подхода к описанию эволюции турбулентных процессов в магнитоактивной высокотемпературной плазме. Результаты опубликованы в статьях [71, 72, 75, 246, 247]. Получены свидетельства о государственной регистрации программы для ЭВМ [41, 74].

Глава 6 посвящена разработке вероятностных моделей и методов исследования метеорологических и океанологических данных на основе теоретических результатов §1.3. Особое внимание уделяется вопросам выявления экстремальных наблюдений в рассматриваемых пространственно-временных рядах. Используются как статистические подходы для оценивания неизвестных параметров, так и широкий набор алгоритмов машинного обучения и нейронных сетей для решения задач заполнения пропусков и прогнозирования.

В §6.1 на основе k-ичной дискретизации исходных непрерывных дан-

ных об объемах осадков решена задача построения вероятностных и нейросетевых прогнозов для подобного рода наблюдений. Продемонстрирована достаточно высокая точность: до 97,1% успехов для однодневных и до 90,1% для двухдневных прогнозов для бинарных паттернов и до 92,2% успехов для однодневных и до 81,7% для двухдневных прогнозов для k-ичных при $k = 10$. При этом для анализа использованы исключительно базовые статистические данные об объемах осадков и не привлекаются какие-либо дополнительные сведения о метеорологических условиях. Продемонстрирована эффективность использования метода случайного поиска для выбора оптимальной конфигурации гиперпараметров для метеорологических данных. Показано, что даже сравнительно небольшое число (порядка десяти) случайно выбранных комбинаций позволяет получить точность, сопоставимую с полным перебором, при этом затраченное время оказывается весьма умеренным. Полученные результаты означают возможность реализовать предложенную методологию паттернов для нейронных сетей в виде исследовательского сервиса цифровой платформы. Результаты опубликованы в статьях [30, 73, 223, 245]. Получены свидетельства о государственной регистрации программы для ЭВМ [28, 41, 47, 58].

В §6.2 решена задача выбора в достаточной степени универсальных с точки зрения эффективности применения методов машинного обучения для заполнения пропусков в пространственно-временных метеорологических данных в произвольных географических регионах. Наилучшие результаты при последовательном решении задач классификации и регрессии получены для экстремального градиентного бустинга. Данный метод обеспечивает высокий базовый уровень при схожих настройках гиперпараметров по сравнению с другими алгоритмами. В отдельных ситуациях, в том числе за счет тонкой настройки и дополнительного расширения признакового пространства, могут быть получены более высокие значения точности, в том числе и иными методами машинного обучения. Результаты опубликованы в статьях [77, 248]. Получено свидетельство о государственной регистрации программы для ЭВМ [76]. Созданные инструменты могут быть успешно использованы и для иных видов наблюдений, в частности, данных экологического мониторинга окружающей среды.

В §6.3 предложено и обосновано использование классических и обобщенных отрицательных биномиальных и гамма-моделей для распределений длительностей «дождливых» периодов (интервалов времени, в кото-

рые осадки регистрировались непрерывно) и соответствующих им объемов осадков. Продемонстрировано высокое соответствие моделей с реальными данными. Разработан эффективный метод функционального оценивания параметров обобщенных распределений. Обобщенная теорема Реньи (теорема 1.9, доказанная в разделе 1.3), использована для обоснования появления дополнительного параметра – показателя степени в экспоненте – как индикатора неоднородности данных за счет глобальных климатических тенденций. Предложен метод оценивания неизвестных параметров в указанной теореме. Полученные результаты являются основой для разработки методов статистического определения экстремальных осадков. Они опубликованы в статьях [29, 224, 232, 292, 295, 413]. Получены свидетельства о государственной регистрации программы для ЭВМ [28, 36, 37, 47].

В §6.4 разработаны статистические методы и алгоритмы обнаружения и идентификации экстремальных наблюдений в различных временных рядах на примере осадков и их интенсивностей. На основе обобщения результатов теорем Реньи и Пикандса–Балкемы–Де Хаана предложены методы определения пороговых уровней, развивающие подходы классической теории экстремальных значений. С помощью проверки в скользящем режиме статистических гипотез об однородности выборки из объемов и интенсивностей создан метод классификации наблюдений как абсолютно, промежуточно и относительно экстремальных. С использованием асимптотического распределения экстремальных наблюдений в случае, если их число является случайным с отрицательным биномиальным распределением, разработан подход к определению экстремальных суточных объемов как превышающих квантили выбранных уровней данного распределения. Эти методы могут быть эффективно использованы и для других пространственно-временных метеорологических и иных данных, удовлетворяющих минимальным модельным предположениям, связанным с отрицательной биномиальностью числа и гамма-распределенностью самих наблюдений. Создание подобных инструментов необходимо для прогнозирования потенциально опасных явлений и процессов в глобальных климатических моделях. В частности, статистические оценки параметров вероятностных моделей могут быть использованы для расширения признакового пространства в задачах машинного обучения без необходимости увеличения объема исходных данных. Результаты опубликованы в статьях [57, 59, 60, 93, 230, 233, 291–293]. Получены свидетельства о государственной регистрации программы для

ЭВМ [32, 33, 39, 40, 44].

В §6.5 продемонстрировано применение CPC-подхода для анализа статистических закономерностей во временной эволюции тепловых потоков между океаном и атмосферой. Показано, что основная компонента с небольшой дисперсией может сопровождаться стохастически развивающимися и исчезающими компонентами с большой дисперсией. Отмечен ряд закономерностей во временной изменчивости моментных характеристик приращений значений процесса тепловых потоков. Развитый в диссертации метод на основе процедуры скользящего разделения смесей и алгоритма определения связности компонент использован для статистического оценивания коэффициентов стохастического дифференциального уравнения Ланжевена для скрытых и явных потоков тепла. На основании упорядочивания весов и дисперсий предложен метод определения доли экстремальных наблюдений в рассматриваемых временных рядах. Продемонстрирована эффективность использования разработанного для осадков и их интенсивностей модифицированного метода превышения порогового значения для выявления аномальных данных и при анализе океанологических рядов. Описан метод анализа характеристик распределений локальных трендов в потоках тепла с помощью аппроксимации обобщенными отрицательным биномиальным и гамма-распределениями. Результаты опубликованы в статьях [66, 94, 95, 294]. Получены свидетельства о государственной регистрации программы для ЭВМ [22, 50].

В главе 7 рассматриваются программные решения и комплексы, которые использовались для анализа неоднородных данных и визуализации результатов в главах 3– 6.

В §7.1 представлены графические интерфейсы для запуска CPC-метода и визуального представления его результатов с помощью динамической и диффузионых компонент, моментных характеристик и квантилей, в том числе с помощью анимированных графиков. Эти инструменты созданы с помощью языка программирования пакета `MATLAB`. Описания разработанных инструментов опубликованы в статьях [16, 23, 222]. Получены свидетельства о государственной регистрации программы для ЭВМ [20–22, 25–27].

В §7.2 описаны функциональные возможности разработанных приложений для анализа распределений длительностей и объемов осадков, реализующих методы оценивания параметров обобщенных отрицательных биномиальных и гамма-распределений, которые были описаны в §6.3. Результаты опубликованы в статье [224]. Получены свидетельства о го-

сударственной регистрации программы для ЭВМ [42, 43].

В §7.3 описана разработанная автором информационная технология для исследования стохастических процессов в плазме на основе спектрального анализа, которая включает в себя инструменты первичной обработки и подготовки данных для анализа, различные модификации EM-алгоритмов, функции для бутстреп-анализа и визуализации результатов. Обсуждаются структура и общая схема функционирования разработанного программного обеспечения. Результаты опубликованы в статьях [14, 237]. Получено свидетельство о государственной регистрации программы для ЭВМ [18].

В §7.4 рассмотрены вопросы реализации развиваемых в диссертации методов в рамках онлайн-системы для анализа информационных потоков с использованием разнообразных вероятностных моделей на основе гетерогенных вычислений, которая может предложить широкие функциональные возможности для различных групп исследователей. Соответствующие результаты опубликованы в статьях [23, 68, 243, 244]. Получены свидетельство о государственной регистрации программ для ЭВМ [69, 70].

В §7.5 обсуждаются вопросы трансформации отдельных программных решений, в том числе описанных в предшествующих разделах, в научно-образовательные сервисы цифровых платформ в полном соответствии с направлениями реализации Стратегии научно-технологического развития Российской Федерации, программой «Цифровая экономика» и общемировыми трендами на цифровизацию науки как отрасли. Результаты опубликованы в статьях [15, 24, 35, 51–53, 80, 81, 225, 433].

В **Заключении** кратко описаны проведенные исследования и полученные результаты, приведены перспективы дальнейшего их развития.

Структура и объем диссертации

Диссертация состоит из введения, **7** глав, разбитых на **33** параграфа, заключения, списка литературы из **447** источников, **28** таблиц, **175** рисунков и **30** вычислительных алгоритмов. Общий объем работы составляет **358** страниц.

Благодарности

Автор выражает искреннюю признательность своему научному консультанту доктору физико-математических наук, профессору **Виктору Юрьевичу Королеву** за полезные обсуждения, ценные рекомендации и плодотворные совместные исследования.

Глава 1

Смешанные законы как предельные в схемах взятия максимума и суммирования для случайных выборок

В этой главе будет получен вид смешанных распределений, которые являются предельными для выборок, объем которых является случайной величиной с обобщенным отрицательным биномиальным распределением. Предельные теоремы доказываются для двух важные схем – максимизации и суммирования. Первая из них ориентирована на поиск асимптотического распределения у наибольшего среди всех наблюдения (то есть, фактически, экстремального элемента), а у второго – у суммы всех наблюдений, при неограниченном росте объема выборки. Будут исследованы свойства предельных распределений, а также получен ряд результатов для важного частного случая – классического отрицательного биномиального распределения, в том числе распределения промежуточных экстремумов и центральных порядковых статистик. Указанные схемы являются корректными и удобными вероятностно-статистическими моделями в рамках анализа реальных данных. Соответствующие задачи будут решены в главе 6. Будет доказан вариант центральной предельной теоремы для случайных сумм в схеме серий, в которой в качестве предельного фигурируют произвольные нормальные смеси. Данный результат будет использован для обоснования вида моделей распределения частиц в селенологических пробах в главе 4.

1.1 Смешанные распределения. Основные определения

Рассмотрим функцию $F(x, \mathbf{y}) : \mathbb{R} \times \mathbb{Y} \to \mathbb{R}$, где $\mathbb{Y} \subseteq \mathbb{R}^m$, $m \geqslant 1$, причем при фиксированном векторе $\mathbf{y}$ она является функцией распределения относительно переменной x, а при фиксированном x – измерима по $\mathbf{y}$. Пусть P – вероятностная мера, заданная на измеримом пространстве $(\mathbb{Y}, \Sigma)$, где Σ – борелевская σ-алгебра для множества $\mathbb{Y}$.

Определение 1.1. [88] *Смесью* $F(x, \mathbf{y})$ по $\mathbf{y}$ относительно P называется функция

$$H(x) = \int\limits_{\mathbb{Y}} F(x, \mathbf{y})\, P(d\mathbf{y}),, \quad x \in \mathbb{R}, \tag{1.1}$$

причем мера P задает *смешивающее* распределение.

Определение 1.2. [88] Пусть случайный вектор $\mathbf{Y}$ имеет дискретное распределение

$$\begin{array}{cccc} \mathbf{y}_1 & \mathbf{y}_2 & \dots \\ p_1 & p_2 & \dots \end{array}$$

Тогда смесь $H(x)$ (1.1) называется *дискретной* и может быть представлена в виде

$$H(x) = \mathbb{E}F(x, \mathbf{Y}) = \sum_{i \geqslant 1} p_i F(x, y_i), \quad x \in \mathbb{R}. \tag{1.2}$$

Здесь величины $F(x, y_i)$ называются компонентами смеси $H(x)$, а $p_i \geqslant 0$ – соответствующими весами. Если число ненулевых весов в представлении (1.2) конечно, то такая смесь называется *конечной*.

Определение 1.3. [88] Пусть $\mathbf{y} = (u, v)$, причем $u > 0$, $v \in \mathbb{R}$ и допустимо представление

$$F(x, \mathbf{y}) = F\left(\frac{x - v}{u}\right), \quad x \in \mathbb{R}.$$

Тогда *сдвиг-масштабной* относительно P называется смесь вида

$$H(x) = \int\limits_{\mathbb{Y}} F\left(\frac{x - v}{u}\right) P(du, dv), \quad x \in \mathbb{R}. \tag{1.3}$$

Отметим, что если у функции $F(x, \mathbf{y})$ почти всюду существует производная $f(x, \mathbf{y})$ относительно переменной x, то выражения (1.1)–(1.3) могут быть переписаны относительно нее. Таким образом, у соответствующей смеси $H(x)$ также существует плотность.

Замечание 1.1. Пусть случайные величины X, U и V заданы на одном вероятностном пространстве, причем X и случайный вектор (U, V) стохастически независимы. Тогда смесь (1.3) может быть записана ввиде

$$H(x) = \mathbb{E}F\left(\frac{x - V}{U}\right), \quad x \in \mathbb{R} \tag{1.4}$$

и является функцией распределением случайной величины $X \cdot U + V$. Очевидно, что сдвиговая смесь $\mathbb{E}F(x - V)$ является функцией распределения $X \cdot U + V$, а масштабная $\mathbb{E}F\left(\frac{x}{U}\right)$ – функцией распределения $X \cdot U$.

Определение 1.4. [88] Пусть случайный вектор $\mathbf{Y}$ имеет следующее дискретное распределение ($k \in \mathbb{N}$)

$$\begin{matrix} (\sigma_1, a_1) & (\sigma_2, a_2) & \dots & (\sigma_k, a_k) \\ p_1 & p_2 & \dots & p_k \end{matrix}.$$

Тогда смесь называется *дискретной сдвиг-масштабной* и может быть представлена в виде

$$H(x) = \sum_{i=1}^{k} p_i F\left(\frac{x - a_i}{\sigma_i}\right), \quad x \in \mathbb{R}. \tag{1.5}$$

Рассмотрим важный частный случай для функции $F(\cdot)$ из формулы (1.5), а именно – нормальное распределение. В дальнейшем для функции распределения стандартного нормального закона $\Phi(x)$ и его плотности будут использоваться стандартные обозначения:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left\{-\frac{t^2}{2}\right\} dt, \quad \varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\}. \tag{1.6}$$

Определение 1.5. [88] *Конечная смесь нормальных законов* задается следующей функцией распределения:

$$F(x, k, \mathbf{a}, \boldsymbol{\sigma}, \mathbf{p}) = \sum_{i=1}^{k} p_i \Phi\left(\frac{x - a_i}{\sigma_i}\right), \quad x \in \mathbb{R}, \tag{1.7}$$

где $k \in \mathbb{N}$ – число компонент смеси с весами $\mathbf{p} = (p_1, \ldots, p_k)$, а величины $\mathbf{a} = (a_1 \ldots, a_k)$ и $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_k)$ – векторы математических ожиданий и средних квадратических отклонений соответственно, причем

$$a_i \in \mathbb{R}, \quad \sigma_i > 0, \quad \sum_{i=1}^{k} p_i = 1, \quad p_i \geqslant 0, \quad i = \overline{1,k}. \tag{1.8}$$

Определение 1.6. [88] Дисперсия случайной величины с распределением (1.7) может быть представлена в виде

$$\sum_{i=1}^{k} p_i \left[a_i - \sum_{i=1}^{k} p_i a_i \right]^2 + \sum_{i=1}^{k} p_i \sigma_i^2, \tag{1.9}$$

в котором первое слагаемое, не зависящее от дисперсий компонент, называется *динамической*, а второе, не зависящее от параметров сдвига, – *диффузионной* компонентами, соответственно.

Анализ данных объектов для последовательных подвыборок исходной выборки составляет суть метода скользящего разделения смесей (СРС) [88], который будет описан в разделе 2.1.

Важной отличительной характеристикой семейства конечных сдвиг-масштабных смесей нормальных законов является их *идентифицируемость*, доказанная Г. Тейчером [403, 404].

Определение 1.7. [88] Семейство конечных смесей (1.5), порожденное ядром F, удовлетворяющее условиям (1.8), *идентифицируемо*, если из равенства (при $k, m \in \mathbb{N}$)

$$\sum_{i=1}^{k} p_i F\left(\frac{x - a_i}{\sigma_i}\right) = \sum_{j=1}^{m} q_j F\left(\frac{x - b_j}{\delta_j}\right)$$

следует, что $k = m$ и для каждого индекса $i = \overline{1,k}$ существует индекс j такой, что $p_i = q_j$, $a_i = b_j$ и $\sigma_i = \delta_j$.

Отметим, что для произвольных сдвиг-масштабных смесей данное свойство не выполнено (см., например, книгу [88]), поэтому важное значение в рамках СРС-метода имеет возможность замены общей некорректно поставленной задачи разделения смесей (то есть статистического оценивания их параметров) аналогичной на основе конечных моделей. Основу данного подхода составляют результаты, связанные с устойчивостью смесей относительно смешивающего распределения (см. теоремы в разделах 2.3 и 2.4).

1.2 Обобщенные отрицательные биномиальные распределения

Перейдем к рассмотрению семейств распределений, которые будут исследованы в этой главе.

Определение 1.8. [390] *Обобщенное гамма-распределение* (GG-распределение) является абсолютно непрерывным и определяется при $r > 0$, $\gamma \in \mathbb{R}$, $\mu > 0$ своей плотностью

$$f^{GG}_{r,\gamma,\mu}(x) = \frac{|\gamma|\mu^r}{\Gamma(r)} x^{\gamma r-1} e^{-\mu x^\gamma}, \quad x \geqslant 0. \tag{1.10}$$

Данный класс был введен Э. Стейси в 1962, как содержащий одновременно распределения гамма и Вейбулла. Семейство GG-распределений содержит множество важных абсолютно непрерывных законов, сосредоточенных на неотрицательной полуоси, в том числе распределения:

– классическое гамма (при этом параметр $\gamma = 1$), включая его специальные случаи – экспоненциальный закон ($r = 1$), распределения Эрланга ($r \in \mathbb{N}$) и χ^2 ($\mu = \frac{1}{2}$);

– Накагами ($\gamma = 2$);

– полунормальное (распределение максимума стандартного винеровского процесса на $[0, 1]$) при $\gamma = 2$ и $r = \frac{1}{2}$;

– Рэлея при $\gamma = 2$ и $r = 1$;

– χ при $\gamma = 2$ и $\mu = 1/\sqrt{2}$;

– Максвелла (распределение абсолютных значений скоростей молекул в разреженном газе) при $\gamma = 2$ и $r = \frac{3}{2}$;

– Вейбулла-Гнеденко (экстремальное типа III) при $r = 1$ и $\gamma > 0$;

– усеченное экспоненциальное ($\gamma > 0$, $r = \frac{1}{\gamma}$);

– обратное гамма ($\gamma = -1$) и его специальный случай – распределение Леви (распределение момента первого достижения уровня броуновским движением) при $r = \frac{1}{2}$;

– Фреше (экстремальное типа II) при $r = 1$, $\gamma < 0$.

Многие из перечисленных распределений являются предельными в асимптотических теоремах теории вероятностей. При некоторых значениях параметров GG-распределение является смешанными экспоненциальным и геометрическим [296]. В этой главе будет получен аналог закона больших чисел (см. теорему 1.9) для случайных сумм, в котором оно является предельным. Поэтому его практическое использование в анализе данных весьма обширно: моделирование распределений размеров

дождевых капель [332], обработка данных, полученных методом радиолокационного синтезирования апертуры [211, 360, 388], анализ частоты наводнений [181], сегментация изображений без учителя [441] и многое другое.

Случайную величину с плотностью (1.10) обозначим $\overline{G}_{r,\gamma,\mu}$. При $\gamma = 1$ GG-распределение превращается в классическое гамма-распределение. Соответствующую случайную величину с плотностью $f^{GG}_{r,1,\mu}(x)$ обозначим $G_{r,\mu}$. Несложно убедиться, что выполнены следующие соотношения:

$$\overline{G}_{r,\gamma,\mu} \stackrel{d}{=} G^{1/\gamma}_{r,\mu} \Longleftrightarrow (\overline{G}_{r,\gamma,\mu})^{\gamma} \stackrel{d}{=} G_{r,\mu}, \tag{1.11}$$

где символ $\stackrel{d}{=}$ обозначает равенство по распределению.

Определение 1.9. [299] Случайная величина $N_{r,\gamma,\mu}$, $r > 0$, $\gamma \in \mathbb{R}$, $\mu > 0$, имеет дискретное распределение, называемое *обобщенным отрицательным биномиальным* (GNB), если оно для всех целых значений k определяется вероятностями

$$\mathbb{P}(N_{r,\gamma,\mu} = k) = \frac{1}{k!} \int\limits_0^{\infty} e^{-z} z^k f^{GG}_{r,\gamma,\mu}(x)\, dz, \tag{1.12}$$

то есть является смешанным пуассоновским со смешивающим GG-распределением (1.10).

В работе [299] В. Ю. Королевым и А. И. Зейфманом было также установлено, что в случае, если $r \in (0, 1]$ и $\gamma \in (0, 1]$, GNB-распределения являются и смешанными геометрическими в смысле определения, данного в статье [90]. Очевидно, что данный класс распределений обобщает классический отрицательный биномиальный закон с параметрами $r > 0$ (формы) и $p \in (0, 1)$ (вероятность успеха)

$$\mathbb{P}(N_{r,p} = k) = \frac{\Gamma(r + k)}{k!\Gamma(r)} \cdot p^r (1 - p)^k, \quad k = 0, 1, 2, \ldots,$$

который можно получить, подставляя в формулу (1.12) следующие значения параметров: $\gamma = 1$, $\mu = p(1 - p)^{-1}$. Введение дополнительного параметра γ позволяет строить на основе GNB-распределения более гибкие вероятностные модели. Семейство GNB-законов включает в себя, в частности: пуассоновское, отрицательное биномиальное (Поля), геометрическое распределения, соответствующие гамма-смешивающему закону; распределения Зихеля [377] (с обратным гамма в качестве смешивающего), Вейбулла-Пуассона [90].

Приведем некоторые обоснования такого определения GNB-распределения. Пуассоновское ядро используется в силу того, что пуассоновские процессы в силу универсального принципа неубывания энтропии в замкнутых системах являются наилучшими моделями однородных стационарных хаотических потоков событий [151]. Распределения, которые характеризуют время между соседними скачками в пуассоновском процессе (экспоненциальное), а также распределение точек, формирующих процесс (равномерное), являются наиболее энтропийными среди всех, соответственно, сосредоточенных на положительной полуоси и обладающих конечным математическим ожиданием, и с ограниченным носителем. В случае, если свойство однородности нарушается, то наилучшей моделью хаотического точечного процесса являются пуассоновские процессы со случайной интенсивностью, называемые дважды стохастическими, или процессами Кокса [151, 253]. Их одномерные распределения являются смешанными пуассоновскими.

Для GNB-распределения можно получить связь между значениями вероятностей того, что случайная величина $N_{r,\gamma,\mu}$ принимает значения k и $k+1$. Здесь и далее рассуждения будут проводиться в терминах случайных величин, поэтом предполагается, что все они заданы на одном и том же вероятностном пространстве $(\Omega, \mathfrak{F}, \mathbb{P})$.

Утверждение 1.1. *Для случайной величины $N_{r,\gamma,\mu}$ с обобщенным отрицательным биномиальным распределением (1.12) справедливо следующее соотношение:*

$$\mathbb{P}(N_{r,\gamma,\mu}=k+1)=\frac{\gamma r+k}{k+1}\,\mathbb{P}(N_{r,\gamma,\mu}=k)-\frac{|\gamma|\mu}{k+1}\,\mathbb{P}(N_{r+1,\,\gamma,\mu}=k). \quad (1.13)$$

Доказательство. Воспользуемся непосредственно определением 1.9. Имеем:

$$\mathbb{P}(N_{r,\gamma,\mu}=k)=\frac{|\gamma|\mu^r}{\Gamma(r)k!}\int\limits_0^\infty e^{-z-\mu z^\gamma}\,z^{\gamma r+k-1}\,dz=\frac{|\gamma|\mu^r}{\Gamma(r)k!}\int\limits_0^\infty e^{-z-\mu z^\gamma}\,d\frac{z^{\gamma r+k}}{\gamma r+k}=$$

$$=\frac{|\gamma|\mu^r}{\Gamma(r)k!}\times\left[\frac{z^{\gamma r+k}}{\gamma r+k}\,e^{-z-\mu z^\gamma}\Big|_0^\infty-\int\limits_0^\infty e^{-z-\mu z^\gamma}\,\frac{z^{\gamma r+k}}{\gamma r+k}(-1-\mu\gamma z^{\gamma-1})\,dz\right]=$$

$$=\frac{|\gamma|\mu^r}{\Gamma(r)k!}\times\left[\frac{1}{\gamma r+k}\int\limits_0^\infty e^{-z-\mu z^\gamma}\,z^{\gamma r+k}\,dz+\frac{\mu\gamma}{\gamma r+k}\int\limits_0^\infty e^{-z-\mu z^\gamma}\,z^{\gamma r+\gamma+k-1}\,dz\right]=$$

$$= \frac{k+1}{\gamma r + k}\,\mathbb{P}(N_{r,\gamma,\mu} = k+1) + \frac{\gamma^2 \mu^{r+1}}{(\gamma r + k)\Gamma(r) k!} \int\limits_0^\infty e^{-z-\mu z^\gamma} z^{\gamma r + k - 1 + \gamma}\, dz =$$

$$= \frac{k+1}{\gamma r + k}\,\mathbb{P}(N_{r,\gamma,\mu} = k+1) + \frac{|\gamma|\mu}{\gamma r + k}\mathbb{P}(N_{r+1,\,\gamma,\mu} = k).$$

Откуда следует справедливость соотношения (1.13). □

Замечание 1.2. Для применения формулы (1.13) необходимо знание распределения случайной величины $N_{r+1,\gamma,\mu}$.

Получим явные выражения для первых двух моментов для $N_{r,\gamma,\mu}$.

Утверждение 1.2. *Для случайной величины $N_{r,\gamma,\mu}$ с GNB-распределением вида (1.12) математическое ожидание и дисперсия имеют вид:*

$$\mathbb{E}N_{r,\gamma,\mu} = \frac{\Gamma(r + 1/\gamma)}{\mu^{1/\gamma}\Gamma(r)}.$$

$$\mathbb{D}N_{r,\gamma,\mu} = \frac{\Gamma(r + 2/\gamma)}{\mu^{2/\gamma}\Gamma(r)} - \frac{\Gamma(r + 1/\gamma)}{\mu^{1/\gamma}\Gamma(r)} \left(\frac{\Gamma(r + 1/\gamma)}{\mu^{1/\gamma}\Gamma(r)} - 1 \right). \tag{1.14}$$

Доказательство. С учетом формул (1.11) для математического ожидания имеем:

$$\mathbb{E}N_{r,\gamma,\mu} = \sum_{k=0}^\infty k \int\limits_0^\infty e^{-x} \frac{x^k}{k!} f^{GG}_{r,\gamma,\mu}(x)\, dx = \int\limits_0^\infty \left[\sum_{k=0}^\infty k e^{-x} \frac{x^k}{k!} \right] f^{GG}_{r,\gamma,\mu}(x)\, dx =$$

$$= \int\limits_0^\infty x f^{GG}_{r,\gamma,\mu}(x)\, dx = \mathbb{E}\overline{G}_{r,\gamma,\mu} = \mathbb{E}G^{1/\gamma}_{r,\mu} = \frac{\mu^r}{\Gamma(r)} \int\limits_0^\infty x^{1/\gamma + r - 1} e^{-\mu x}\, dx =$$

$$= \frac{\mu^r}{\mu^{1/\gamma + r}\Gamma(r)} \int\limits_0^\infty x^{1/\gamma + r - 1} e^{-x}\, dx = \frac{\Gamma(r + 1/\gamma)}{\mu^{1/\gamma}\Gamma(r)}.$$

Аналогично для второго момента получим

$$\mathbb{E}N^2_{r,\gamma,\mu} = \sum_{k=0}^\infty k^2 \int\limits_0^\infty e^{-x} \frac{x^k}{k!} f^{GG}_{r,\gamma,\mu}(x)\, dx = \int\limits_0^\infty \left[\sum_{k=0}^\infty k^2 e^{-x} \frac{x^k}{k!} \right] f^{GG}_{r,\gamma,\mu}(x)\, dx =$$

$$= \int\limits_0^\infty (x^2 + x) f^{GG}_{r,\gamma,\mu}(x)\, dx = \mathbb{E}\overline{G}^2_{r,\gamma,\mu} + \mathbb{E}\overline{G}_{r,\gamma,\mu} = \mathbb{E}G^{2/\gamma}_{r,\mu} + \mathbb{E}G^{1/\gamma}_{r,\mu} =$$

$$= \frac{\Gamma(r + 2/\gamma)}{\mu^{2/\gamma}\Gamma(r)} + \frac{\Gamma(r + 1/\gamma)}{\mu^{1/\gamma}\Gamma(r)}.$$

Пользуясь стандартным определением $\mathbb{D}N_{r,\gamma,\mu} = \mathbb{E}N^2_{r,\gamma,\mu} - (\mathbb{E}N_{r,\gamma,\mu})^2$,и выполняя тривиальные преобразования, приходим к выражению (1.14). □

Замечание 1.3. Из вывода формулы для $\mathbb{E}N^2_{r,\gamma,\mu}$ следует, что моменты произвольных целых порядков q GNB-распределения могут быть получены на основании факториальных моментов для распределения Пуассона и математических ожиданий случайных величин $G^{q/\gamma}_{r,\mu}$.

Сформулируем ряд известных результатов, которые будут использованы далее при доказательства теорем в этой главе.

Случайную величину с распределением Вейбулла с функцией распределения $\left[1 - e^{-x^\gamma}\right]\mathcal{I}_{x\geqslant 0}(x)$, $\gamma > 0$, будем обозначать W_γ. Здесь и далее через $\mathcal{I}_A(x)$ обозначается индикаторная функция множества A, равная единице, если $x \in A$, и нулю, если $x \notin A$.

Рассмотрим случайную величину

$$Z_{r,\mu} = \frac{\mu(G_{r,1} + G_{1-r,1})}{G_{r,1}} \stackrel{d}{=} \mu Z_{r,1} \stackrel{d}{=} \mu\big(1 + \tfrac{1-r}{r} Q_{1-r,r}\big), \tag{1.15}$$

для $r \in (0,1)$, в которой случайные величины $G_{r,1}$ and $G_{1-r,1}$ независимы, $\mu > 0$, а $Q_{1-r,r}$ – случайная величина с распределением Снедекора-Фишера, определяемая плотностью

$$q(x; 1-r, r) = \frac{(1-r)^{1-r} r^r}{\Gamma(1-r)\Gamma(r)} \cdot \frac{1}{x^r [r + (1-r)x]}, \quad x \geqslant 0. \tag{1.16}$$

Определение 1.10. [447] Случайная величина $S_{\alpha,\theta}$ имеет *строго устойчивое распределение*, если ее характеристическая функция представима в виде

$$\mathfrak{f}_{\alpha,\theta}(t) = \exp\left\{-|t|^\alpha \exp\{-\tfrac{1}{2} i\pi\theta\alpha \operatorname{sign} t\}\right\}, \quad t \in \mathbb{R},$$

где $0 < \alpha \leqslant 2$, $|\theta| \leqslant \min\{1, \frac{2}{\alpha} - 1\}$.

Лемма 1.1. [299] *Для случайной величины $N_{r,\gamma,\mu}$, $r > 0$, $\gamma \in \mathbb{R}$, $\mu > 0$, с обобщенным отрицательным биномиальным распределением при $\mu \to 0$ имеет место слабая сходимость $\mu^{1/\gamma} N_{r,\gamma,\mu} \Longrightarrow \overline{G}_{r,\gamma,1}$. Если $r \in (0,1]$ и $\gamma \in (0,1]$, то предельный закон может быть представлен в виде*

$$\overline{G}_{r,\gamma,1} \stackrel{d}{=} \frac{W_1}{S_{\gamma,1} Z^{1/\gamma}_{r,1}} \stackrel{d}{=} \frac{W_1^{1/\gamma}}{Z^{1/\gamma}_{r,1}} \stackrel{d}{=}$$

$$\stackrel{d}{=} \left(\frac{W_1 G_{r,1}}{G_{r,1} + G_{1-r,1}}\right)^{1/\gamma} \stackrel{d}{=} W_1^{1/\gamma} \cdot \big(1 + \tfrac{1-r}{r} Q_{1-r,r}\big)^{-1/\gamma}, \tag{1.17}$$

где в каждом выражении все случайные величины считаются независимыми.

Для построения асимптотических аппроксимаций для выборок большого размера рассмотрим в качестве третьего параметра в распределении 1.10 величину $\mu n^{-\gamma} > 0$. Тогда справедливы следующие равенства по распределению:

$$n^{-1}\overline{G}_{r,\gamma,\mu/n^{\gamma}} \stackrel{d}{=} \overline{G}_{r,\gamma,\mu} \stackrel{d}{=} \mu^{-1/\gamma}\overline{G}_{r,\gamma,1} \stackrel{d}{=} \mu^{-1/\gamma}G_{r,1}^{1/\gamma}. \tag{1.18}$$

Лемма 1.2. [288] *Пусть $\Lambda_1, \Lambda_2, \ldots$ – последовательность положительных случайных величин, таких, что для любого $n \in \mathbb{N}$ случайная величина Λ_n не зависит от стандартного пуассоновского процесса $P(t)$, $t \geqslant 0$. Сходимость $n^{-1}P(\Lambda_n) \Longrightarrow \Lambda$ при $n \to \infty$ к неотрицательной случайной величине Λ имеет место тогда и только тогда, когда*

$$n^{-1}\Lambda_n \Longrightarrow \Lambda, \quad n \to \infty. \tag{1.19}$$

Здесь и далее символ $\Longrightarrow$ обозначает слабую сходимость. Введем следующие обозначения:

$$\operatorname{rext}(F) = \sup\{x : F(x) < 1\}, \quad F^{-1}(a) = \inf\{x : F(x) \geqslant a\}.$$

Лемма 1.3. [297] *Пусть последовательность положительных случайных величин $\Lambda_1, \Lambda_2, \ldots$ такова, что для каждого $n \in \mathbb{N}$ случайная величина Λ_n не зависит от стандартного пуассоновского процесса $P(t)$, $t \geqslant 0$. Пусть $N_n = P(\Lambda_n)$. Предположим, что существует неотрицательная случайная величина Λ такая, что выполнено условие (1.19). Пусть $X_1, X_2, \ldots$ – последовательность независимых случайных величин с общей функцией распределения $F(x)$, и N_n при любом $n \in \mathbb{N}$ не зависит от нее. Пусть также $\operatorname{rext}(F) = \infty$, и существует такое число $\alpha > 0$, что для каждого $x > 0$ выполнено*

$$\lim_{y\to\infty} \frac{1 - F(xy)}{1 - F(y)} = x^{-\alpha}. \tag{1.20}$$

Тогда для распределения величины $M_n = \max\{X_1, \ldots, X_{N_n}\}$ справедливо следующее выражение:

$$\lim_{n\to\infty} \sup_{x\geqslant 0} \left| \mathbb{P}\left(\frac{M_n}{F^{-1}(1 - \frac{1}{n})} < x \right) - \int_0^{\infty} e^{-zx^{-\alpha}} d\,\mathbb{P}(\Lambda < z) \right| = 0.$$

Лемма 1.4. [97] *Пусть $\lambda > 0$, $X_1, X_2, \ldots$ независимые одинаково распределенные случайные величины с общей функцией распределения $F(x)$, и $P(t)$ – независящий от них стандартный пуассоновский процесс. Предположим, что существует функция распределения $H(x)$ такая, что для любого $x \in \mathbb{R}$*

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{1}{F^{-1}(1-\frac{1}{n})} \max_{1\leqslant k\leqslant n} X_k < x\right) = H(x). \tag{1.21}$$

Тогда для любого $n \in \mathbb{N}$

$$\left|\mathbb{P}\left(\frac{1}{F^{-1}(1-\frac{1}{n})} \max_{1\leqslant k\leqslant P(n\lambda)} X_k < x\right) - H^{\lambda}(x)\right| \leqslant$$
$$\leqslant \left|n\left[1 - F\left(xF^{-1}(1-\tfrac{1}{n})\right)\right] - \log H(x)\right| \lambda H^{\lambda}(x).$$

Лемма 1.5. [85, 86] *Рассмотрим последовательность случайных величин $Y_1, Y_2, \ldots$, а также натуральнозначные величины $N_1, N_2, \ldots$ такие, что для каждого $n \in \mathbb{N}$ случайная величина N_n не зависит от последовательности $Y_1, Y_2, \ldots$ Предположим, что существует неограниченно возрастающая последовательность положительных чисел $\{b_n\}_{n\geqslant 1}$ и случайная величина Y такая, что $b_n^{-1} Y_n \Longrightarrow Y$ при $n \to \infty$. Тогда, если существует неограниченно возрастающая последовательность положительных чисел $\{d_n\}_{n\geqslant 1}$ и случайная величина N такая, что*

$$d_n^{-1} b_{N_n} \Longrightarrow N, \quad n \to \infty \tag{1.22}$$

то

$$d_n^{-1} Y_{N_n} \Longrightarrow Y \cdot N, \quad n \to \infty \tag{1.23}$$

причем величины в правой части (1.23) независимы.

1.3 Асимптотические распределения для выборок с обобщенным отрицательным биномиальным объемом

В данном разделе получим вид асимптотического распределения для величины $M_n = \max\{X_1, \ldots, X_{N_n}\}$ (см. лемму 1.3) при условии, что объем рассматриваемой выборки N_n является случайной величиной с обобщенным отрицательным биномиальным распределением, и исследуем его свойства. Также будет рассмотрен случай классического закона,

для которого предельное распределение допускает достаточно простое аналитическое представление, которое удобно использовать при решении прикладных задач вероятностно-статистического моделирования.

1.3.1 Асимптотические распределения максимума

Теорема 1.1. *Пусть* $n \in \mathbb{N}$, $r > 0$, $\gamma > 0$, $\mu > 0$ *и* $N_{r,\gamma,\mu/n^\gamma}$ *– случайная величина, имеющая обобщенное отрицательное биномиальное распределение вида* (1.12). *Пусть* $X_1, X_2, \ldots$ *– независимые одинаково распределенные случайные величины с общей функцией распределения* $F(x)$. *Предположим, что* $\operatorname{rext}(F) = \infty$ *и существует такое число* $\lambda > 0$, *что при любом* $x > 0$ *справедливо соотношение* (1.20). *Тогда*

$$\lim_{n\to\infty} \sup_{x\geqslant 0} \left| \mathbb{P}\left(\frac{\max\{X_1, \ldots, X_{N_{r,\gamma,\mu/n^\gamma}}\}}{F^{-1}(1-\frac{1}{n})} < x \right) - F_{\lambda,\gamma,\mu,r}(x) \right| = 0,$$

где

$$F_{\lambda,\gamma,\mu,r}(x) = \int_0^\infty e^{-zx^{-\lambda}} f^{GG}_{r,\gamma,\mu}(z)\, dz \equiv \mathbb{P}(M_{\lambda,\gamma,\mu,r} < x), \quad x \geqslant 0. \tag{1.24}$$

При этом предельная случайная величина $M_{\lambda,\gamma,\mu,r}$ *допускает следующие представления:*

$$M_{\lambda,\gamma,\mu,r} \stackrel{d}{=} \frac{\overline{G}_{r,\lambda\gamma,\mu}}{W_\lambda} \stackrel{d}{=} \left(\frac{\overline{G}_{r,\gamma,\mu}}{W_1} \right)^{1/\lambda} \stackrel{d}{=} \mu^{-1/\lambda\gamma} \left(\frac{G_{r,1}}{W_\gamma} \right)^{1/\lambda\gamma}, \tag{1.25}$$

причем все случайные величины являются независимыми.

Доказательство. GNB-распределение (1.12), согласно определению 1.9, является смешанным пуассоновским со смешивающим обобщенным гамма. Поэтому $N_{r,\gamma,\mu/n^\gamma} \stackrel{d}{=} P\left(\overline{G}_{r,\gamma,\mu/n^\gamma}\right)$. Тогда с учетом выражений (1.18), леммы 1.2, в которой считаем $\Lambda_n = \overline{G}_{r,\gamma,\mu/n^\gamma}$ и результата леммы 1.3) с учетом абсолютой непрерывности предельного распределения следует справедливость соотношений (1.24).

Легко видеть, что $M_{\lambda,\gamma,\mu,r} \stackrel{d}{=} \overline{G}^{1/\lambda}_{r,\gamma,\mu} {W_\lambda}^{-1}$, то есть предельный закон представляет собой масштабную смесь распределений Фреше (обратного распределения Вейбулла $e^{-x^{-\lambda}}$, $\lambda > 0$), в которой смешивающим выступает обобщенное гамма. С учетом соотношения $\overline{G}_{r,\gamma,\mu} \stackrel{d}{=} G^{1/\gamma}_{r,\mu}$ следует

справедливость представлений (1.25) (независимость соответствующих случайных величин предполагается).

□

Приведенный далее результат является следствием теоремы 1.1, однако в силу своей важности при анализе реальных процессов (см. главу 6) будет сформулирован как теорема.

Теорема 1.2. *Пусть выполнены условия теоремы 1.1, однако объем выборки является отрицательным биномиальным, то есть рассматривается случайная величина N_{r,p_n}, имеющая отрицательное биномиальное распределение с параметрами $r > 0$ и $p_n = \min\{q, \mu/n\}$, где $q \in (0,1)$, $n \in \mathbb{N}$, $\mu > 0$. Тогда*

$$\lim_{n\to\infty} \sup_{x\geqslant 0} \left| \mathbb{P}\left(\frac{\max\{X_1, \ldots, X_{N_{r,p_n}}\}}{F^{-1}(1-\frac{1}{n})} < x \right) - F_{\lambda,\mu,r}(x) \right| = 0,$$

где

$$F_{\lambda,\mu,r}(x) = \left(\frac{\mu x^\lambda}{1+\mu x^\lambda} \right)^r \equiv \mathbb{P}(M_{\lambda,\mu,r} < x), \quad x \geqslant 0. \tag{1.26}$$

При этом предельная случайная величина $M_{\lambda,\mu,r}$ допускает следующие представления:

$$M_{\lambda,\mu,r} \stackrel{d}{=} \frac{G_{r,\mu}^{1/\lambda}}{W_\lambda} \stackrel{d}{=} \left(\frac{Q_{r,1}}{\mu r} \right)^{1/\lambda},$$

причем все случайные величины являются независимыми.

Доказательство. Данный результат соответствует случаю $\gamma = 1$ в теореме 1.1, а также может быть получен непосредственно с учетом того, что $N_{r,p_n} \stackrel{d}{=} P(G_{r,p_n/(1-p_n)})$. Аналитический вид предельного распределения $F_{\lambda,\mu,r}(x)$ (1.26) следует из следующих соотношений:

$$\frac{\mu^r}{\Gamma(r)} \int_0^\infty e^{-z(\mu+x^{-\gamma})} z^{r-1} dz = \frac{\mu^r}{\Gamma(r)(\mu+x^{-\gamma})^r} \int_0^\infty e^{-z} z^{r-1} dz = \left(\frac{\mu x^\gamma}{1+\mu x^\gamma} \right)^r.$$

В силу равенства $W_\lambda^{-1} \stackrel{d}{=} G_{1,1}^{-1/\lambda}$, то получим

$$M_{r,\gamma,\mu} \equiv \left(\frac{G_{r,\mu}}{G_{1,1}} \right)^{1/\gamma} \stackrel{d}{=} \left(\frac{G_{r,1}}{\mu G_{1,1}} \right)^{1/\gamma} \stackrel{d}{=} \left(\frac{Q_{r,1}}{\mu r} \right)^{1/\gamma},$$

где участвующие случайные величины независимы.

□

Замечание 1.4. Плотность предельного распределения $F_{\lambda,\mu,r}(x)$ (1.26) имеет вид

$$f_{\lambda,\mu,r}(x) = \frac{r\lambda\mu^r x^{\lambda r-1}}{(1+\mu x^\lambda)^{r+1}} = \frac{\lambda r \mu^r}{x^{1+\lambda}(\mu + x^{-\lambda})^{r+1}}, \quad x > 0.$$

Очевидно, что $f_{\lambda,\mu,r}(x) = O(x^{-1-\lambda})$ при $x \to \infty$. Поэтому у предельного распределения существуют лишь моменты порядков $\delta < \lambda$.

Изучим более подробно свойства предельного распределения для обобщенного случая. Во всех теоремах далее подразумевается, что все случайные величины независимы.

Теорема 1.3. *Распределение случайной величины $M_{\lambda,\gamma,\mu,r}$ допускает следующие представления.*
(i) Если $r \in (0,1]$, то оно является масштабной смесью отношений двух независимых вейбулловских случайных величин и $Z_{r,1}$ (1.15):

$$M_{\lambda,\gamma,\mu,r} \stackrel{d}{=} (\mu Z_{r,1})^{-1/\lambda\gamma} \cdot \frac{W_{\lambda\gamma}}{W_\gamma}.$$

(ii) Если $\gamma \in (0,1]$, то оно является масштабной смесью темперированного распределения Снедекора-Фишера с параметрами r and 1 и положительного строго устойчивого закона

$$M_{\lambda,\gamma,\mu,r} \stackrel{d}{=} \left(\frac{S_{\gamma,1}}{\mu r} \cdot Q_{r,1}\right)^{1/\lambda\gamma}.$$

(iii) Если $\gamma \in (0,1]$ и $r \in (0,1]$, то оно является смесью распределений Парето ($\mathbb{P}(\Pi_\lambda > x) = (x^\lambda + 1)^{-1}$, $x \geqslant 0$):

$$M_{\lambda,\gamma,\mu,r} \stackrel{d}{=} \Pi_\lambda \left(S_{\gamma,1} Z_{r,1}^{1/\gamma}\right)^{-1/\lambda}.$$

(iv) Если $r \in (0,1]$ и $\lambda\gamma \in (0,1]$, то оно представимо в виде смеси полунормальных законов:

$$M_{\lambda,\gamma,\mu,r} \stackrel{d}{=} |X| \cdot \frac{\sqrt{2W_1}}{\mu^{1/\lambda\gamma} W_\lambda S_{\lambda\gamma,1} Z_{r,1}^{1/\lambda\gamma}}.$$

Доказательство. Для доказательства пункта (i) достаточно рассмотреть величину, стоящую в правой части (1.25) и использовать соотношения $W_1^{1/\gamma} \stackrel{d}{=} W_\gamma$ и $G_{r,\mu} \stackrel{d}{=} W_1 Z_{r,\mu}^{-1}$ (по условиям теоремы $0 < r < 1$ и

случайные величины W_1 and $Z_{r,\mu}$ независимы, то есть выполнены условия теоремы Глезера [215] о представимости гамма-распределения в виде смешанного экспоненциального).

Для доказательства пункта (ii) необходимо записать величину, стоящую в правой части (1.25), с учетом соотношения $W_\gamma \stackrel{d}{=} W_1 \cdot S_{\gamma,1}^{-1}$, которое выполняется для $\gamma \in (0,1]$, если случайные величины в правой части независимы [290], и использовать определение распределения Снедекора-Фишера как отношения двух независимых гамма-распределенных случайных величин [275].

Для доказательства пункта (iii) достаточно преобразовать второе выражение в представлениях (1.25) с учетом формулы (1.17), а также учесть, что распределение отношения двух экспоненциально распределенных случайных величин совпадает с распределением Π_1.

Для доказательства пункта (iv) достаточно преобразовать второе выражение в представлениях (1.25) с учетом формулы (1.17), а также учесть, что $W_1 \stackrel{d}{=} |X|\sqrt{2W_1}$ [290].

□

Полученные представления могут быть использованы для компьютерной симуляции наблюдений из предельного распределения случайной величины $M_{\lambda,\gamma,\mu,r}$, поскольку они основаны во многом на стандартных законах, которые обычно включаются в статистические библиотеки и пакеты.

Теорема 1.4. *Если $r \in (0,1]$, $\mu > 0$ и $\lambda\gamma \in (0,1]$, то функция распределения $F_{\lambda,\gamma,\mu,r}(x)$ является смешанной экспоненциальной:*

$$F_{\lambda,\gamma,\mu,r}(x) = 1 - \int_0^\infty e^{-ux}\, dF_A(u), \quad x \geqslant 0,$$

где $F_A(u) = \mathbb{P}\left(\mu^{1/\lambda\gamma} W_\lambda S_{\lambda\gamma,1} Z_{r,1}^{1/\lambda\gamma} < u\right)$, $u \geqslant 0$, и все указанные случайные величины независимы.

Доказательство. Для доказательства необходимо воспользоваться преобразованием (1.17) для второго выражения в (1.25), ведущего к представлению

$$M_{\lambda,\gamma,\mu,r} \stackrel{d}{=} \frac{W_1}{\mu^{1/\lambda\gamma} W_\lambda S_{\lambda\gamma,1} Z_{r,1}^{1/\lambda\gamma}}, \tag{1.27}$$

которое и завершает доказательство.

□

Следствие 1.1. *В условиях теоремы 1.4 функция распределения* $F_{\lambda,\gamma,\mu,r}(x)$ *является безгранично делимой.*

Доказательство. Доказательство вытекает из известного результата Ч. Голди [219] о том, что распределение произведения двух неотрицательных независимых случайных величины является безгранично делимым, если одна из них – экспоненциальная. Это условие, очевидно, в данном случае выполнено (см. представление (1.27)).

□

Теперь получим явный вид моментов случайной величины $M_{\lambda,\gamma,\mu,r}$.

Теорема 1.5. *Для моментов порядка* $0 < \delta < \lambda$ *случайной величины* $M_{\lambda,\gamma,\mu,r}$ *справедливо следующее представление:*

$$\mathbb{E}M^{\delta}_{\lambda,\gamma,\mu,r} = \frac{\Gamma\left(r+\frac{\delta}{\lambda\gamma}\right)\Gamma\left(1-\frac{\delta}{\lambda}\right)}{\mu^{\delta/\lambda\gamma}\Gamma(r)}. \tag{1.28}$$

Доказательство. Из выражения (1.17) следует, что

$$\mathbb{E}M^{\delta}_{\lambda,\gamma,\mu,r} = \mu^{-\delta/\lambda\gamma}\mathbb{E}G^{\delta/\lambda\gamma}_{r,1}\cdot\mathbb{E}W_1^{-\delta/\lambda}. \tag{1.29}$$

Непосредственно можно убедиться в справедливости следующих соотношений для моментов распределений гамма и Фреше:

$$\mathbb{E}G^{\delta/\lambda\gamma}_{r,1} = \Gamma\left(r+\frac{\delta}{\lambda\gamma}\right)/\Gamma(r), \quad EW_1^{-\delta/\lambda} = \Gamma\left(1-\frac{\delta}{\lambda}\right),$$

которые при подстановке в формулу (1.29) ведут к выражению (1.28). □

Рассмотрим вопросы скорости сходимости к предельному распределению в теореме 1.1.

Теорема 1.6. *Пусть в условиях теоремы 1.1 случайные величины* $X_1, X_2, \ldots$ *имеют одинаковое распределение Парето вида*

$$F(x) = 1 - \frac{c}{ax^{\lambda}+c}, \quad x \geqslant 0. \tag{1.30}$$

для некоторых $a > 0$, $c > 0$ *и* $\lambda > 0$. *Тогда для любого* $x \in \mathbb{R}$

$$\left|\mathbb{P}\left(\left[\frac{a}{c(n-1)}\right]^{1/\gamma}\max_{1\leqslant k\leqslant N_{r,\gamma,\mu/n^{\gamma}}}X_k < x\right) - F_{\lambda,\gamma,\mu,r}(x)\right| \leqslant$$

$$\leqslant \left|\frac{x^{\lambda}-1}{x^{\lambda}(n-1)+1}\right|\cdot\frac{\Gamma(r+\frac{1}{\gamma})}{\mu^{1/\gamma}\Gamma(r)}.$$

Доказательство. Прежде всего, проверим, что для распределения Парето (1.30) выполнено условие (1.20). Имеем

$$\frac{1-F(xy)}{1-F(y)} = \frac{ay^\lambda + c}{ax^\lambda y^\lambda + c} \longrightarrow x^{-\lambda}$$

при $y \to \infty$. Таким образом, условие (1.20) выполнено, что означает, что в выражении (1.21) распределение $H(x) = e^{-x^{-\lambda}}$ в соответствии с классической экстремальной теорией [5]. В рассматриваемом случае:

$$F^{-1}\Big(1-\frac{1}{n}\Big) = \left[\frac{c(n-1)}{a}\right]^{1/\lambda},\ F\left(xF^{-1}\Big(1-\frac{1}{n}\Big)\right) = 1 - \frac{1}{x^\lambda(n-1)+1},$$
$$n\left[1 - F\left(xF^{-1}\left(1-\frac{1}{n}\right)\right)\right] - \log H(x) = \frac{x^\lambda - 1}{x^\lambda(n-1)+1}.$$

Тогда, согласно лемме 1.4, имеем:

$$\left|\mathbb{P}\left(\left[\frac{a}{c(n-1)}\right]^{1/\gamma} \max_{1\leqslant k\leqslant N_{r,\gamma,\mu/n^\gamma}} X_k < x\right) - F_{\lambda,\gamma,\mu,r}(x)\right| \leqslant$$
$$\leqslant \int_0^\infty \left|\mathbb{P}\left(\left[\frac{a}{c(n-1)}\right]^{1/\gamma} \max_{1\leqslant k\leqslant P(nz)} X_k < x\right) - e^{-zx^{-\lambda}}\right| f^{GG}_{r,\gamma,\mu}(z)\,dz \leqslant$$
$$\leqslant \left|\frac{x^\lambda - 1}{x^\lambda(n-1)+1}\right| \cdot \int_0^\infty \lambda e^{-zx^{-\lambda}} f^{GG}_{r,\gamma,\mu}(z)\,dz \leqslant \left|\frac{x^\lambda - 1}{x^\lambda(n-1)+1}\right| \cdot \mathbb{E}\overline{G}_{r,\gamma,\mu} =$$
$$= \left|\frac{x^\lambda - 1}{x^\lambda(n-1)+1}\right| \cdot \frac{\Gamma(r+\frac{1}{\gamma})}{\mu^{1/\gamma}\Gamma(r)}.$$

Таким образом, теорема доказана. □

Теорема 1.6 означает, что скорость сходимости к предельному распределению в теореме 1.1 составляет $O(\mu^{1/\gamma}n^{-1})$ при $\mu/n^\gamma \to 0$.

Замечание 1.5. Рассуждая аналогичным образом, в условиях теоремы 1.2 можно показать, что для $0 < p \leqslant \frac{1}{2}$:

$$\sup_{x\geqslant 0}\left|\mathbb{P}\left(\frac{p}{1-p}\max_{1\leqslant k\leqslant N_{r,p}} X_k < x\right) - \left(\frac{x^\lambda}{1+x^\lambda}\right)^r\right| \leqslant$$
$$\leqslant \frac{pr}{1-2p}\cdot\sup_{y\geqslant 0}\frac{y^r}{(1+y)^{r+1}} = \frac{p}{1-2p}\cdot\left(\frac{r}{r+1}\right)^r,$$

то есть в этом случае скорость сходимости составляет $O(p)$ при $p \to 0$.

1.3.2 Асимптотические распределения порядковых статистик и выборочных квантилей

Пусть $X_1, X_2, \ldots$ – независимые одинаково распределенные случайные величины с общей функцией распределения $F(x)$, удовлетворяющей условию (1.20). Рассмотрим порядковые статистики $X_{(1)}, \ldots, X_{(n)}$, $n \in \mathbb{N}$, построенные по выборке неслучайного объема $X_1, \ldots, X_n$. Пусть $m \in \mathbb{N}$, причем $m < n$. Рассмотрим m-ю «экстремальную» порядковую статистику $X_{(n-m+1)}$. Отметим, что подобная задача может быть актуальна в прикладных исследованиях в ситуации, когда в выборке есть несколько относительно близких «больших» значений. Пусть N – натуральнозначная случайная величина, не зависящая от последовательности $X_1, X_2, \ldots$. Обозначим

$$M_{N,m} = \begin{cases} X_{(N-m+1)}, & \text{если } N > m, \\ 0, & \text{если } N < m. \end{cases}$$

Как показано в книге [97], лемма 1.3 остается справедливой и для величин $M_{N,m}$, а предельное распределение имеет вид

$$H^{(m)}(x) = \sum_{j=0}^{m-1} \frac{x^{-\gamma j}}{j!} \int_0^\infty z^j e^{-zx^{-\gamma}} d\mathbb{P}(\Lambda < z), \quad x \geqslant 0.$$

С учетом этого факта, повторяя рассуждения из теоремы 1.2, получим следующий результат.

Теорема 1.7. *В условиях теоремы 1.2 справедливо следующее соотношение*

$$\lim_{n\to\infty} \sup_{x\geqslant 0} \left| \mathbb{P}\left(\frac{M_{N_{r,p_n},m}}{F^{-1}(1-\frac{1}{n})} < x \right) - F^{(m)}_{\lambda,\mu,r}(x) \right| = 0,$$

где

$$F^{(m)}_{\lambda,\mu,r}(x) = \left(\frac{\lambda x^\gamma}{1+\lambda x^\gamma} \right)^r \sum_{j=0}^{m-1} \frac{\Gamma(r+j)}{\Gamma(r) j! (1+\lambda x^\gamma)^j}, \quad x \geqslant 0.$$

Пусть теперь у случайных величин $X_1, X_2, \ldots$ есть плотность $p(x)$. Для некоторого $\alpha \in (0,1)$ обозначим квантиль порядка α распределения случайной величины X_1 через τ_α, а выборочный аналог – $X_{([\alpha n]+1)}$.

Теорема 1.8. *Пусть случайная величина N_{r,p_n} имеет отрицательное биномиальное распределение с параметрами $r > 0$ и $p_n = \min\{q, \mu/n\}$, где $q \in (0,1)$, $n \in \mathbb{N}$, $\mu > 0$, а последовательность случайных величин $X_1, X_2, \ldots$ является независимой и одинаково распределенной с общей плотностью $p(x)$. Кроме того, пусть эта плотность является дифференцируемой в некоторой окрестности точки τ_α, причем $p(\tau_\alpha) > 0$. Тогда*

$$\lim_{n\to\infty} \sup_{x\in\mathbb{R}} \left| \mathbb{P}\left(\frac{\sqrt{n}p(\tau_\alpha)}{\sqrt{\alpha(1-\alpha)}} \left(X_{([\alpha N_n]+1)} - \tau_\alpha \right) < x \right) - T_{\mu,r}(x) \right| = 0, \quad (1.31)$$

где $T_{\mu,r}(x)$ – распределение Стьюдента с плотностью

$$t_{\mu,r}(x) = \frac{\mu^r \Gamma(r+\frac{1}{2})}{\sqrt{2\pi}\Gamma(r)\left(\mu + \frac{x^2}{2}\right)^{r+1/2}}, \quad x \in \mathbb{R}.$$

Доказательство. В статье [288] показано, что условие (1.31) выполняется для некоторой случайной величины W тогда и только тогда, когда существует неотрицательная случайная величина U, такая, что $\mathbb{P}(W < x) \equiv \mathbb{E}\Phi\left(x\sqrt{U}\right)$ и $n^{-1}N_n \Longrightarrow U$, при $n \to \infty$. С учетом леммы 1.2 и выражения (1.18) в данном случае положим $U \stackrel{d}{=} G_{r,\mu}$. Распределение Стьюдента в этой ситуации появляется вследствие прямого вычисления вида масштабной смеси нормальных законов с указанным смешивающим распределением. □

1.3.3 Теорема Реньи для обобщенных отрицательных биномиальных сумм

Классическая теорема Реньи о сходимости прореженного точечного процесса восстановления к пуассоновскому [366] устанавливает, что распределение геометрических сумм, нормированное математическим ожиданием, сходится к экспоненциальному закону при неограниченном росте математического ожидания суммы. Данный результат представляет собой вариант закона больших чисел (ЗБЧ) для случайных сумм и играет большую роль в вопросах моделирования редких событий. В данном разделе рассмотрим ЗБЧ для сумм с обобщенным отрицательным биномиальным распределением, причем для слагаемых не предполагается независимость и одинаковая распределенность.

Теорема 1.9. *Пусть для случайных величин $X_1, X_2, \ldots$ (не обязательно независимых и одинаково распределенных) при $n \to \infty$ выполнено условие*

$$\frac{1}{n^\beta} \sum_{j=1}^{n} X_j \Longrightarrow a \tag{1.32}$$

для некоторых конечных параметров $\beta > 0$ и $a > 0$. Пусть величины $r > 0$, γ и $\mu > 0$ произвольны. Пусть для каждого $n \in \mathbb{N}$ $N_{r,\gamma,\mu/n^\gamma}$ – случайная величина, имеющая обобщенное отрицательное биномиальное распределение вида (1.12), независимая от последовательности $X_1, X_2, \ldots$. Тогда при $n \to \infty$

$$\frac{a\mu^{\beta/\gamma}}{n^\beta} \sum_{j=1}^{N_{r,\gamma,\mu/n^\gamma}} X_j \Longrightarrow \overline{G}_{r,\gamma/\beta,1} \stackrel{d}{=} G_{r,1}^{\beta/\gamma}.$$

Доказательство. Из представлений (1.18) следует, что

$$\frac{\mu^{1/\gamma}}{n} \cdot N_{r,\gamma,\mu/n^\gamma} \Longrightarrow \overline{G}_{r,\gamma,1} \tag{1.33}$$

при $n \to \infty$. В лемме 1.5 положим, основываясь на условии (1.32), $b_n = n^\beta/a$, $N_n \stackrel{d}{=} N_{r,\gamma,\mu/n^\gamma}$. Тогда $b_{N_n} = \frac{1}{a} N^\beta_{r,\gamma,\mu/n^\gamma}$. Из сходимости (1.33) следует, что при $n \to \infty$

$$\frac{1}{a} N^\beta_{r,\gamma,\mu/n^\gamma} \cdot \frac{\mu^{\beta/\gamma}}{n^\beta} \Longrightarrow \frac{1}{a} \overline{G}^\beta_{r,\gamma,1} \stackrel{d}{=} \frac{1}{a} \overline{G}_{r,\gamma/\beta,1} \stackrel{d}{=} \frac{1}{a} G^{\beta/\gamma}_{r,1}. \tag{1.34}$$

Полагая $d_n = n^\beta/\mu^{\beta/\gamma}$ и используя выражение (1.34) в качестве (1.22) в лемме 1.5, получим соотношение (1.23) в следующем виде

$$\frac{\mu^{\beta/\gamma}}{n^\beta} \sum_{j=1}^{N_{r,\gamma,\mu/n^\gamma}} X_j \Longrightarrow \frac{1}{a} \overline{G}_{r,\gamma/\beta,1} \stackrel{d}{=} \frac{1}{a} G^{\beta/\gamma}_{r,1},$$

что завершает доказательство теоремы. □

Если в теореме 1.9 $r = \gamma = \beta = 1$, то данный результат обобщает результат Реньи на случай разнораспределенных и не обязательно одинаково распределенных слагаемых [279]. При $\beta = 1$ теорема 1.9 представляет собой ЗБЧ для обобщенных отрицательных биномиальных сумм [299], а в случае $\gamma = 1$ – ЗБЧ для отрицательных биномиальных сумм [293], в том числе и для случая, когда $\beta \neq 1$.

1.4 Центральная предельная теорема для случайных сумм в схеме серий

Пусть $n, k \in \mathbb{N}$ и $\{S_{n,k}\}$ – схема серий случайных величин, а $a_{n,k}$ и $b_{n,k} > 0$ – действительные числа, $n, k \in \mathbb{N}$. Независимость строк $\{S_{n,k}\}_{k \geqslant 1}$ не предполагается. Обозначим через

$$\left\{ Y_{n,k} \equiv \frac{S_{n,k} - a_{n,k}}{b_{n,k}} \right\}_{n,k \in \mathbb{N}} \tag{1.35}$$

семейство случайных величин с некоторыми характеристическими функциями $f_{Y_{n,k}}(t)$, $t \in \mathbb{R}$. Рассмотрим семейство $\{N_n\}_{n \in \mathbb{N}}$ неотрицательных целочисленных случайных величин, причем для каждого из номеров n и k случайные величины N_n и $S_{n,k}$ независимы. Пусть Y – некоторая случайная величина с функцией распределения $F_Y(x)$.

Определение 1.11. [298] Будем говорить, что для семейства случайных величин $Y_{n,k}$ (1.35) и некоторой случайной величины Y выполнено *условие согласованности*, если для любого $T \in (0, \infty)$ для их характеристических функций $f_{Y_{n,k}}(t)$ и $f_Y(t)$, $t \in \mathbb{R}$, выполнено соотношение:

$$\lim_{n \to \infty} \mathbb{E} \sup_{|t| \leqslant T} \left| f_{Y_{n,N_n}}(t) - f_Y(t) \right| = 0. \tag{1.36}$$

Отметим, что при наличии построчной сходимости $Y_{n,k}$ к Y для любых $n \in \mathbb{N}$ и $T \geqslant 0$ выполнено [6] условие вида:

$$\lim_{k \to \infty} \sup_{|t| \leqslant T} |f_{Y_{n,k}}(t) - f_Y(t)| = 0.$$

Пусть c_n и $d_n > 0$ – вещественные числа. Рассмотрим следующие случайные величины ($n \in \mathbb{N}$):

$$U_n = \frac{b_{n,N_n}}{d_n}, \quad V_n = \frac{a_{n,N_n} - c_n}{d_n}, \quad Z_n = \frac{S_{n,N_n} - c_n}{d_n}.$$

Пусть $\{X_{n,j}\}_{j \geqslant 1}$, $n \in \mathbb{N}$, схема серий построчно независимых необязательно одинаково распределенных случайных величин с функциями распределения $F_{n,j}(x)$. Обозначим

$$S_{n,k} = X_{n,1} + ... + X_{n,k}, \quad n, k \in \mathbb{N}. \tag{1.37}$$

Следующая теорема переноса устанавливает вид предельной функции распределения для случайных величин Z_n в рамках рассматриваемой схемы.

Лемма 1.6. [298] *Пусть в рамках рассматриваемой схемы выполнено условие согласованности* (1.36). *Если существуют случайные величины* U *и* V *такие, что имеет место сходимость* $(U_n, V_n) \Longrightarrow (U, V)$ *при* $n \to \infty$, *то*

$$Z_n \Longrightarrow Z \stackrel{d}{=} Y \cdot U + V \tag{1.38}$$

при $n \to \infty$, *где случайная величина* Y *не зависит от пары* (U, V).

Таким образом, из леммы 1.6 следует (см. также формулу (1.4)), что функции распределения и характеристические функции случайных величин Z и Y связаны следующим образом:

$$\begin{aligned} F_Z(x) &= \mathbb{E} F_Y\left(\frac{x-V}{U}\right), \quad x \in \mathbb{R}, \\ f_Z(t) &= \mathbb{E} f_Y(tU) e^{itV}, \quad t \in \mathbb{R}. \end{aligned} \tag{1.39}$$

То есть предельным распределением для случайных сумм Z_n со случайным индексом является сдвиг-масштабная смесь нормальных законов, предельная для последовательности случайных величин $Y_{n,k}$ с неслучайным индексом.

Для случайных величин Z и Y определим множество $\mathcal{W}(Z|Y)$ как содержащее все пары случайных величин (U, V), такие, что характеристическая функция $f_Z(t)$ представима в виде (1.39) и $\mathbb{P}(U \geqslant 0) = 1$. Какими бы ни были случайные величины Z и Y, множество $\mathcal{W}(Z|Y)$ всегда не пусто, так как всегда содержит пару $(0, Z)$. Множество $\mathcal{W}(Z|Y)$ может состоять и из большего числа элементов. Например, если Y – стандартная нормальная случайная величина и $Z \stackrel{d}{=} W_1 - W_2$, где W_1 и W_2 – независимые стандартные случайные величины со стандартным экспоненциальным распределением, то $\mathcal{W}(Z|Y) = \{(0, W_1 - W_2), \left(\sqrt{W_1}, 0\right)\}$. В этом случае $F_Z(x)$ является симметричным распределением Лапласа.

Определение 1.12. [321] Семейство случайных величин $\{X_j\}_{j \in \mathbb{N}}$ называется *слабо относительно компактным*, если каждая последовательность его элементов содержит слабо сходящуюся подпоследовательность. В конечномерном случае она эквивалентна свойству *плотности* данного семейства:

$$\lim_{R \to \infty} \sup_{n \in \mathbb{N}} \mathbb{P}(|X_n| > R) = 0$$

Пусть некоторая вероятностная метрика $L_2\left((X_1, X_2), (Y_1, Y_2)\right)$ использована для метризации слабой сходимости в пространстве двумерных случайных векторов (например, Леви-Прохорова [447]).

Лемма 1.7. [298] *Пусть случайные величины $S_{n,k}$ имеют вид (1.37), семейство случайных величин $\{Y_{n,k}\}_{n,k\in\mathbb{N}}$ слабо относительно компактно и выполнено условие согласованности (1.36). Тогда сходимость (1.38) нормализованных случайных сумм Z_n к случайной величине Z имеет место с некоторым $c_n \in \mathbb{R}$ тогда и только тогда, когда существует слабо относительно компактная последовательность пар $(U'_n, V'_n) \in \mathcal{W}(Z|Y)$, $n \in \mathbb{N}$, такая, что выполнено условие*

$$\lim_{n\to\infty} L_2\left((U_n, V_n), (U'_n, V'_n)\right) = 0. \tag{1.40}$$

Обозначим $\mu_{n,j} = \mathbb{E}X_{n,j}$, $\sigma^2_{n,j} = \mathbb{D}X_{n,j}$, причем $0 < \sigma^2_{n,j} < \infty$, $n, j \in \mathbb{N}$. Математическое ожидание $\mathbb{E}S_{n,k}$ и дисперсию $\mathbb{D}S_{n,k}$ запишем как

$$A_{n,k} = \mu_{n,1} + ... + \mu_{n,k}, \quad B^2_{n,k} = \sigma^2_{n,1} + ... + \sigma^2_{n,k}.$$

Легко видеть, что $\mathbb{E}S_{n,N_n} = \mathbb{E}A_{n,N_n}$, $\mathbb{D}S_{n,N_n} = \mathbb{E}B^2_{n,N_n} + \mathbb{D}A_{n,N_n}$, $n \in \mathbb{N}$. Для формулировки центральной предельной теоремы для случайных сумм с произвольной нормальной смесью в качестве предельного закона, рассмотрим центрированные и нормированные неслучайные суммы $Y_{n,k}$ со следующими величинами: $a_{n,k} = A_{n,k}$ и $b^2_{n,k} = B^2_{n,k}$, $n, k \in \mathbb{N}$. Тогда $c_n = \mathbb{E}A_{n,N_n}$ и $d^2_n = \mathbb{E}B^2_{n,N_n} + \mathbb{D}A_{n,N_n}$.

Теорема 1.10. *Пусть выполнено случайное условие Линдеберга: для любого $\varepsilon > 0$*

$$\lim_{n\to\infty} \mathbb{E}\frac{1}{B^2_{n,N_n}} \sum_{j=1}^{N_n} \int\limits_{|x-\mu_{n,j}|>\varepsilon B_{n,N_n}} (x-\mu_{n,j})^2 dF_{n,j}(x) = 0. \tag{1.41}$$

Тогда сходимость (1.38) Z_n к случайной величине Z имеет место для некоторых $c_n \in \mathbb{R}$ и $d_n > 0$ тогда и только тогда, когда существует слабо относительно компактная последовательность пар $\{(U'_n, V'_n)\}_{n\geqslant 1}$, такая что для любого $n \in \mathbb{N}$ выполнены условия (1.40) и

$$\mathbb{P}(Z < x) = \mathbb{E}\Phi\left(\frac{x - V'_n}{U'_n}\right), \quad x \in \mathbb{R}, \tag{1.42}$$

где $\Phi(\cdot)$ обозначает функцию функции распределения стандартного нормального закона (1.6).

Доказательство. Сначала проверим, что семейство $\{Y_{n,k}\}_{n,k\in\mathbb{N}}$ является слабо относительно компактным. Для некоторого $R > 0$ по неравенству Чебышева имеем:

$$\mathbb{P}\left(|Y_{n,k}| > R\right) = \mathbb{P}\left(\left|\frac{S_{n,k} - A_{n,k}}{B_{n,k}}\right| > R\right) \leqslant \frac{1}{R^2} \longrightarrow 0$$

при $R \to \infty$, что означает плотность указанного семейства.

В статье [298] показано, что случайное условие Линдеберга (1.41) влечет выполнение условия согласованности (1.36) с $f_Y(t) = e^{-t^2/2}$, $t \in \mathbb{R}$. Применение леммы 1.7 завершает доказательство.

□

Теорема 1.10 означает, что при выполнении достаточно общего случайного условия Линдеберга (1.41) распределение суммы случайного числа независимых случайных величин с конечными вторыми моментами приближается нормальными сдвиг-масштабными смесями, таким образом при конечных n может быть выбрана смесь для использования в качестве асимптотической аппроксимации.

Поскольку в отличие от однопараметрических нормальных смесей (чисто масштабные или сдвиговые, дисперсионно-сдвиговые (2.56)) произвольные двухпараметрические нормальные смеси не являются идентифицируемыми [88], условие (1.40), описывающее сближение последовательности пар случайных величин $\{(U_n, V_n)\}_{n \geqslant 1}$ со специальным набором пар, нельзя заменить условием сходимости этой последовательности к элементу этого множества, так чтобы характер теоремы 1.10, устанавливающей необходимые и достаточные условия, был сохранен. Однако последнее условие, являющееся более сильным по сравнению с (1.40), вместе со случайным условием Линдеберга (1.41) являются достаточными для сходимости (1.38). Сформулируем этот результат как следствие, уточняющее лемму 1.6.

Следствие 1.2. *Пусть выполнено случайное условие Линдеберга* (1.41) *и существуют* $c_n \in \mathbb{R}$, $d_n > 0$ *и пары случайных величин* (U, V) *такие, что при* $n \to \infty$

$$\left(\frac{B_{n,N_n}}{d_n}, \frac{A_{n,N_n} - c_n}{d_n} \right) \Longrightarrow (U, V).$$

Тогда при $n \to \infty$

$$\mathbb{P}\left(\frac{S_{n,N_n} - c_n}{d_n} < x \right) \Longrightarrow \mathbb{E}\Phi\left(\frac{x - V}{U} \right).$$

Глава 2

Аналитические свойства смешанных нормальных и гамма-моделей

В данной главе описан метод скользящего разделения смесей и предложено его использование в качестве базовой процедуры для статистической оценки распределений случайных коэффициентов в стохастическом дифференциальном уравнении Ланжевена. Приведены сведения о важных модификациях EM-алгоритма –медианных, которые ведут к робастным оценкам, а также стохастических, позволяющих эффективнее выбирать в качестве решений глобальные, а не локальные максимумы. Сформулированы теоремы о свойствах стохастического EM-алгоритма, об асимптотически оптимальных критериях проверки гипотез о числе компонент смеси и устойчивости конечных масштабных смесей нормальных законов относительно смешивающего распределения. Продемонстрирован вывод формул для итерационных шагов метода скользящего разделения конечных гамма-смесей и модельный пример их применения для анализа данных биржевой книги заявок. Получены новые результаты в задачах устойчивости конечных сдвиговых и дисперсионно-сдвиговых смесей нормальных законов относительно изменений параметров смешивающего распределения, являющиеся развитием результатов кандидатской диссертации. Разработаны теоретические подходы к устранению ошибок в специальной смешанной модели округления данных.

2.1 Статистическое разделение смесей

Для обнаружения и отслеживания изменений в структуре изучаемых стохастических процессов, развивающихся в течение некоторого време-

ни, особенно длительного, достаточно разумно предположить, что и описывающее их смешанное распределение (а точнее, его параметры) не остается постоянным, а эволюционирует некоторым образом. Поэтому для максимального корректного учета данного обстоятельства был предложен следующий алгоритм [88], названный методом скользящего разделения смесей (СРС). Исходная выборка – реализация изучаемого случайного процесса – разбивается на подвыборки одинаковой длины (возможно, пересекающиеся), называемые *окнами*, в рамках которых характеристики процесса считаются неизменными. Для каждого подобного окна производится оценивание параметров смешанного распределения, например, с помощью одной из модификаций EM-алгоритма, о которых будет сказано далее. После это окно сдвигается по ряду в направлении изменения астрономического времени на некоторое количество наблюдений – одно, несколько или совпадающее с размером подвыборок, в зависимости от того, насколько гладкой должна быть эволюции параметров смеси в рамках конкретного исследования. Выбор размера окна представляет собой отдельную задачу, так как слишком маленькие подвыборки имеют недостаточный объем для корректного применения статистических процедур, в то время как для больших нарушается указанное выше свойство локальной «однородности».

2.1.1 Статистическое оценивание распределений коэффициентов в уравнении Ланжевена

Во введении была отмечена важность статистического оценивания случайных коэффициентов в стохастическом дифференциальном уравнении (СДУ) Ланжевена, которое имеет следующий вид:

$$dX(t) = a(t)dt + b(t)dW, \tag{2.1}$$

определяющее случайный процесс $X(t)$, где $W(t)$ – винеровский процесс, а коэффициенты сдвига (дрейфа) $a(t)$ и масштаба (диффузии) $b(t)$ – случайны. Уравнения вида (2.1) широко используются, например, в задаче ассимиляции данных при анализе разномасштабной изменчивости геофизических переменных [149]. В финансовой математике известны специальные версии [121] уравнения (2.1), в частности, весьма популярна модель геометрического броуновского движения вида $dX(t) = aX(t)dt + bX(t)dW$, где $a \in \mathbb{R}$, $b > 0$.

При отсутствии априорной информации о «физической» структуре процесса $X(t)$ для успешного изучения и прогнозирования его эволюции

первостепенную важность приобретает задача статистического оценивания функциональных коэффициентов $a(t)$ и $b(t)$. В силу их случайности данная задача допускает как минимум две принципиально разные формулировки. Во-первых, можно пытаться найти точечные аппроксимации функций $a(t)$ и $b(t)$, и, во-вторых, статистически оценить распределения случайных величин $a(t)$ и $b(t)$. При этом дополнительные сведения о свойствах этих коэффициентов, например структура их функциональной зависимости от исходного процесса $X(t)$ [149], позволяют найти оценки числовых параметров этих моделей.

Пусть $n \geqslant 1$ и $t_0 = 0 < t_1 < \ldots < t_n$ – моменты времени, в которые наблюдается процесс $X(t)$. Для простоты предположим, что $t_i - t_{i-1} = 1$ для любого $i \geqslant 1$. Из уравнения (2.1) следует, что

$$\mathbb{P}\left(X(t_i) - X(t_{i-1}) < x\right) \approx \mathbb{E}\Phi\left(\frac{x - A_i}{B_i}\right), \tag{2.2}$$

где A_i – вещественнозначные, а B_i – положительные случайные величины. В свою очередь, для распределений случайных величин A_i и B_i, по отношению к которым берется математическое ожидание в формуле (2.2), можно использовать дискретную аппроксимацию. Тогда вместо выражения (2.2) можно применить приближение вида

$$\mathbb{P}\left(X(t_i) - X(t_{i-1}) < x\right) \approx \sum_{k=1}^{K} p_k \Phi\left(\frac{x - a_k}{b_k}\right), \tag{2.3}$$

то есть модель конечной смеси нескольких нормальных распределений (1.7) с ограничениями типа (1.8) на параметры, изменяющиеся при переходе от t_i к t_{i+1}. Очевидно, параметры p_k, a_k и b_k, формирующие так называемые динамические и диффузионные компоненты [88], зависят также от i и изменяются при переходе от t_i к t_{i+1}. Для статистического оценивания параметров в (2.3) можно использовать СРС-метод. Действительно, статистические закономерности поведения рассматриваемых процессов $X(t)$, $a(t)$, $b(t)$ изменяются во времени, нерегулярным образом, поэтому нет универсального смешивающего закона. Следовательно, изучение динамики изменения статистических закономерностей в поведении исследуемого процесса должно проводиться на последовательных интервалах времени с помощью EM-алгоритма или каких-либо из его модификаций, которые будут рассмотрены в следующих разделах.

Для изучения закономерностей эволюции процесса $X(t)$ (2.1) необходимо знать, как ведут себя случайные коэффициенты $a(t)$ и $b(t)$ как

функции времени. Таким образом, решается задача приближенного восстановления двумерного распределения

$$F_t(x,y) \equiv \mathbb{P}\left(a(t) < x, b(t) < y\right).$$

С этой целью при каждом $i = \overline{1,n}$ ищется дискретная аппроксимация распределения $F_{t_i}(x,y)$, а затем осуществляется интерполяция этой функции для остальных t. Для интерполяции необходимо установить соответствие между множествами возможных значений $\{a_k, b_k : k = \overline{1,K}\}$ случайных коэффициентов $a(t_i)$ и $b(t_i)$ при разных i, то есть восстановить эволюцию компонент аппроксимирующей смеси (2.3) во времени.

Эволюцию оценок коэффициентов сдвига (дрейфа) и масштаба (диффузии) естественным образом можно ассоциировать с описанными в определении 1.6 динамической и диффузионной компонентам дисперсии (1.9). Отметим, что первая из них зачастую учитывает характер локальных трендов в процессе, а вторая – наличие значительного количества случайных факторов, оказывающих существенное влияние на функционирование системы. Получаемые оценки коэффициентов позволяют содержательно расширять признаковое пространство в методах машинного обучения за счет использования характеристик адекватных математических моделей. В разделе 5.2 будет продемонстрировано на примере физических данных, как подобное добавление признаков на основе моментов конечных нормальных смесей, выражающихся через оценки коэффициентов 3.1, позволяет повысить эффективность прогнозирования нестационарных данных с помощью нейронных сетей.

2.1.2 EM-алгоритм для разделения конечных смесей

Классический EM-алгоритм [192] является двухэтапной итерационной процедурой поиска оценок максимального правдоподобия вектора $\boldsymbol{\theta}$ неизвестных параметров. В данном разделе рассматриваются явные выражения итерационных оценок EM-алгоритма для двух важных частных случаев конечных смесей: нормальных и гамма-распределений.

Сначала рассмотрим конечные нормальные смеси (1.7) с ограничениями (1.8). Введем обозначение

$$g_{ij}^{(m)} = \frac{\frac{p_i^{(m)}}{\sigma_i^{(m)}}\varphi\Big(\frac{x_j - a_i^{(m)}}{\sigma_i^{(m)}}\Big)}{\sum\limits_{r=1}^{k}\frac{p_r^{(m)}}{\sigma_r^{(m)}}\varphi\Big(\frac{x_j - a_r^{(m)}}{\sigma_r^{(m)}}\Big)} = \frac{\frac{p_i^{(m)}}{\sigma_i^{(m)}}\exp\Big\{-\frac{1}{2}\Big(\frac{x_j - a_i^{(m)}}{\sigma_i^{(m)}}\Big)^2\Big\}}{\sum\limits_{r=1}^{k}\frac{p_r^{(m)}}{\sigma_r^{(m)}}\exp\Big\{-\frac{1}{2}\Big(\frac{x_j - a_r^{(m)}}{\sigma_r^{(m)}}\Big)^2\Big\}}, \tag{2.4}$$

где верхний индекс (m), $m \in \mathbb{N}$ указывает на соответствие номеру итерационного шага в EM-алгоритме, $j = \overline{1,n}$ – наблюдению в скользящем CPC-окне, а $i = \overline{1,k}$ – компоненте нормальной смеси. В этом случае $\boldsymbol{\theta} = (p_1, \ldots, p_k, a_1, \ldots, a_k, \sigma_1, \ldots, \sigma_k)$, а для его оценивания используется последовательность EM-оценок вида $\boldsymbol{\theta}^{(m)} = (p_1^{(m)}, \ldots, p_k^{(m)}, a_1^{(m)}, \ldots, a_k^{(m)}, \sigma_1^{(m)}, \ldots, \sigma_k^{(m)})$.

Утверждение 2.1. [88] *С учетом обозначений (2.4) значения параметров p_i, a_i и σ_i, $i = \overline{1,k}$, на $(m+1)$-м шаге EM-алгоритма имеют вид ($j = \overline{1,n}$)*

$$p_i^{(m+1)} = \frac{1}{n}\sum_{j=1}^{n} g_{ij}^{(m)}, \quad a_i^{(m+1)} = \frac{1}{\sum_{j=1}^{n} g_{ij}^{(m)}}\sum_{j=1}^{n} g_{ij}^{(m)} x_j,$$
$$\sigma_i^{(m+1)} = \left[\frac{1}{\sum_{j=1}^{n} g_{ij}^{(m)}}\sum_{j=1}^{n} g_{ij}^{(m)}(x_j - a_i^{(m+1)})^2\right]^{1/2}.$$

В качестве критерия останова итерационных шагов используется близость значений этапах с номерами (m) и $(m+1)$ в смысле некоторой метрики, например, для некоторого заданного малого значения $\varepsilon > 0$

$$\|\boldsymbol{\theta}\|_\infty = \max_{j=\overline{1,3k}} \left|\theta_j^{(m+1)} - \theta_j^{(m)}\right| < \varepsilon, \tag{2.5}$$

где $\theta_j^{(m)}$ – элемент из вектора оцениваемых параметров $\boldsymbol{\theta}^{(m)}$ на (m)-м итерационном шаге. Обычно в качестве величины ε для конечных нормальных смесей используются значения 10^{-5}–10^{-8}, в зависимости от исходных данных.

Рассмотрим случай конечных гамма-смесей для строго положительных наблюдений, то есть будем использовать в выражении (1.2) соответствующую функцию распределения. Аналог величин (2.4) имеет вид

$$\widetilde{g}_{ij}^{(m)} = \frac{p_i^{(m)} \dfrac{\left(\mu_i^{(m)}\right)^{r_i^{(m)}}}{\Gamma\left(r_i^{(m)}\right)} x_j^{r_i^{(m)}-1} e^{-\mu_i^{(m)} x_j}}{\sum\limits_{l=1}^{k} p_l^{(m)} \dfrac{\left(\mu_l^{(m)}\right)^{r_l^{(m)}}}{\Gamma\left(r_l^{(m)}\right)} x_j^{r_l^{(m)}-1} e^{-\mu_l^{(m)} x_j}}, \tag{2.6}$$

где верхний индекс (m), $m \in \mathbb{N}$ соответствует номеру итерационного

шага в EM-алгоритме, а $\Gamma(r)$ – гамма-функция вида

$$\Gamma(r) = \int_0^\infty x^{r-1} e^{-x} dx.$$

В этом случае $\boldsymbol{\theta} = (p_1, \ldots, p_k, r_1, \ldots, r_k, \mu_1, \ldots, \mu_k)$, а для его оценивания используется последовательность EM-оценок вида $\boldsymbol{\theta}^{(m)} = (p_1^{(m)}, \ldots, p_k^{(m)}, r_1^{(m)}, \ldots, r_k^{(m)}, \mu_1^{(m)}, \ldots, \mu_k^{(m)})$.

Утверждение 2.2. *С учетом обозначений (2.6) значения параметров p_i, r_i и μ_i, $i = \overline{1,k}$, на $(m+1)$-м шаге EM-алгоритма могут быть определены из следущих соотношений ($j = \overline{1,n}$)*

$$p_i^{(m+1)} = \frac{1}{n} \sum_{j=1}^{n} \widetilde{g}_{ij}^{(m)}, \tag{2.7}$$

$$\mu_i^{(m+1)} = r_i^{(m)} \sum_{j=1}^{n} \widetilde{g}_{ij}^{(m)} \left(\sum_{j=1}^{n} x_j \widetilde{g}_{ij}^{(m)} \right)^{-1}, \tag{2.8}$$

а величины $r_1^{(m+1)}$ являются численным решением уравнения

$$\psi(r_i^{(m+1)}) - \log r_i^{(m+1)} = \sum_{j=1}^{n} \widetilde{g}_{ij}^{(m)} \log \frac{x_j \sum\limits_{j=1}^{n} \widetilde{g}_{ij}^{(m)}}{\sum\limits_{j=1}^{n} x_j \widetilde{g}_{ij}^{(m)}} \left(\sum_{j=1}^{n} \widetilde{g}_{ij}^{(m)} \right)^{-1}, \tag{2.9}$$

в котором через $\psi(x) = \Gamma'(x)(\Gamma(x))^{-1}$ обозначена дигамма-функция.

Доказательство. Выражение (2.7) получается естественным образом, исходя из смысла величин $\widetilde{g}_{ij}^{(m)}$, которые представляют собой апостериорные вероятности того, что на (m)-м шаге наблюдение x_j соответствует i-й компоненте смеси. Также, как показано в книге [88], для отыскания EM-оценок остальных параметров, необходимо определить экстремумы функции, которая в случае конечных гамма-смесей записывается в виде:

$$f(r_i^{(m)}, \mu_i^{(m)}) = \sum_{i,j} \widetilde{g}_{ij}^{(m)} \log \left(\frac{\left(\mu_i^{(m)}\right)^{r_i^{(m)}}}{\Gamma\left(r_i^{(m)}\right)} x_j^{r_i^{(m)}-1} e^{-\mu_i^{(m)} x_j} \right).$$

Рассмотрим ее частные производные по каждому из параметров. Сначала выпишем производную относительно $\mu_i^{(m)}$ и приравняем ее к нулю.

Имеем

$$f'_{\mu_i^{(m)}}(r_i^{(m)}, \mu_i^{(m)}) = \sum_j \widetilde{g}_{ij}^{(m)} \left(\frac{r_i^{(m)}}{\mu_i^{(m)}} - x_j \right) = 0,$$

откуда очевидным образом следует представление (2.8). Производная относительно $r_i^{(m)}$ имеет вид

$$f'_{r_i^{(m)}}(r_i^{(m)}, \mu_i^{(m)}) = \sum_j \widetilde{g}_{ij}^{(m)} \left(\log \mu_i^{(m)} - \frac{\Gamma'(x)}{\Gamma(x)} - \log x_j \right) = 0,$$

которое с учетом полученного выше выражения (2.8) для связи параметров $\mu_i^{(m)}$ и $r_i^{(m)}$, приводит к уравнению (2.9). Аналитическая форма его решения в общем случае не существует, поэтому в данном случае требуется применения каких-либо численных методов.

□

В качестве критерия останова, с очевидными модификациями, в данном случае также может быть использовано условия (2.5), при этом значения ε могут быть уменьшены до 10^{-3}–10^{-4} для уменьшения общей вычислительной нагрузки, связанной с необходимостью решения уравнения (2.9) на каждом итерационном шаге.

Такие распределения оказываются весьма полезным при анализе данных информационных систем, например, интернет-трафика [229] или биржевой книги заявок [227] с применением программных инструментов для реализации работы со смесями с поддержкой векторных вычислений [38].

На рисунках 2.1 и 2.2 продемонстрирована эволюция аппроксимирующей конечной гамма-смеси в зависимости от положения скользящего окна CPC-метода для финансовых данных указанного типа.

На нижнем графике на рисунке 2.2 представлена аппроксимация гистограммы соответствующей плотностью для скользящего окна с номером 221. Наглядно видны нетривиальные компоненты в смеси, отличающие ее от классического гамма-распределения, которые позволяют лучше учесть неоднородную структуру данных, в том числе и на хвостах. Рисунки также демонстрируют несколько возможных вариаций цветовых шкал для вывода трехмерных плоскостей.

2.1.3 Медианные модификации EM-алгоритма

Классический EM-алгоритм зачастую не обладает свойством устойчивости оценок, например, на практике встречаются примеры, когда за-

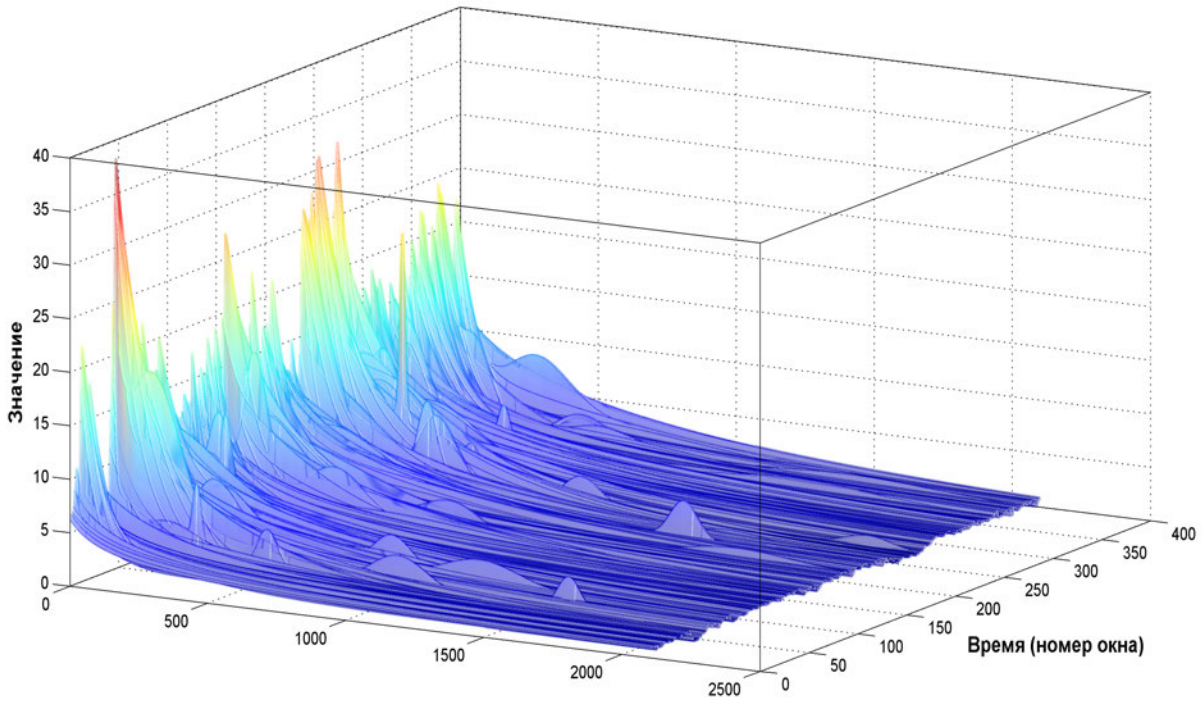

Рис. 2.1. Эволюция плотности смеси гамма-распределений

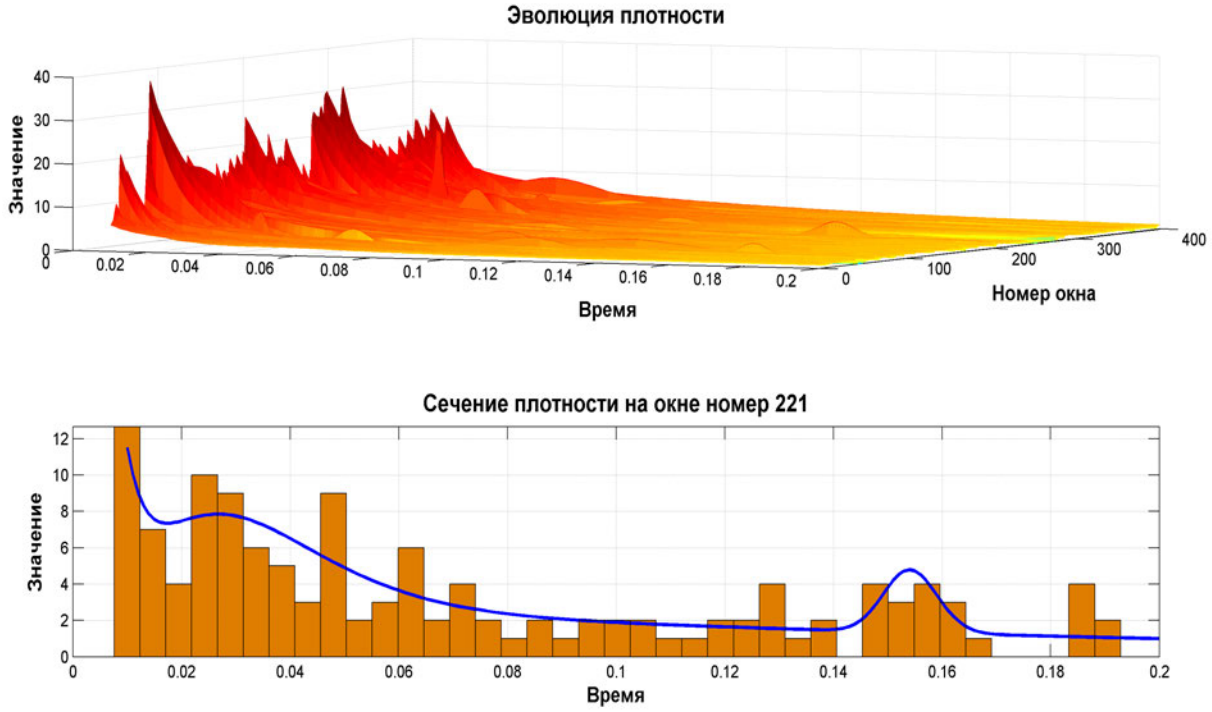

Рис. 2.2. Аппроксимирующая плотность и гистограмма (окно 221)

мена одного наблюдения в выборке из 200 элементов существенным образом влияла на получаемые результаты [88]. Для противодействия данному недостатку можно использование на итерационных шагах робастные оценки (то есть устойчивые к малым отклонениям), одной из которых,

как показал П. Хьюбер [117], являются медианы.

Автором было показано [62], что медианные оценки естественным образом возникают на E-шаге в задаче разделения конечных смесей двойных экспоненциальных распределений (Лапласа)

$$L(x) = \begin{cases} \frac{1}{2}e^{\sqrt{2}x}, & \text{если } x \leqslant 0, \\ 1 - \frac{1}{2}e^{-\sqrt{2}x}, & \text{если } x > 0, \end{cases}$$

с теми же самыми значениями параметров сдвига и масштаба компонент, что и у исходной смеси нормальных законов. В свою очередь, двойное экспоненциальное распределение можно представить в виде масштабной смеси нормальных законов при стандартном показательном смешивающем распределении [88]: если U – случайная величина такая, что $\mathbb{P}(U < x) = (1 - e^{-x})\mathcal{I}_{\{x \geqslant 0\}}(x)$, то $L(x) = \mathbb{E}\Phi\big(x(\sqrt{U})^{-1}\big)$.

Таким образом, медианная модификация EM-алгоритма сводится к замене исходной задачи разделения конечных смесей нормальных законов задачей разделения конечных смесей распределений Лапласа с аналогичными параметрами. При указанной замене исходные данные представляются в виде «зашумленной» выборки, которая производится за счет умножения параметров масштаба компонент на случайную величину со стандартным показательным распределением. При этом оценки, получаемые с помощью медианной версии EM-алгоритма в задаче разделения конечных смесей нормальных законов, приближают оцениваемые параметры, так как соответствующая последовательность оценок, получаемая EM-алгоритмом, сходится к оценкам максимального правдоподобия аналогичных параметров в модели вида конечных смесей распределения Лапласа.

Использование робастных медианных оценок параметров компонент конечных нормальных смесей позволяет получить менее зашумленные результаты. Это было подтверждено при анализе реальных данных – финансовых индексов и турбулентной плазмы [1, 65].

2.1.4 Стохастические модификации EM-алгоритма

Классический EM-алгоритм выбирает первый попавшийся локальный максимум. То есть, являясь методом локальной оптимизации, он приводит не к глобальному максимуму функции правдоподобия, а к тому локальному максимуму, который является ближайшим к начальному приближению. Для преодоления данного недостатка М. Бронятовски,

Ж. Селё и Ж. Диболт предложили [164] добавить дополнительный S-шаг (стохастический) к традиционным этапам EM-алгоритма, который реализует случайное «встряхивание» выборки (подробнее см. книгу [88]). Также ими были установлены некоторые свойства для двухкомпонентной смеси без возможности обобщения на случай произвольного числа компонент [172, 191], а С. Ф. Нильсеном [343] предложен метод доказательства сходимости для произвольного числа компонент, а также продемонстрирована асимптотическая нормальность оценок, однако при выполнении достаточно сложно верифицируемых на практике условий. В наиболее общем виде и без введения дополнительных предположений ключевые свойства стохастической модификации EM-алгоритма были установлены автором в статье [8] как следствия, вытекающие из следующей теоремы.

Теорема 2.1. [8] *Последовательность оценок* $\{\boldsymbol{\theta}^{(m)}\}$, *получаемая стохастическим EM-алгоритмом в задаче разделения идентифицируемых смесей с произвольным конечным числом компонент, представляет собой конечную однородную апериодическую эргодическую марковскую цепь*

Приведенный результат справедлив для любой версии стохастических EM-алгоритмов, в том числе и для медианных модификаций для конечных смесей нормальных законов, описанных в предыдущем разделе.

Замечание 2.1. Кратко сформулируем вытекающие из теоремы 2.1 важные свойства стохастического EM-алгоритма.

1. В случае разрешения существования пустых кластеров SEM-цепь становится стационарной только в случае попадания в поглощающее состояние. При этом можно говорить о поточечной сходимости последовательности $\{\boldsymbol{\theta}^{(m)}\}$.
2. Скорость сходимости SEM-алгоритма в случае разрешения существования пустых кластеров определяется временем попадания в поглощающее состояние.
3. В качестве оценки параметров в случае разрешения существования пустых кластеров при достижении стационарного режима берутся оценки выборки по соответствующим модификации формулам. Так можно оценить компоненту волатильности, вносящую наиболее значительный вклад.

4. Запрет существования пустых кластеров позволяет подгонять к данным модель строго k-компонентной смеси.

2.2 Асимптотически оптимальные критерии проверки гипотез о числе компонент

В модели (1.7) параметр k, определяющий число компонент в конечной смеси, на практике обычно неизвестен. Выбор слишком малых значений для него в большинстве случаев значительно ухудшает качество аппроксимации распределения данных, в то время как использование слишком больших величин увеличивает число параметров в модели, а значит, ведет существенному росту вычислительной сложности методов их оценивания и в ряде случаев – к переобучению. Таким образом, необходимо развитие подходов для корректного оценивания числа компонент.

Многие существующие подходы к определению числа компонент смеси основываются на понятии расстояния Кульбака–Лейблера [303] и носят название информационных (так как данную величину также называют энтропией по Кульбаку). В качестве примеров можно привести критерий Акаике [125], байесовский информационный критерий [373], критерий Ло [319, 320]. Их общим недостатком является то, что для корректности их применения требуется выполнение достаточно жестких условий регулярности, которые на практике обычно нарушаются (например,конечность функции правдоподобия для смесей нормальных законов), поэтому формальное применение данных критериев может приводить к ошибочным результатам.

Чтобы минимизировать возможные ошибки, возникающие из-за необходимости задавать в явном виде точное число компонент алгоритмам EM-типа, автором был предложен статистический подход для определения числа компонент по выборке [2, 9] на основе использования отношении правдоподобия и асимптотического подхода [355]. Значимый вклад в развитии этой области принадлежит Дж. Л. Ходжесу и Э. Л. Леману [314–316], Г. Е. Ноэзеру [344], В. Элберсу [127, 128], Л. ЛеКаму [309], Д. М. Чибисову [182], В. Е. Бенингу [150]. В рамках указанного подхода размер и мощность критерия одновременно отделены от нуля, при этом важную роль играют асимптотический дефект и потеря мощности. При этом распределение статистики и мощность критерия зависят от некоторого неизвестного параметра t, а величина, определяющая по-

терю мощности, позволяет сравнить мощность некоторого критерия, не зависящего от неизвестного параметра t, с наиболее мощным критерием, зависящим от t. Таким образом, можно гарантировать, что полученный критерий является асимптотически наиболее мощным, и при этом корректен для прикладного использования. Величина же дефекта критерия говорит о том, сколько дополнительных наблюдений необходимо для достижения наибольшей возможной мощности.

В данном разделе более подробно приведем соответствующие результаты из статей [2, 9], в которых получены асимптотически оптимальные (в смысле максимизации предельной мощности критерия) критерии проверки гипотез о числе компонент смеси вероятностных распределений . Будут рассмотрены два часто встречающихся при анализе реальных данных случая, соответствующие моделям добавления [2] и расщепления [9] компонент.

Итак, предположим, что каждое из независимых наблюдений $\mathbf{X}_n = (X_1, \ldots, X_n)$ имеет плотность вида (1.7) с K компонентами. Пусть $k \in \mathbb{N}$ – некоторое заданное число. Требуется проверить гипотезу вида $H_0 : K = k$ против альтернативы $H_1 : K = k + 1$. Подобная формулировка позволяет проверить значимость $(k+1)$-й компоненты в предположении, что веса p_i, $i = \overline{1, k+1}$, в модели (1.7), без ограничения общности, упорядочены по убыванию. Таким образом, возможно статистически удостовериться в значимости компоненты смеси с малым весом (то есть в необходимости добавления компоненты) или ответить на вопрос об объединении нескольких компонент с близкими параметрами в одну без существенного снижения качества аппроксимации. Упомянутые выше модели добавления и расщепления компоненты используются для сведения задачи проверки гипотез о значении дискретного параметра K к задаче проверки гипотез для некоторого непрерывного параметра $\theta \in [0, 1]$, соответствующего весу дополнительной компоненты.

2.2.1 Модель добавления компоненты

Для модели добавления компоненты предполагается, что случайная величина X_1 имеет плотность

$$p(x, \theta) = (1 - \theta) \cdot \sum_{i=1}^{k} p_i \psi_i(x) + \theta \cdot \psi_{k+1}(x) = \\ = (1 - \theta) \cdot f(x) + \theta \cdot g(x). \tag{2.10}$$

При этом рассматриваются такие версии плотностей $\psi_i(x)$, $i = \overline{1,k}$, что функция $f(x)$ строго положительна. Тогда сформулированная выше статистическая задача может быть переформулирована в виде проверки простой гипотезы $H_0^* : \theta = 0$ (то есть рассматриваемая смесь является k-компонентной) против последовательности сложных альтернатив (в смеси (2.10) $k+1$ значимая компонента)

$$H_{n,1} : \theta = \frac{t}{\sqrt{n}},$$

причем параметр $0 < t \leqslant C$, $C > 0$ неизвестен.

Введем следующее обозначение:

$$\Psi_s = \mathbb{E}_0 \left(\frac{g(X_1)}{f(X_1)} \right)^s = \int\limits_{-\infty}^{+\infty} \frac{g^s(x)}{f^{s-1}(x)}\, dx, \tag{2.11}$$

где $s = 2, 3, 4$, а функции $f(x)$ и $g(x)$ определены в выражении (2.10). Для этой модели справедлив следующий результат.

Теорема 2.2. [2] *Пусть моментные характеристики Ψ_s (2.11), $s = 2, 3, 4$, конечны, а смесь в выражении (2.10) идентифицируема. Тогда для модели добавления компоненты (2.10) критерий проверки гипотезы H_0^* против последовательности сложных альтернатив $H_{n,1}$ основан на статистике*

$$T = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{g(X_i)}{f(X_i)} - 1 \right),$$

обладающей следующими свойствами:

1. При справедливости нулевой гипотезы H_0^* эта статистика имеет нормальное распределение с параметрами 0 и $\Psi_2 - 1$ при $n \to \infty$:

$$\mathfrak{L}(T \mid H_0^*) \to N(0, \Psi_2 - 1).$$

2. При справедливости альтернативы эта статистика имеет нормальное распределение с параметрами $t\,(\Psi_2 - 1)$ и $\Psi_2 - 1$ при $n \to \infty$:

$$\mathfrak{L}(T \mid H_{n,1}) \to N(t\,(\Psi_2 - 1)\,, \Psi_2 - 1).$$

3. Данный критерий является асимптотически наиболее мощным критерием с предельной мощностью (для заданного уровня $\alpha \in (0,1)$) вида

$$\beta^*(t) = \Phi(t\sqrt{\Psi_2 - 1} - u_\alpha).$$

4. Потеря мощности этого критерия равна

$$r(t) = \frac{t^3\varphi\left(u_\alpha - t\sqrt{\Psi_2 - 1}\right)}{8\sqrt{\Psi_2 - 1}}\left(\Psi_4 + 2\Psi_3 - \Psi_2^2 - \Psi_2 - \frac{(\Psi_3 - 1)^2}{\Psi_2 - 1} - 1\right).$$

5. Асимптотический дефект этого критерия равен

$$d = \frac{t^2}{4\left(\Psi_2 - 1\right)} \cdot \left(\Psi_4 + 2\Psi_3 - \Psi_2^2 - \Psi_2 - \frac{(\Psi_3 - 1)^2}{\Psi_2 - 1} - 1\right).$$

В упомянутой статье [2] автором выписаны примеры для нескольких важных случаев, прежде всего, конечных нормальных и гамма-смесей, с проверкой корректности всех условий теоремы 2.2.

2.2.2 Модель расщепления компоненты

Для модели расщепления компоненты предполагается, что случайная величина X_1 имеет плотность

$$p(x,\theta) = \sum_{i=1}^{k-1} p_i\psi_i(x) + (p_k - \theta)\psi_k(x) + \theta\psi(x) =$$

$$= \sum_{i=1}^{k} p_i\psi_i(x) + \theta\left(\psi(x) - \psi_k(x)\right) = f(x) + \theta g(x), \qquad (2.12)$$

причем функция $\psi(x)$ является плотностью из того же семейства распределений, что и все $\psi_i(x)$. Рассматриваются такие версии плотностей ψ_i, $i = \overline{1,k}$, что функция $f(x)$ строго положительна. Отметим, что в отличие от рассмотренного в предыдущем разделе 2.10 случая, функция $g(x)$, вообще говоря, не является плотностью какого-либо распределения, поэтому нельзя осуществить непосредственный перенос результатов. Однако статистическая задача формулируется схожим образом в терминах проверки простой гипотезы H_0^* против последовательности сложных альтернатив $H_{n,t}$. Величины (2.11) имеют тот же вид в терминах функций $f(x)$ и $g(x)$, однако теперь они задаются в выражении (2.12), поэтому их аналог в дальнейшем для корректности будем обозначать как $\widetilde{\Psi}_s$.

Теорема 2.3. [9] *Пусть моментные характеристики $\widetilde{\Psi}_s$, $s = 2,3,4$, конечны, а смесь в выражении (2.12) идентифицируема. Тогда для модели*

расщепления компоненты (2.12) *критерий проверки гипотезы* H_0^* *против последовательности сложных альтернатив* $H_{n,1}$ *основан на статистике*

$$T = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{g(X_i)}{f(X_i)},$$

обладает следующими свойствами:

1. При справедливости нулевой гипотезы эта статистика имеет нормальное распределение с параметрами 0 *и* $\widetilde{\Psi}_2$ *при* $n \to \infty$*:*

$$\mathfrak{L}(T \mid H_0) \to N(0, \widetilde{\Psi}_2).$$

2. При справедливости альтернативы эта статистика имеет нормальное распределение с параметрами $t\widetilde{\Psi}_2$ *и* $\widetilde{\Psi}_2$ *при* $n \to \infty$*:*

$$\mathfrak{L}(T \mid H_{n,1}) \to N(t\widetilde{\Psi}_2, \widetilde{\Psi}_2).$$

3. Данный критерий является асимптотически наиболее мощным критерием с предельной мощностью (для заданного уровня $\alpha \in (0,1)$*) вида*

$$\beta^*(t) = \Phi(t\sqrt{\widetilde{\Psi}_2} - u_\alpha).$$

4. Потеря мощности для этого критерия составляет

$$r(t) = \frac{t^3}{8\sqrt{\widetilde{\Psi}_2}} \varphi\left(u_\alpha - t\sqrt{\widetilde{\Psi}_2}\right)\left(\widetilde{\Psi}_4 - \widetilde{\Psi}_2^2 - \frac{\widetilde{\Psi}_2^3}{\widetilde{\Psi}_2}\right).$$

5. Асимптотический дефект для этого критерия равен

$$d = \frac{t^2}{4\widetilde{\Psi}_2}\left(\widetilde{\Psi}_4 - \widetilde{\Psi}_2^2 - \frac{\widetilde{\Psi}_2^3}{\widetilde{\Psi}_2}\right).$$

В статье [9] автором рассмотрены примеры нескольких важных случаев, прежде всего, конечных нормальных и гамма-смесей, с проверкой корректности всех условий теоремы 2.3.

Отметим, что был проведен статистический эксперимент [7], продемонстрировавший эффективность предложенных критериев на тестовых выборках с различными характеристиками. Данные результаты вполне востребованы при реализации алгоритмов EM-типа. В них обычно задается некоторая оценка сверху для числа компонент в смеси, которое может уменьшаться в процессе итерационных шагов. Использование приведенных критериев позволяет проводить подобную процедуру статистически корректно.

2.3 Устойчивость конечных масштабных смесей нормальных законов относительно смешивающего распределения

Рассмотрим ряд известных результатов, которые были получены автором в области оценивания изменчивости смесей относительно возмущений параметров смешивающего распределения [10]. Аналогично разделу 2.2, будут рассмотрены модели добавления (2.10) и расщепления (2.12) компоненты, однако для важного частного случая – конечных масштабных смесей нормальных распределений.

Предположим, что каждое из независимых наблюдений $\mathbf{X}_n = (X_1, \ldots, X_n)$ имеет распределение вида (1.7) с нулевыми математическими ожиданиями у компонент, то есть

$$G(x) = \sum_{i=1}^{k} p_i \Phi(x\sigma_i) = \mathbb{E}\Phi(Ux), \tag{2.13}$$

где U – дискретная случайная величина, принимающая значения σ_i с вероятностями p_i, то есть

$$U : \begin{matrix} \sigma_1 & \sigma_2 & \ldots & \sigma_k \\ p_1 & p_2 & \ldots & p_k. \end{matrix} \tag{2.14}$$

Обозначим через $\rho(F, G)$ равномерное расстояние между функциями распределения $F(x)$ и $G(x)$:

$$\rho(F, G) = \sup_{x \in \mathbb{R}} |F(x) - G(x)|, \tag{2.15}$$

а через $L(F, G)$ – метрику Леви между функциями распределения $F(x)$ и $G(x)$:

$$L(F, G) = \inf\{h : \; G(x - h) - h \leqslant F(x) \leqslant G(x + h) + h, \; \forall x \in \mathbb{R}\}. \tag{2.16}$$

Модели добавления и расщепления компоненты могут быть представлены в виде

$$G_p(x) = \mathbb{E}\Phi(U_p\, x), \tag{2.17}$$

где дискретная случайная величина U_p будет указана в дальнейшем в явном виде для каждой модели.

2.3.1 Модель добавления компоненты

Модель добавления компоненты формализуется следующим образом. Предполагается, что каждое из независимых наблюдений $\mathbf{X}_n = (X_1, \ldots, X_n)$ имеет распределение, представимое в виде

$$G_p(x) = (1-p)\sum_{i=1}^{k} p_i \Phi(x\sigma_i) + p\Phi(x\sigma), \tag{2.18}$$

где все величины σ_i, p_i, $i = 1, \ldots, k$, считаем известными, а σ и p считаем параметрами модели, при этом $\sigma > 0$, $0 \leqslant p \leqslant 1$. Без ограничения общности для определенности будем считать, что выполнены соотношения

$$0 < \sigma \leqslant \sigma_1 \leqslant \sigma_2 \leqslant \ldots \leqslant \sigma_k. \tag{2.19}$$

Отметим, что условие отделенности параметров от нуля в формуле (2.19) также является достаточно общим и означает, что рассматриваются невырожденные нормальные законы с конечными дисперсиями.

Для данной модели дискретная случайная величина U_p в выражении (2.17) принимает следующий вид:

$$U_p : \begin{array}{ccccc} \sigma & \sigma_1 & \sigma_2 & \ldots & \sigma_k \\ p & p_1(1-p) & p_2(1-p) & \ldots & p_k(1-p). \end{array} \tag{2.20}$$

Тогда справедлива следующая теорема.

Теорема 2.4. [10] *В рамках модели добавления компоненты* (2.18) *при выполнении условий* (2.19) *справедливы следующие соотношения:*

$$L(G, G_p) \leqslant L(U, U_p) \leqslant C_1^{[1]}(\sigma_k) L^{1/2}(G, G_p),$$

где коэффициент $C_1^{[1]}(\sigma_k)$ *зависит только от известной величины* σ_k *и имеет вид*

$$C_1^{[1]}(\sigma_k) = \frac{1}{\sqrt{\varphi(\sigma_k)}} \left(1 + \frac{\sigma_k}{\sqrt{2\pi}}\right)^{1/2}. \tag{2.21}$$

Пусть существует еще одна смесь указанного типа, отличающаяся от (2.18) только весом, то есть

$$G_q(x) = (1-q)\sum_{i=1}^{k} p_i \Phi(x\sigma_i) + q\Phi(x\sigma), \tag{2.22}$$

где $0 \leqslant q \leqslant 1$. Для нее дискретная случайная величина U_q в выражении (2.17) принимает следующий вид:

$$U_q : \begin{matrix} \sigma & \sigma_1 & \sigma_2 & \dots & \sigma_k \\ q & p_1(1-q) & p_2(1-q) & \dots & p_k(1-q). \end{matrix} \tag{2.23}$$

Тогда справедлива следующая теорема.

Теорема 2.5. [10] *В рамках модели добавления компоненты* (2.18) *при выполнении условий* (2.19) *справедливы следующие соотношения:*

$$L(G_p, G_q) \leqslant L(U_p, U_q) \leqslant C_1^{[1]}(\sigma_k) L^{1/2}(G_p, G_q),$$

где коэффициент $C_1^{[1]}(\sigma_k)$ *определяется формулой* (2.21).

2.3.2 Модель расщепления компоненты

Модель расщепления компоненты формализуется следующим образом. Предполагается, что каждое из независимых наблюдений $\mathbf{X}_n = (X_1, \dots, X_n)$ имеет распределение, представимое в виде

$$G_p(x) = \sum_{i=1}^{k-1} p_i \Phi(x\sigma_i) + (p_k - p)\Phi(x\sigma_k) + p\Phi(x\sigma), \tag{2.24}$$

где все величины σ_i, p_i, $i = 1, \dots, k$, считаем известными, а σ и p считаем параметрами модели, при этом $\sigma > 0$, $0 \leqslant p \leqslant p_k$. Без ограничения общности для определенности будем считать, что выполнены соотношения

$$0 < \sigma_1 \leqslant \sigma_2 \leqslant \dots \leqslant \sigma_{k-1} \leqslant \sigma \leqslant \sigma_k. \tag{2.25}$$

Отметим, что условие отделенности параметров от нуля в формуле (2.25) также является достаточно общим и означает, что рассматриваются невырожденные нормальные законы с конечными дисперсиями.

Для данной модели дискретная случайная величина U_p в выражении (2.17) принимает следующий вид

$$U_p : \begin{matrix} \sigma_1 & \sigma_2 & \dots & \sigma & \sigma_k \\ p_1 & p_2 & \dots & p & p_k - p. \end{matrix} \tag{2.26}$$

Тогда справедлива следующая теорема.

Теорема 2.6.[10] *В рамках модели расщепления компоненты* (2.24) *при выполнении условий* (2.25) *справедливы следующие соотношения:*

$$C_2^{[2]}(\sigma_1,\sigma_k)L(G,G_p)\leqslant L(U,U_p)\leqslant C_1^{[2]}(\sigma_k)L^{1/2}(G,G_p),$$

где коэффициенты $C_j^{[2]}$, $j=1,2$, *не зависят от величин* p *и* σ *и имеют вид*

$$C_1^{[2]}(\sigma_k)=\varphi^{-1/2}(\sigma_k)\left(1+\frac{\sigma_k}{\sqrt{2\pi}}\right)^{1/2}, \tag{2.27}$$

$$C_2^{[2]}(\sigma_1,\sigma_k)=\frac{\sigma_1\sqrt{2\pi e}}{\max\{1,\sigma_k\}}. \tag{2.28}$$

Пусть существует еще одна смесь данного типа, отличающаяся от (2.24) только весом, то есть

$$G_q(x)=\sum_{i=1}^{k-1}p_i\Phi(x\sigma_i)+(p_k-q)\Phi(x\sigma_k)+q\Phi(x\sigma), \tag{2.29}$$

где $0\leqslant q\leqslant p_k$. Для нее дискретная случайная величина U_q в выражении (2.17) принимает следующий вид:

$$U_q:\begin{matrix}\sigma_1 & \sigma_2 & \dots & \sigma & \sigma_k\\ p_1 & p_2 & \dots & q & p_k-q.\end{matrix} \tag{2.30}$$

Тогда справедлива следующая теорема.

Теорема 2.7.[10] *В рамках модели расщепления компоненты* (2.24) *при выполнении условий* (2.25) *справедливы следующие соотношения:*

$$C_2^{[2]}(\sigma_1,\sigma_k)L(G_p,G_q)\leqslant L(U_p,U_q)\leqslant C_1^{[2]}(\sigma_k)L^{1/2}(G_p,G_q),$$

где коэффициенты $C_j^{[2]}$, $j=1,2$, *определяются формулами* (2.27) *и* (2.28).

Теоремы 2.4 и 2.6 означают, что близость смешивающих распределений влечет близость смесей, и наоборот, близость смесей влечет близость смешивающих распределений. Теоремы 2.5 и 2.7 означают, что если рассматриваемые смеси близки по параметрам p и q, то близки и их смешивающие распределения, и наоборот, если смешивающие распределения близки по параметрам p и q, то близки и соответствующие смеси. Близость смешивающих распределений (то есть стремление к 0 веса

p дополнительной компоненты в обеих моделях) влечет близость итоговых смесей (то есть число компонент смеси равно k, а не $k+1$) в терминах расстояния Леви. Причем справедливо и обратное утверждение. Таким образом, устанавливается взаимно однозначное соответствие между значением параметра веса и числом компонент в смеси, и поэтому становится возможным сведение задачи проверки гипотез о значении дискретного параметра к значению непрерывного параметра, как было продемонстрировано в разделе 2.2.

2.4 Устойчивость конечных сдвиговых нормальных смесей по отношению к изменениям смешивающего распределения

В данном разделе будут получены новые результаты, связанные с устойчивостью конечных сдвиговых смесей нормальных законов относительно изменений параметров смешивающего распределения. Модели такого типа возникают, например, при решении оптимизационных задач для управлении запасами, при моделировании потоков страховых выплат, при прогнозировании надежности различных систем.

2.4.1 Постановка задачи

Предположим, что каждое из независимых наблюдений $\mathbf{X}_n = (X_1, \ldots, X_n)$ имеет распределение, представимое в виде конечной сдвиговой смеси нормальных законов (то есть вида (1.7) с единичными среднеквадратическими отклонениями при справедливости всех условий (1.8)):

$$G(x) = \sum_{i=1}^{k} p_i \Phi(x - a_i) = \mathbb{E}\Phi(x - V), \tag{2.31}$$

где V – дискретная случайная величина, принимающая значения a_i с вероятностями p_i, то есть

$$V : \begin{matrix} a_1 & a_2 & \ldots & a_k \\ p_1 & p_2 & \ldots & p_k. \end{matrix} \tag{2.32}$$

Модели добавления и расщепления компоненты могут быть представлены в виде $G_p(x) = \mathbb{E}\Phi(x - V_p)$, где дискретная случайная величина V_p

определяется для каждой из моделей по-разному. Необходимо получить соотношения, связывающие расстояния Леви (2.16) между смешивающими распределениями и смесями. Перейдем к рассмотрению каждой из моделей.

2.4.2 Модель добавления компоненты

Модель добавления компоненты формализуется следующим образом. Предполагается, что каждое из независимых наблюдений $\mathbf{X}_n = (X_1, \ldots, X_n)$ имеет распределение, представимое в виде

$$G_p(x) = (1-p)\sum_{i=1}^{k} p_i \Phi(x - a_i) + p\Phi(x - a), \tag{2.33}$$

где все величины $a_i \in \mathbb{R}$, $p_i \geqslant 0$, $i = 1, \ldots, k$, считаются известными, а a и p являются параметрами модели, при этом $a \in \mathbb{R}$, $0 \leqslant p \leqslant 1$. Без ограничения общности для определенности будем считать, что выполнены соотношения

$$a_0 \leqslant a \leqslant a_1 \leqslant a_2 \leqslant \ldots \leqslant a_k. \tag{2.34}$$

Левое неравенство означает достаточно естественное для практики предположение, что рассматриваются конечные математические ожидания. Поэтому в дальнейшем считаем a_0 известным параметром модели (так как он может быть указан из некоторых разумных предположений для каждого конкретного случая).

В модели добавления компоненты дискретная случайная величина V_p имеет следующий вид:

$$V_p : \begin{array}{cccccc} a & a_1 & a_2 & \ldots & a_k \\ p & p_1(1-p) & p_2(1-p) & \ldots & p_k(1-p). \end{array} \tag{2.35}$$

Отметим, что расстояние Леви $L(V, V_p)$ (2.16) не превосходит величины p, так как расстояние между ступеньками функций распределения составляет в точности p на сегменте $[a, a_1]$ и pp_i на сегментах $[a_i, a_{i+1}]$, $i = 1, \ldots, k-1$. Изменяться могут лишь параметры a и p, величины a_i, p_i, $i = 1, \ldots, k$, считаем постоянными. Однако при фиксированном параметре p и при стремлении $a \to a_1$ очевидно, что $L(V, V_p)$ к нулю не стремится. Таким образом, без ограничения общности считаем, что $0 \leqslant p \leqslant a_1 - a$. Поэтому

$$L(V, V_p) = p. \tag{2.36}$$

Тогда справедлива следующая теорема.

Теорема 2.8. *В рамках модели добавления компоненты* (2.33) *при выполнении условий* (2.34) *и* (2.36) *расстояние Леви* $L(V, V_p)$ *между смешивающими распределениями* V *из соотношения* (2.32) *и* V_p *из соотношения* (2.35) *и расстояние Леви* $L(G, G_p)$ *между истинным распределением* $G(x)$ *из соотношения* (2.31) *и приближающей смесью* $G_p(x)$ *из соотношения* (2.33) *связывают неравенства*

$$C_1^{[1]}(a_k, a_0)L(G, G_p) \leqslant L(V, V_p) \leqslant C_2^{[1]}(a_k, a_0)L^{1/2}(G, G_p),$$

где коэффициенты $C_j^{[1]}(a_k, a_0)$, $j = 1, 2$, *зависящие только от известных величин* a_k *и* a_0, *имеют вид*

$$C_1^{[1]}(a_k, a_0) = \max\left\{1, \frac{\sqrt{2\pi}}{a_k - \min\{0, a_0\}}\right\}, \tag{2.37}$$

$$C_2^{[1]}(a_k, a_0) = \varphi^{-1/2}\Big(a_k + |a_k| - \min\{0, a_0\}\Big)\left(1 + \frac{1}{\sqrt{2\pi}}\right)^{1/2}. \tag{2.38}$$

Доказательство. Запишем оценки снизу для равномерного расстояния (2.15) между функциями распределения $G(x)$ и $G_p(x)$, воспользовавшись формулой Лагранжа:

$$\rho(G, G_p) = \sup_x |G(x) - G_p(x)| =$$

$$= \sup_x |G(x) - G(x) + p(G(x) - \Phi(x - a))| =$$

$$= p \sup_x |G(x) - \Phi(x - a)| \geqslant p|G(x_0 - a_i) - \Phi(x_0 - a)| =$$

$$= p\left|\sum_{i=1}^{k} p_i \left(\Phi(x_0 - a_i) - \Phi(x_0 - a)\right)\right| =$$

$$= p\left|\sum_{i=1}^{k} p_i(a - a_i)\varphi(\theta_i(x_0 - a_i) + (1 - \theta_i)(x_0 - a))\right| =$$

$$= p\left|\sum_{i=1}^{k} p_i(a_i - a)\varphi(x_0 - a - \theta_i(a_i - a))\right|. \tag{2.39}$$

Неравенство в соотношении (2.39) справедливо для любого x_0. Выберем значение данной величины так, чтобы воспользоваться свойством монотонного убывания плотности стандартного нормального распределения $\varphi(x)$ от положительного аргумента.

А именно потребуем выполнения условия

$$x_0 - a - \theta_i(a_i - a) \geqslant 0.$$

Откуда следует (с учетом того, что выражение в скобках в силу условий (2.34) неотрицательно и $0 \leqslant \theta_i \leqslant 1$), что

$$x_0 \geqslant a_i \tag{2.40}$$

сразу для всех номеров i. Тогда в качестве x_0 возьмем величину

$$x_0 = a_k + |a_k|. \tag{2.41}$$

Очевидно, что условие (2.40) выполняется, при этом $x_0 \geqslant 0$ и $x_0 - a \geqslant 0$. Тогда, продолжая (2.39) с учетом соотношений (2.34) и (2.36), получим

$$\rho(G, G_p) \geqslant p \left| \sum_{i=1}^{k} p_i(a_i - a)\varphi\Big(a_k + |a_k| - a - \theta_i(a_i - a)\Big) \right| \geqslant$$

$$\geqslant p \left| \sum_{i=1}^{k} p_i(a_i - a)\varphi\Big(a_k + |a_k| - a\Big) \right| \geqslant$$

$$\geqslant p \left| \sum_{i=1}^{k} p_i(a_i - a)\varphi\Big(a_k + |a_k| - \min\{0, a_0\}\Big) \right| \geqslant$$

$$\geqslant p \sum_{i=1}^{k} p_i(a_1 - a)\varphi\Big(a_k + |a_k| - \min\{0, a_0\}\Big) =$$

$$= p(a_1 - a)\varphi\Big(a_k + |a_k| - \min\{0, a_0\}\Big) \geqslant L^2(V, V_p)\varphi\Big(a_k + |a_k| - \min\{0, a_0\}\Big).$$

Воспользуемся известным неравенством для метрики Леви (см., например, книгу [**?**])

$$L(G, G_p) \leqslant \rho(G, G_p) \leqslant (1 + \max_x G'(x))L(G, G_p). \tag{2.42}$$

Воспользуемся правым неравенством из соотношения (2.42). Имеем

$$L^2(V, V_p)\varphi\Big(a_k + |a_k| - \min\{0, a_0\}\Big) \leqslant$$

$$\leqslant \rho(G, G_p) \leqslant (1 + \max_x G'(x))L(G, G_p) =$$

$$= \left(1 + \max_x \left(\sum_{i=1}^{k} p_i\varphi(x - a_i)\right)\right) L(G, G_p) \leqslant$$

$$\leqslant \left(1 + \sum_{i=1}^{k} p_i \frac{1}{\sqrt{2\pi}}\right) L(G, G_p) = \left(1 + \frac{1}{\sqrt{2\pi}}\right) L(G, G_p).$$

Окончательно получаем следующую оценку сверху для $L(V, V_p)$:

$$L(V, V_p) \leqslant \varphi^{-1/2}\Big(a_k + |a_k| - \min\{0, a_0\}\Big) \left(1 + \frac{1}{\sqrt{2\pi}}\right)^{1/2} L^{1/2}(G, G_p) =$$
$$= C_2^{[1]}(a_k, a_0) L^{1/2}(G, G_p).$$

Оценка снизу для $L(V, V_p)$ может быть найдена из соотношений

$$L(G, G_p) \leqslant \rho(G, G_p) = \sup_x |G(x) - G_p(x)| =$$
$$= p \sup_x \left| \sum_{i=1}^{k} p_i (\Phi(x - a_i) - \Phi(x - a)) \right| \leqslant$$
$$\leqslant p \sup_x \sum_{i=1}^{k} p_i |\Phi(x - a_i) - \Phi(x - a)| \leqslant$$
$$\leqslant p \sum_{i=1}^{k} p_i \sup_x |\Phi(x - a_i) - \Phi(x - a)| \leqslant p \sum_{i=1}^{k} p_i = L(V, V_p).$$

Однако можно провести оценивание и другим путем. Найдем точки экстремума функции $\Phi(x - a) - \Phi(x - a_i)$ из условия

$$\varphi(x - a) - \varphi(x - a_i) = 0.$$

Максимум достигается в точке

$$x_i^* = \frac{a + a_i}{2}.$$

Тогда, учитывая четность функции $\varphi(x)$, получим

$$p \sum_{i=1}^{k} p_i \sup_x |\Phi(x - a_i) - \Phi(x - a)| \leqslant p \sup_x \sum_{i=1}^{k} p_i |\Phi(x - a_i) - \Phi(x - a)| \leqslant$$
$$\leqslant p \sup_x \sum_{i=1}^{k} p_i |\Phi(x_i^* - a_i) - \Phi(x_i^* - a)| =$$
$$= p \sum_{i=1}^{k} p_i (a_i - a) \varphi\Big(\theta(x_i^* - a) + (1 - \theta)(x_i^* - a_i)\Big) \leqslant$$
$$\leqslant p \frac{a_k - \min\{0, a_0\}}{\sqrt{2\pi}} = L(V, V_p) \frac{a_k - \min\{0, a_0\}}{\sqrt{2\pi}}.$$

Окончательно

$$L(V,V_p)\geqslant \max\left\{1,\frac{\sqrt{2\pi}}{a_k-\min\{0,a_0\}}\right\}L(G,G_p)=C_1^{[1]}(a_k,a_0)L(G,G_p).$$

□

Рассмотрим следующее обобщение модели (2.33). Пусть имеется еще одна смесь данного типа, отличающаяся от (2.33) только весом, то есть (при этом $0\leqslant q\leqslant 1$)

$$G_q(x)=(1-q)\sum_{i=1}^{k}p_i\Phi(x-a_i)+q\Phi(x-a). \tag{2.43}$$

Для $G_q(x)$ дискретная случайная величина V_q имеет следующий вид:

$$V_q:\begin{array}{cccccc} a & a_1 & a_2 & \dots & a_k \\ q & p_1(1-q) & p_2(1-q) & \dots & p_k(1-q). \end{array} \tag{2.44}$$

Рассуждая как описано выше, получим, что $|p-q|\leqslant a_1-a$. В этом случае расстояние Леви $L(V_p,V_q)$ примет вид

$$L(V_p,V_q)=|p-q|. \tag{2.45}$$

Тогда справедлива следующая теорема.

Теорема 2.9. *В рамках модели добавления компоненты* (2.33) *при выполнении условий* (2.34) *и* (2.45) *расстояние Леви* $L(V_p,V_q)$ *между смешивающими распределениями* V_p *из соотношения* (2.35) *и* V_q *из соотношения* (2.44) *и расстояние Леви* $L(G_p,G_q)$ *между распределениями* $G_p(x)$ *из соотношения* (2.33) *и* $G_q(x)$ *из соотношения* (2.43) *связывают неравенства*

$$C_1^{[1]}(a_k,a_0)L(G_p,G_q)\leqslant L(V_p,V_q)\leqslant C_2^{[1]}(a_k,a_0)L^{1/2}(G_p,G_q),$$

где коэффициенты $C_j^{[1]}(a_k,a_0)$, $j=1,2$, *зависящие только от известных величин* a_k *и* a_0, *определяются формулами* (2.37) *и* (2.38).

Доказательство. Рассуждая аналогично доказательству теоремы 2.8, найдем оценки снизу для равномерного расстояния между функциями распределения $G_p(x)$ и $G_q(x)$. Имеем:

$$\rho(G_p,G_q)=\sup_x\left|(q-p)\sum_{i=1}^{k}p_i\Phi(x-a_i)+(p-q)\Phi(x-a)\right|=$$

$$= |p-q| \sup_x \left| \sum_{i=1}^{k} p_i \Phi(x-a_i) - \Phi(x-a) \right| \geqslant$$

$$\geqslant |p-q| \left| \sum_{i=1}^{k} p_i (\Phi(x-a_i) - \Phi(x-a)) \right| \geqslant$$

$$\geqslant L^2(V_p, V_q) \varphi\Big(a_k + |a_k| - \min\{0, a_0\}\Big).$$

Оценим максимум производной для функций G_p и G_q. Запишем выражения, например, для функции G_p (для функции G_q оценка получается аналогично). Имеем

$$\max_x G_p'(x) = \max_x \left((1-p) \sum_{i=1}^{k} p_i \varphi(x-a_i) + p\varphi(x-a) \right) \leqslant$$

$$\leqslant \frac{1-p}{\sqrt{2\pi}} \sum_{i=1}^{k} p_i + \frac{p}{\sqrt{2\pi}} = \frac{1}{\sqrt{2\pi}}.$$

Пользуясь правым неравенством в формуле (2.42), приходим к следующему результату:

$$L(V_p, V_q) \leqslant \varphi^{-1/2}\Big(a_k + |a_k| - \min\{0, a_0\}\Big) \left(1 + \frac{1}{\sqrt{2\pi}}\right)^{1/2} L^{1/2}(G_p, G_q) =$$

$$= C_2^{[1]}(a_k, a_0) L^{1/2}(G_p, G_q).$$

Оценка снизу для $L(V_p, V_q)$ может быть найдена из следующих соотношений:

$$L(G_p, G_q) \leqslant \rho(G_p, G_q) = |p-q| \sup_x \left| \sum_{i=1}^{k} p_i \Phi(x-a_i) - \Phi(x-a) \right| \leqslant$$

$$\leqslant |p-q| \sup_x \sum_{i=1}^{k} p_i |\Phi(x-a_i) - \Phi(x-a)| \leqslant$$

$$\leqslant |p-q| \sum_{i=1}^{k} p_i \sup_x |\Phi(x-a_i) - \Phi(x-a)| \leqslant |p-q| \sum_{i=1}^{k} p_i = L(V_p, V_q).$$

Аналогично доказательству теоремы 2.8 получим

$$L(V_p, V_q) \geqslant \max\left\{1, \frac{\sqrt{2\pi}}{a_k - \min\{0, a_0\}}\right\} L(G_p, G_q) = C_1^{[1]}(a_k, a_0) L(G_p, G_q),$$

которое завершает доказательство данной теоремы. $\square$

2.4.3 Модель расщепления компоненты

Модель расщепления компоненты формализуется следующим образом. Предполагается, что каждое из независимых наблюдений $\mathbf{X}_n = (X_1, \ldots, X_n)$ имеет распределение, представимое в виде

$$G_p(x) = \sum_{i=1}^{k-1} p_i \Phi(x - a_i) + (p_k - p)\Phi(x - a_k) + p\Phi(x - a), \tag{2.46}$$

где все величины $a_i \in \mathbb{R}$, $0 \leqslant p_i \leqslant 1$, $i = 1, \ldots, k$, считаются известными, a и p являются параметрами модели, при этом $0 \leqslant p \leqslant p_k$. Без ограничения общности для определенности будем считать, что выполнены соотношения

$$a_1 \leqslant a_2 \leqslant \ldots \leqslant a_{k-1} \leqslant a \leqslant a_k. \tag{2.47}$$

Для данной модели дискретная случайная величина V_p имеет вид

$$V_p : \begin{matrix} a_1 & a_2 & \ldots & a & a_k \\ p_1 & p_2 & \ldots & p & p_k - p. \end{matrix} \tag{2.48}$$

Воспользовавшись геометрической интерпретацией расстояния Леви, можно получить, что

$$L(V, V_p) = \min\{a_k - a, p\}. \tag{2.49}$$

В этой ситуации оба условия: $a \to a_k$ при фиксированном параметре p и $p \to 0$ при фиксированном a – влекут справедливость соотношения $L(V, V_p) \to 0$. Тогда справедлива следующая теорема.

Теорема 2.10. *В рамках модели расщепления компоненты* (2.46) *при выполнении условий* (2.47) *расстояние Леви* $L(V, V_p)$ *из соотношения* (2.49) *между смешивающими распределениями* V *из соотношения* (2.32) *и* V_p *из соотношения* (2.48) *и расстояние Леви* $L(G, G_p)$ *между истинным распределением* $G(x)$ *из соотношения* (2.31) *и приближающей смесью* $G_p(x)$ *из соотношения* (2.46) *связывают неравенства*

$$C_1^{[2]}(a_{k-1}, a_k) L(G, G_p) \leqslant L(V, V_p) \leqslant C_2^{[2]}(a_{k-1}, a_k) L^{1/2}(G, G_p),$$

где коэффициенты $C_j^{[2]}(a_{k-1}, a_k)$, $j = 1, 2$, *не зависят от величин* a, p *и*

имеют вид

$$C_1^{[2]}(a_{k-1}, a_k) = \frac{\sqrt{2\pi}}{\max\{1, a_k - a_{k-1}\}}, \tag{2.50}$$

$$C_2^{[2]}(a_{k-1}, a_k) = \varphi^{-1/2}\Big(a_k + |a_k| - \min\{0, a_{k-1}\}\Big)\left(1 + \frac{1}{\sqrt{2\pi}}\right)^{1/2}. \tag{2.51}$$

Доказательство. Запишем оценки снизу для равномерного расстояния между функциями распределения $G(x)$ и $G_p(x)$, воспользовавшись формулой Лагранжа, свойством монотонного убывания плотности стандартного нормального распределения $\varphi(x)$ от положительного аргумента и соотношениями (2.41), (2.47) и (2.49):

$$\rho(G, G_p) = \sup_x |G(x) - G_p(x)| =$$

$$= \sup_x \left| \sum_{i=1}^{k} p_i \Phi(x - a_i) - \sum_{i=1}^{k} p_i \Phi(x - a_i) + p\Phi(x - a_k) - p\Phi(x - a) \right| =$$

$$= p \sup_x |\Phi(x - a_k) - \Phi(x - a)| \geqslant p|\Phi(x_0 - a_k) - \Phi(x_0 - a)| =$$

$$= p|(a_k - a)\varphi(\theta(x_0 - a_k) + (1 - \theta)(x_0 - a))| \geqslant$$

$$\geqslant p(a_k - a)\varphi\Big(a_k + |a_k| - \min\{0, a_{k-1}\}\Big) \geqslant$$

$$\geqslant L^2(V, V_p)\varphi\Big(a_k + |a_k| - \min\{0, a_{k-1}\}\Big).$$

Чтобы оценить сверху $L(V, V_p)$ воспользуемся правым неравенством из соотношения (2.42) и найденной в доказательстве теоремы 2.8 оценкой для максимума производной, а также неравенствами (2.47). Имеем

$$L^2(V, V_p)\varphi\Big(a_k + |a_k| - \min\{0, a_{k-1}\}\Big) \leqslant \left(1 + \frac{1}{\sqrt{2\pi}}\right) L(G, G_p).$$

Откуда

$$L(V, V_p) \leqslant \varphi^{-1/2}\Big(a_k + |a_k| - \min\{0, a_{k-1}\}\Big)\left(1 + \frac{1}{\sqrt{2\pi}}\right)^{1/2} L^{1/2}(G, G_p) =$$

$$= C_2^{[2]}(a_{k-1}, a_k) L^{1/2}(G, G_p).$$

Выпишем оценку снизу для $L(V, V_p)$. С этой целью заметим, что

$$L(G, G_p) \leqslant \rho(G, G_p) = p \sup_x |\Phi(x - a_k) - \Phi(x - a)| =$$

$$= p \sup_x \Big(\Phi(x - a) - \Phi(x - a_k)\Big). \tag{2.52}$$

Найдем точки экстремума функции $\Phi(x-a)-\Phi(x-a_k)$ из условия

$$\varphi(x-a)-\varphi(x-a_k)=0.$$

Максимум достигается в точке

$$x^*=\frac{a+a_k}{2}.$$

Подставляя это значение в (2.52), получим (учитывая четность функции $\varphi(x)$)

$$p\sup_x\Big(\Phi(x-a)-\Phi(x-a_k)\Big)=p\Big(\Phi(x^*-a)-\Phi(x^*-a_k)\Big)=$$
$$=p(a_k-a)\varphi\Big(\theta(x^*-a)+(1-\theta)(x^*-a_k)\Big)=$$
$$=p(a_k-a)\varphi\Big((a_k-a)\Big|\theta-\frac{1}{2}\Big|\Big)\leqslant$$
$$\leqslant L(V,V_p)\max\{p,a_k-a\}\frac{1}{\sqrt{2\pi}}\leqslant L(V,V_p)\max\{1,a_k-a_{k-1}\}\frac{1}{\sqrt{2\pi}}.$$

Окончательно получаем неравенство

$$L(V,V_p)\geqslant\frac{\sqrt{2\pi}}{\max\{1,a_k-a_{k-1}\}}L(G,G_p)=C_1^{[2]}(a_{k-1},a_k)L(G,G_p),$$

которое завершает доказательство теоремы. □

Рассмотрим следующее обобщение модели (2.46). Пусть имеется еще одна смесь данного типа, отличающаяся от (2.46) только весом, то есть (при этом $0\leqslant q\leqslant p_k$)

$$G_q(x)=\sum_{i=1}^{k-1}p_i\Phi(x-a_i)+(p_k-q)\Phi(x-a_k)+q\Phi(x-a). \tag{2.53}$$

Для $G_q(x)$ дискретная случайная величина V_q имеет вид

$$V_q:\ \begin{matrix}a_1 & a_2 & \dots & a & a_k\\ p_1 & p_2 & \dots & q & p_k-q.\end{matrix} \tag{2.54}$$

Воспользовавшись геометрической интерпретацией расстояния Леви, можно получить, что

$$L(V_p,V_q)=\min\{a_k-a,|p-q|\}. \tag{2.55}$$

Тогда справедлива следующая теорема.

Теорема 2.11. *В рамках модели расщепления компоненты* (2.46) *при выполнении условий* (2.47) *расстояние Леви* $L(V_p, V_q)$ *из соотношения* (2.55) *между смешивающими распределениями* V_p *из соотношения* (2.48) *и* V_q *из соотношения* (2.54) *и расстояние Леви* $L(G_p, G_q)$ *между распределениями* $G_p(x)$ *из соотношения* (2.46) *и* $G_q(x)$ *из соотношения* (2.53) *связывают неравенства*

$$C_1^{[2]}(a_{k-1}, a_k)L(G_p, G_q) \leqslant L(V_p, V_q) \leqslant C_2^{[2]}(a_{k-1}, a_k)L^{1/2}(G_p, G_q),$$

где коэффициенты $C_j^{[2]}(a_{k-1}, a_k)$, $j = 1, 2$, *не зависят от величин* a, p *и определяются формулами* (2.50) *и* (2.51).

Доказательство. Рассуждая аналогично доказательству теоремы 2.10, найдем оценки снизу для равномерного расстояния между функциями распределения $G_p(x)$ и $G_q(x)$. Имеем

$$\rho(G_p, G_q) = \sup_x |G_p(x) - G_q(x)| =$$
$$= |p - q| \sup_x |\Phi(x - a_k) - \Phi(x - a)| \geqslant p|\Phi(x_0 - a_k) - \Phi(x_0 - a)| \geqslant$$
$$\geqslant L^2(V_p, V_q)\varphi\Big(a_k + |a_k| - \min\{0, a_{k-1}\}\Big).$$

Оценим максимум производной для функций G_p и G_q. Имеем

$$\max_x G_p'(x) = \max_x \Big(\sum_{i=1}^{k-1} p_i\varphi(x - a_i) + (p_k - p)\varphi(x - a_k) + p\varphi(x - a)\Big) \leqslant$$
$$\leqslant \frac{1}{\sqrt{2\pi}}\sum_{i=1}^{k-1} p_i + \frac{(p_k - p)}{\sqrt{2\pi}} + \frac{p}{\sqrt{2\pi}} = \frac{1}{\sqrt{2\pi}}.$$

Пользуясь правым неравенством в формуле (2.42), приходим к следующему результату:

$$L(V_p, V_q) \leqslant \varphi^{-1/2}\Big(a_k + |a_k| - \min\{0, a_{k-1}\}\Big)\left(1 + \frac{1}{\sqrt{2\pi}}\right)^{1/2} L^{1/2}(G_p, G_q) =$$
$$= C_2^{[2]}(a_{k-1}, a_k)L^{1/2}(G_p, G_q).$$

Оценка снизу для $L(V_p, V_q)$ может быть найдена из следующих соотношений:

$$L(G_p, G_q) \leqslant \rho(G_p, G_q) = |p - q| \sup_x |\Phi(x - a_k) - \Phi(x - a)|.$$

Повторяя рассуждения из доказательства теоремы 2.10, получаем неравенство

$$L(V_p, V_q) \geqslant \frac{\sqrt{2\pi}}{\max\{1, a_k - a_{k-1}\}} L(G, G_p) = C_1^{[2]}(a_{k-1}, a_k) L(G_p, G_q),$$

которое завершает доказательство данной теоремы. □

2.5 Устойчивость дисперсионно-сдвиговых смесей нормальных законов относительно смешивающего распределения

Рассмотрим важный класс сдвиг-масштабных смесей (1.4), называемых *дисперсионно-сдвиговыми смесями нормальных законов* [143]:

$$\Phi_{\alpha,\sigma,F_A}(x) = \int_0^\infty \Phi\left(\frac{x - \alpha u}{\sigma\sqrt{u}}\right) dF_A(u), \quad \alpha \in \mathbb{R}, \quad \sigma > 0, \tag{2.56}$$

где $F_A(u)$ – функция распределения положительной с вероятностью единица случайной величины. Таким образом, рассматриваются степенные смеси (в этом легко убедиться, записав характеристическую функцию случайной величины X с нормальным распределением в случайной почти наверное неотрицательной степени U, распределением которой является $F_A(u)$). В смеси (2.56) смешивание происходит одновременно по параметрам сдвига и масштаба, связанным между собой. К данному классу относятся широко используемые в анализе данных обобщенные гиперболические, дисперсионные гамма- и NIG-распределения.

Лемма 2.1. *Если $F(x)$ и $G(x)$ – две дифференцируемые функции распределения, то $\rho(F, G)$ реализуется (достигается $\sup_x |F(x) - G(x)|$) в одной из точек, в которых $F'(x) = G'(x)$.*

Доказательство. Очевидно, что в рассматриваемой ситуации

$$\rho(F, G) = \sup_x |F(x) - G(x)| = \\ = \max\left\{ \max_x \left[F(x) - G(x)\right], \max_x \left[G(x) - F(x)\right] \right\},$$

а экстремум каждого из выражений в фигурных скобках в правой части достигается в точке, где производная соответствующего выражения

равна нулю, что эквивалентно равенству производных функций распределения F и G, то есть совпадению соответствующих плотностей. $\square$

Теорема 2.12. *Предположим, что F_A и F_B – функции распределения с точками роста, расположенными на неотрицательной полуоси, и по крайней мере F_A имеет плотность, ограниченную некоторым числом $0 < a < \infty$. Тогда*

$$L(\Phi_{\alpha,\sigma,F_A}, \Phi_{\alpha,\sigma,F_B}) \leqslant 2(1+a)L(F_A, F_B).$$

Доказательство. Пусть x_0 – точка, в которой реализуется $\rho(\Phi_{\alpha,\sigma,F_A}, \Phi_{\alpha,\sigma,F_A})$. Тогда

$$\rho(\Phi_{\alpha,\sigma,F_A}, \Phi_{\alpha,\sigma,F_B}) = \left| \int_0^\infty \Phi\left(\frac{x_0-\alpha u}{\sigma\sqrt{u}}\right) dF_A(u) - \int_0^\infty \Phi\left(\frac{x_0-\alpha u}{\sigma\sqrt{u}}\right) dF_B(u) \right| =$$
$$= \left| \int_0^\infty \Phi\left(\frac{x_0-\alpha u}{\sigma\sqrt{u}}\right) d\left[F_A(u) - F_B(u)\right] \right|. \qquad (2.57)$$

Интегрируя по частям выражение, стоящее справа в формуле (2.57), получим

$$\left| \Phi\left(\frac{x_0-\alpha u}{\sigma\sqrt{u}}\right) [F_A(u) - F_B(u)] \Big|_{u=0}^\infty - \int_0^\infty [F_A(u) - F_B(u)]\, d_u\Phi\left(\frac{x_0-\alpha u}{\sigma\sqrt{u}}\right) \right| =$$
$$= \left| \int_0^\infty [F_A(u) - F_B(u)]\, d_u\Phi\left(\frac{x_0-\alpha u}{\sigma\sqrt{u}}\right) \right| \leqslant \rho(F_A, F_B) \cdot \int_0^\infty \left| d_u\Phi\left(\frac{x_0-\alpha u}{\sigma\sqrt{u}}\right) \right|.$$

Оценка в правой части основывается на результате леммы 2.1. Рассмотрим поведение аргумента функции $\Phi\left((x_0-\alpha u)/(\sigma\sqrt{u})\right)$ в зависимости от u. Имеем:

$$\frac{d}{du}\left(\frac{x_0-\alpha u}{\sigma\sqrt{u}}\right) = \frac{d}{du}\left(\frac{x_0}{\sigma\sqrt{u}} - \alpha\sqrt{u}\right) =$$
$$= -\frac{x_0}{2\sigma u^{3/2}} - \frac{\alpha}{2\sqrt{u}} = -\frac{1}{2\sqrt{u}}\left(\frac{x_0}{\sigma u} + \alpha\right),$$

Производная меняет знак не более чем в одной точке $u_0 = -x_0(\alpha\sigma)^{-1}$.

При этом функция

$$g(u) = \frac{x_0-\alpha u}{\sigma\sqrt{u}}$$

монотонна при $u \neq u_0$. Так как $\Phi(\cdot)$ – монотонная функция распределения, то и $\Phi\left((x_0 - \alpha u)(\sigma\sqrt{u})^{-1}\right)$ как функция переменной u обладает аналогичным свойством на тех же множествах. Поэтому

$$\int_0^{\infty} \left| d_u \Phi\left(\frac{x_0 - \alpha u}{\sigma\sqrt{u}}\right)\right| \leqslant \int_0^{u_0} \left| d_u \Phi\left(\frac{x_0 - \alpha u}{\sigma\sqrt{u}}\right)\right| + \int_{u_0}^{\infty} \left| d_u \Phi\left(\frac{x_0 - \alpha u}{\sigma\sqrt{u}}\right)\right| \leqslant 2,$$

так как суммарное приращение любой функции распределения не превосходит единицы. Таким образом,

$$\rho(\Phi_{\alpha,\sigma,F_A}, \Phi_{\alpha,\sigma,F_B}) \leqslant 2\rho(F_A, F_B),$$

откуда с учетом соотношений (2.42) (причем в правом неравенстве для общего случая можно использовать супремум вместо максимума) получаем неравенства

$$L(\Phi_{\alpha,\sigma,F_A}, \Phi_{\alpha,\sigma,F_B}) \leqslant \rho(\Phi_{\alpha,\sigma,F_A}, \Phi_{\alpha,\sigma,F_B}) \leqslant 2\rho(F_A, F_B) \leqslant 2(1+a)L(F_A, F_B),$$

завершающие доказательство теоремы. □

Таким образом, близость смешивающих распределений в смысле расстояния Леви необходимо влечет и близость соответствующих смесей. Использование метрики Леви в данном случае объясняется тем обстоятельством, что одно из двух смешивающих распределений в приведенной выше теореме может быть дискретным. Данный результат об устойчивости дисперсионно-сдвиговых смесей вида (2.56) относительно возмущений смешивающего распределения F_A важен для обоснования вычислительного метода разделения подобных смесей (см., например, статью [96]).

2.6 Зашумление данных конечными смесями нормальных и гамма-распределений для случая округленных наблюдений

Одной из проблем, возникающих при статистическом анализе современных сложных стохастических систем, является ограниченная точность данных при их регистрации и хранении. На практике все данные представлены в округленном виде. Это связано как с тем, что аппаратура, считывающая данные, имеет ограниченную разрешающую способность, так и с тем, что при хранении больших и сверхбольших объемов

данных приходится производить их группировку, которая тоже может рассматриваться как своеобразное округление. Стандартные математические методы решения статистических задач, которые не учитывают округленную природу данных, приводят к значительным погрешностям. Это недопустимо в современных условиях. Таким образом, актуальным является разработка методов, которые ориентированы на анализ именно округленных данных. Среди задач обработки округленных данных особое место занимают такие, в которых возможно то или иное воздействие на измеряемую величину до ее регистрации. Наиболее естественным и просто осуществляемым таким воздействием является наложение аддитивного случайного шума. В этом случае появляется возможность уменьшения ошибок округления за счет возрастания случайных ошибок измерения.

Для решения данной проблемы развивались различные подходы, в том числе на основе смешанных моделей (см., например, статью [422], в которой различные компоненты используются для представления уровней округления). В работе [140] приводятся результаты для моделей авторегрессии и скользящего среднего для округленных данных, а в статье [437] эти результаты развиваются и исследуются их асимптотические свойства. В статье [438] исследован метод оценивания параметров конечных смесей вероятностных распределений (в том числе и многомерных) на основе использования EM-алгоритма с целью получения состоятельных и асимптотически нормальных оценок.

В данном разделе изучаются теоретические основы для устранения ошибок в модели округления данных [114, 115, 411], для которой будут получены оценки для неизвестного математического ожидания наблюдений в предположении, что исходные данные зашумлены с помощью случайных величин, имеющих распределения типа конечных смесей нормальных и гамма-законов. Такой подход позволяет учесть большее число случайных факторов, влияющих на величину «дополнительной» ошибки. Также будут построены доверительные интервалы для неизвестного математического ожидания.

2.6.1 Предположения и базовые отношения

Для сокращения формулировок теорем в следующих разделах сделаем ряд предположений, на которые будем ссылаться в дальнейшем:
A: $X_1, X_2, \ldots$ – независимые одинаково распределенные случайные ве-

личины с неизвестным математическим ожиданием $E_X < +\infty$.
B: $\varepsilon_1, \varepsilon_2, \ldots$ – независимые одинаково распределенные случайные величины с математическим ожиданием $E_\varepsilon < +\infty$.
C: $X_1, X_2, \ldots$ и $\varepsilon_1, \varepsilon_2, \ldots$ являются независимыми.
D: $Y_j = \left[X_j + \varepsilon_j + \frac{1}{2}\right]$ для всех $j = 1, 2, \ldots$ представляют собой округление значения суммы случайных величин $X_j + \varepsilon_j$ до ближайшего целого сверху (при этом запись $[\cdot]$ соответствует целой части выражения).

В рамках данных предположений в статье будут рассмотрены вопросы качества приближения неизвестного математического ожидания E_X для исходных данных в ситуации, когда наблюдения для анализа получены с аддитивной ошибкой с известными распределениями (см. предположение B) и дополнительно округляются до ближайшего целого (см. предположение D).

Заметим, что вследствие усиленного закона больших чисел справедливы следующие выражения:

$$\begin{aligned} \frac{1}{n}\sum_{j=1}^{n} Y_j \xrightarrow[n\to\infty]{\text{п. н.}} E_Y \equiv \mathbb{E}\left[X_1 + \varepsilon_1 + \frac{1}{2}\right] = \\ = \mathbb{E}\left(X_j + \varepsilon_j + \frac{1}{2}\right) - \mathbb{E}\left\{X_j + \varepsilon_j + \frac{1}{2}\right\} = \\ = E_X + E_\varepsilon + \frac{1}{2} - \mathbb{E}\left\{X_j + \varepsilon_j + \frac{1}{2}\right\}. \end{aligned} \tag{2.58}$$

Запись $\{\cdot\}$ в формуле (2.58) соответствует дробной части выражения, а `п.н.` обозначает сходимость в смысле почти наверное.

Для доказательства результатов в дальнейшем потребуется следующее представления для дробной части абсолютно непрерывной случайной величины Z с абсолютно интегрируемой характеристической функцией $\varphi_Z(t)$ (см., например, лемму 4 в работе [411]):

$$\mathbb{E}\{Z\} = \frac{1}{2} - \sum_{n=1}^{\infty} \frac{\mathrm{Im}(\varphi_Z(2\pi n))}{\pi n}. \tag{2.59}$$

Через $\mathrm{Im}(\cdot)$ в формуле (2.59) обозначена мнимая часть соответствующей функции.

При построении доверительных интервалов в дальнейшем будет использована следующая оценка, справедливая для любой случайной величины Z:

$$\mathbb{D}[Z] \leqslant \left(\sqrt{\mathbb{D}Z} + \frac{1}{2}\right)^2. \tag{2.60}$$

Она может быть проверена непосредственно с учетом представления $\mathbb{D}[Z] = \mathbb{D}\left(Z - \{Z\}\right)$, неравенства Коши–Буняковского для ковариации и соотношения $\mathbb{D}\{Z\} \leqslant \frac{1}{4}$, справедливого для любой случайной величины Z (см., например, статью [411]). Отметим, что данная оценка является более точной по сравнению с использованным для аналогичных целей в работе [411] соотношением $\mathbb{D}[Z] \leqslant 2\mathbb{D}Z + \frac{1}{2}$. Действительно,

$$2\mathbb{D}Z + \frac{1}{2} - \left(\sqrt{\mathbb{D}Z} + \frac{1}{2}\right)^2 = \left(\sqrt{\mathbb{D}Z} - \frac{1}{2}\right)^2 \geqslant 0,$$

причем для всех $\sqrt{\mathbb{D}Z} \neq \frac{1}{2}$ данное неравенство является строгим.

2.6.2 Конечные смеси нормальных законов

Для случайной величины X, имеющей распределение типа конечной смеси нормальных законов (1.7), параметры которой удовлетворяют соотношениям (1.8), характеристическая функция имеет вид

$$\varphi_X(t) = \int\limits_{-\infty}^{+\infty} e^{itx} \left\{ \sum_{j=1}^{k} \frac{p_j}{\sigma_j\sqrt{2\pi}}\, e^{-\dfrac{(x-a_j)^2}{2\sigma_j^2}} \right\} dx = \sum_{j=1}^{k} p_j e^{ita_j - \frac{1}{2}\sigma_j^2 t^2}. \quad (2.61)$$

Абсолютная интегрируемость $\varphi_X(t)$ вытекает из свойств характеристической функции нормального распределения. Заметим, что в точке $t = 2\pi n$ выражение (2.61) принимает следующий вид:

$$\varphi_X(2\pi n) = \sum_{j=1}^{k} p_j e^{-2\pi^2\sigma_j^2 n^2}. \quad (2.62)$$

Рассмотрим вопрос точности оценивания неизвестного математического ожидания E_X при добавлении зашумления.

Теорема 2.13. *Пусть выполнены предположения A–D, причем случайные величины ε_j, $j = 1, 2, \ldots$, имеют распределение типа конечной k-компонентной смеси нормальных законов вида (1.7) с параметрами* $\mathbf{a}$, $\boldsymbol{\sigma}$ *и* $\mathbf{p}$. *Тогда*

$$|E_Y - E_X| \leqslant A + \frac{1}{\pi}\left(1 + \frac{1}{4\pi^2\sigma^2}\right) e^{-2\pi^2\sigma^2}, \quad (2.63)$$

где $A = \max(|a_1|, \ldots, |a_k|)$, $\sigma = \min(\sigma_1, \ldots, \sigma_k)$.

Доказательство. С учетом представлений (2.58), (2.59), (2.62), ограниченности модуля характеристической функции, а также независимости случайных величин X_j и ε_j, имеем:

$$|E_Y - E_X| = \left| E_\varepsilon + \frac{1}{2} - \mathbb{E}\left\{ X_j + \varepsilon_j + \frac{1}{2} \right\} \right| =$$

$$= \left| E_\varepsilon + \sum_{n=1}^{\infty} \frac{\operatorname{Im}\left(\varphi_{X_j}(2\pi n) \varphi_{\varepsilon_j}(2\pi n) \varphi_{1/2}(2\pi n) \right)}{\pi n} \right| =$$

$$= \left| E_\varepsilon + \sum_{n=1}^{\infty} \frac{\operatorname{Im}\left(\varphi_{X_j}(2\pi n) \cdot \sum\limits_{j=1}^{k} p_j e^{-2\pi^2 \sigma_j^2 n^2} \cdot e^{\pi n} \right)}{\pi n} \right| =$$

$$= \left| E_\varepsilon + \sum_{n=1}^{\infty} \frac{(-1)^n \sum\limits_{j=1}^{k} p_j e^{-2\pi^2 \sigma_j^2 n^2} \operatorname{Im}\left(\varphi_{X_j}(2\pi n) \right)}{\pi n} \right| \leqslant$$

$$\leqslant |E_\varepsilon| + \left| \sum_{j=1}^{k} p_j \sum_{n=1}^{\infty} \frac{1}{\pi n} e^{-2\pi^2 \sigma_j^2 n^2} \right| \leqslant$$

$$\leqslant \max\left(|a_1|, \ldots, |a_k| \right) + \sum_{j=1}^{k} \frac{p_j}{\pi} \left(1 + \frac{1}{4\pi^2 \sigma_j^2} \right) e^{-2\pi^2 \sigma_j^2} \leqslant$$

$$\leqslant A + \frac{1}{\pi} \left(1 + \frac{1}{4\pi^2 \sigma^2} \right) e^{-2\pi^2 \sigma^2}.$$

Справедливость использованной оценки

$$\sum_{n=1}^{\infty} \frac{e^{-2\pi^2 \sigma_j^2 n^2}}{n} \leqslant \left(1 + \frac{1}{4\pi^2 \sigma_j^2} \right) e^{-2\pi^2 \sigma_j^2}$$

может быть проверена непосредственно (например, см. доказательство теоремы 6 в статье [411]). □

Замечание 2.2. В случае, если зашумление производится нормально распределенными случайными величинами с нулевыми средними (то есть в формуле (2.63) необходимо считать $A = 0$, $k = 1$), теорема 2.13 совпадает с результатом, полученным в работе [411].

Рассмотрим вопросы построения доверительного интервала для неизвестного математического ожидания E_X в предположении, что случайные величины X_j не содержат ошибок измерения, а все погрешности учтены исключительно в зашумляющих элементах ε_j.

Теорема 2.14. *Пусть выполнены предположения A–D, причем случайные величины ε_j, $j = 1, 2, \ldots$, имеют распределение типа конечной k-компонентной смеси нормальных законов вида (1.7) с параметрами* $\mathbf{a}$, $\boldsymbol{\sigma}$ *и* $\mathbf{p}$, *а случайные величины* $X_j \overset{\text{п. н.}}{=} E_X$, $j = 1, 2, \ldots$ *Тогда доверительный интервал для* E_X *уровня* $1 - \alpha$, $0 < \alpha < 1$, *имеет вид*

$$\left[\hat{E}_X - f(\mathbf{a}, \boldsymbol{\sigma}, \alpha, n), \hat{E}_X + f(\mathbf{a}, \boldsymbol{\sigma}, \alpha, n)\right], \tag{2.64}$$

где

$$\hat{E}_X = \frac{1}{n}\sum_{j=1}^{n}\left[E_X + \varepsilon_j + \frac{1}{2}\right], \tag{2.65}$$

$$f(\mathbf{a}, \boldsymbol{\sigma}, \alpha, n) = \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}}\left(\sqrt{A^2 + \Sigma^2} + \frac{1}{2}\right) + \\ + A + \frac{1}{\pi}\left(1 + \frac{1}{4\pi^2\sigma^2}\right)e^{-2\pi^2\sigma^2}, \tag{2.66}$$

$z_{1-\frac{\alpha}{2}}$ – $\left(1 - \frac{\alpha}{2}\right)$-*квантиль стандартного нормального распределения,* $A = \max(|a_1|, \ldots, |a_k|)$, $\Sigma = \max(\sigma_1, \ldots, \sigma_k)$, $\sigma = \min(\sigma_1, \ldots, \sigma_k)$.

Доказательство. Из центральной предельной теоремы с учетом условия A следует, что величина $\hat{E}_X$ (2.65) асимптотически нормальна со следующими математическим ожиданием и дисперсией:

$$E = \mathbb{E}\left[E_X + \varepsilon_1 + \frac{1}{2}\right], \quad \frac{1}{n}D = \frac{1}{n}\mathbb{D}\left[E_X + \varepsilon_1 + \frac{1}{2}\right]. \tag{2.67}$$

Воспользовавшись оценкой (2.60), получим

$$D \leqslant \left(\sqrt{\mathbb{D}\left(E_X + \varepsilon_1 + \frac{1}{2}\right)} + \frac{1}{2}\right)^2 = \left(\sqrt{\mathbb{D}\varepsilon_1} + \frac{1}{2}\right)^2 = \\ = \left(\sqrt{\sum_{j=1}^{k} p_j\left(\left(a_j - \sum_{t=1}^{k} p_t a_t\right)^2 + \sigma_j^2\right)} + \frac{1}{2}\right)^2 \leqslant \left(\sqrt{A^2 + \Sigma^2} + \frac{1}{2}\right)^2.$$

Тогда доверительный интервал уровня $1 - \alpha$ для математического ожидания E имеет вид

$$\mathbb{P}\left(\left|\hat{E}_X - E\right| \leqslant \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}}\left(\sqrt{A^2 + \Sigma^2} + \frac{1}{2}\right)\right) \geqslant 1 - \alpha.$$

Откуда следует справедливость соотношения (2.64) с учетом очевидного неравенства

$$\left|\hat{E}_X - E_X\right| \leqslant \left|\hat{E}_X - E\right| + |E - E_X|$$

и оценки (2.63) из теоремы 2.13. □

2.6.3 Конечные смеси гамма-распределений

Для случайной величины X, имеющей распределение типа конечной смеси гамма-распределений с параметрами $\mathbf{r} = (r_1, \ldots, r_k)$, $r_j > 0$, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k)$, $\lambda_j > 0$, $\mathbf{p} = (p_1, \ldots, p_k)$, $p_j \geqslant 0$, $\sum\limits_{j=1}^{k} p_j = 1$, плотность которого задается выражением

$$f_X(x) = \sum_{j=1}^{k} p_j \frac{\lambda_j^{r_j} e^{-\lambda_j x}}{\Gamma(r_j)} x^{r_j - 1}, \tag{2.68}$$

характеристическая функция задается следующим выражением:

$$\varphi_X(t) = \int\limits_{-\infty}^{+\infty} e^{itx} f_X(x)\, dx = \sum_{j=1}^{k} p_j \left(1 - \frac{it}{\lambda_j}\right)^{-r_j}. \tag{2.69}$$

Отметим, что подобные модели зашумления разумно использовать в случае, если известно, что данные сосредоточены на положительной полуоси, например, при анализе различных информационных потоков (в частности, см. работу [229]).

Проверим абсолютную интегрируемость функции $\varphi_X(t)$ (2.69). Имеем:

$$\int\limits_{-\infty}^{+\infty} |\varphi_X(t)|\, dt \leqslant \sum_{j=1}^{k} p_j \int\limits_{-\infty}^{+\infty} \left|\left(1 - \frac{it}{\lambda_j}\right)^{-r_j}\right| dt =$$

$$= \sum_{j=1}^{k} p_j \int\limits_{-\infty}^{+\infty} \left|\left(\frac{\lambda_j(\lambda_j + it)}{\lambda_j^2 + t^2}\right)^{r_j}\right| dt \leqslant \sum_{j=1}^{k} p_j \lambda_j \int\limits_{-\infty}^{+\infty} \left(1 + y^2\right)^{-\frac{r_j}{2}} dy.$$

Подынтегральное выражение при $r_j \geqslant 2$ может быть оценено сверху функцией $\dfrac{1}{1+y^2}$, при этом соответствующий интеграл равен π, что влечет абсолютную интегрируемость характеристической функции для конечной смеси гамма-распределений. Поэтому в дальнейшем будем предполагать, что $r_j \geqslant 2$ для всех возможных значений $j = 1, 2, \ldots$

Рассмотрим вопрос точности оценивания неизвестного математического ожидания $E_X > 0$ при добавлении зашумления.

Теорема 2.15. *Пусть выполнены предположения A–D, причем случайные величины ε_j, $j = 1, 2, \ldots$, имеют распределение типа конечной k-компонентной смеси гамма-распределений вида (2.68) с параметрами* $\mathbf{r}$, $\boldsymbol{\lambda}$ *и* $\mathbf{p}$. *Тогда*

$$|E_Y - E_X| \leqslant \frac{R}{\lambda} + \frac{\Lambda^R}{2^r \pi^{r+1}} \left(1 + \frac{1}{r}\right), \tag{2.70}$$

где $r = \min(r_1, \ldots, r_k)$, $R = \max(r_1, \ldots, r_k)$, $\lambda = \max(\lambda_1, \ldots, \lambda_k)$, $\Lambda = \max(\lambda_1, \ldots, \lambda_k)$.

Доказательство. С учетом представлений (2.58), (2.59), ограниченности модуля характеристической функции, перехода от тригонометрической к показательной записи комплексных чисел, а также независимости случайных величин X_j и ε_j имеем:

$$|E_Y - E_X| \leqslant |E_\varepsilon| + \left| \sum_{n=1}^{\infty} \frac{(-1)^n \mathrm{Im}\left(\sum\limits_{j=1}^{k} p_j \varphi_{X_j}(2\pi n) \left(1 - i\frac{2\pi n}{\lambda_j}\right)^{-r_j} \right)}{\pi n} \right| =$$

$$= |E_\varepsilon| + \left| \sum_{n=1}^{\infty} \frac{(-1)^n \mathrm{Im}\left(\sum\limits_{j=1}^{k} p_j \left(1 + \frac{4\pi^2 n^2}{\lambda_j^2}\right)^{-\frac{r_j}{2}} \varphi_{X_j}(2\pi n)\, e^{-ir_j \mathrm{arctg}\frac{t}{\lambda_j}} \right)}{\pi n} \right| \leqslant$$

$$\leqslant |E_\varepsilon| + \sum_{j=1}^{k} p_j \sum_{n=1}^{\infty} \frac{1}{\pi n} \left(1 + \frac{4\pi^2 n^2}{\lambda_j^2}\right)^{-\frac{r_j}{2}} \leqslant \frac{R}{\lambda} + \sum_{j=1}^{k} p_j \sum_{n=1}^{\infty} \left(\frac{1}{\pi n} \cdot \frac{\lambda_j^{r_j}}{(2\pi)^{r_j} n^{r_j}} \right) \leqslant$$

$$\leqslant \frac{R}{\lambda} + \sum_{j=1}^{k} p_j \frac{\lambda_j^{r_j}}{2^{r_j} \pi^{r_j+1}} \left(1 + \int_1^{\infty} \frac{1}{x^{r_j+1}}\, dx \right) \leqslant \frac{R}{\lambda} + \frac{\Lambda^R}{2^r \pi^{r+1}} \left(1 + \frac{1}{r}\right).$$

При переходе от суммы к интегралу используется факт убывания функции как переменной n (или x). □

Замечание 2.3. Теорема 2.15 описывает соответствующий результат для гамма-распределенных зашумляющих случайных величин, если положить $k = 1$ в выражении (2.70). При этом, очевидно, $r \equiv R$ и $\lambda \equiv \Lambda$.

Рассмотрим вопросы построения доверительного интервала для неизвестного математического ожидания $E_X > 0$ в предположении, что случайные величины X_j не содержат ошибок измерения, а все погрешности учтены исключительно в зашумляющих элементах ε_j.

Теорема 2.16. *Пусть выполнены предположения A–D, причем случайные величины* ε_j, $j = 1, 2, \ldots$, *имеют распределение типа конечной* k*-компонентной смеси гамма-распределений вида* (2.68) *с параметрами* $\mathbf{r}$, $\boldsymbol{\lambda}$ *и* $\mathbf{p}$, *а случайные величины* $X_j \overset{\text{п. н.}}{=} E_X$, $j = 1, 2, \ldots$ *Тогда доверительный интервал для* E_X *уровня* $1 - \alpha$, $0 < \alpha < 1$, *имеет вид*

$$\left[\hat{E}_X - f(\mathbf{r}, \boldsymbol{\lambda}, \alpha, n), \hat{E}_X + f(\mathbf{r}, \boldsymbol{\lambda}, \alpha, n)\right], \tag{2.71}$$

где

$$\hat{E}_X = \frac{1}{n}\sum_{j=1}^{n}\left[E_X + \varepsilon_j + \frac{1}{2}\right], \tag{2.72}$$

$$f(\mathbf{r}, \boldsymbol{\lambda}, \alpha, n) = \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}}\left(\sqrt{\frac{R(R+1)}{\lambda^2} - \frac{r^2}{\Lambda^2}} + \frac{1}{2}\right) + \frac{R}{\lambda} + \frac{\Lambda^R}{2^r\pi^{r+1}}\left(1 + \frac{1}{r}\right),$$

$z_{1-\frac{\alpha}{2}}$ – $\left(1 - \frac{\alpha}{2}\right)$*-квантиль стандартного нормального распределения,* $r = \min(r_1, \ldots, r_k)$, $R = \max(r_1, \ldots, r_k)$, $\lambda = \max(\lambda_1, \ldots, \lambda_k)$, $\Lambda = \max(\lambda_1, \ldots, \lambda_k)$.

Доказательство. Из центральной предельной теоремы с учетом условия A следует, что величина $\hat{E}_X$ (2.72) асимптотически нормальна с математическим ожиданием E и дисперсией $\frac{1}{n}D$ (2.67) (при этом в данном случае распределение величин ε_1 определяется формулой (2.68)). Пользуясь определением и свойствами гамма-функции, а также оценкой (2.60) получим:

$$D_Y \leqslant \left(\sqrt{\sum_{j=1}^{k} p_j \frac{\lambda_j^{r_j}}{\Gamma(r_j)} \int_0^{+\infty} e^{\lambda_j x} x^{r_j+1}\, dx} + \frac{1}{2}\right)^2 =$$

$$= \left(\sqrt{\sum_{j=1}^{k} p_j \frac{r_j(r_j+1)}{\lambda_j^2} - \left(\sum_{j=1}^{k} p_j \frac{r_j}{\lambda_j}\right)^2} + \frac{1}{2}\right)^2 \leqslant \left(\sqrt{\frac{R(R+1)}{\lambda^2} - \frac{r^2}{\Lambda^2}} + \frac{1}{2}\right)^2.$$

Аналогично доказательству теоремы 2.14 с учетом оценки (2.70) отсюда следует справедливость соотношения (2.71).

□

Глава 3

Методы анализа данных на основе скользящего разделения смесей

В данной главе разработаны новые алгоритмы анализа данных, в основу которых положен метод скользящего разделения смесей. В частности, рассмотрены подходы к обработке наблюдений, из которых нужно удалить шумовую составляющую, и в то же время предложен метод повышения точности аппроксимации с помощью конечных нормальных смесей с помощью дополнительного зашумления наблюдений. Изучаются двухсторонние подходы к детектированию различных событий в данных. Предложен метод автоматического выделения непрерывных компонент на основе жадного алгоритма в комбинации с методами машинного обучения, составляющего основу метода статистического оценивания распределений случайных коэффициентов в уравнении Ланжевена (см. параграф 2.1). Примеры применения этих методов будут рассмотрены далее в разделах 5.2 и 6.5.

3.1 Матричные представления моментов конечных нормальных смесей

Пусть случайная величина Z_n имеет функцию распределения $F(x, k(n), \mathbf{a}_n, \boldsymbol{\sigma}_n, \mathbf{p}_n)$ (1.7), параметры $\mathbf{a}_n = \{a_i(n)\}$, $\boldsymbol{\sigma}_n = \{\sigma_i(n)\}$ и $\mathbf{p}_n = \{p_i(n)\}$ которой удовлетворяют соотношениям (1.8), а n здесь обозначает время (номер окна) для метода скользящего разделения смесей и подчеркивает тот факт, что распределение изменяется для каждого по-

ложения окна (возможно, весьма существенным образом). В частности, параметр $k(n)$, описывающий число компонент в смеси вида (1.7), при аппроксимации реальных данных может принимать различные значения с течением времени (см., например, статьи [64, 65]). Это обстоятельство значительно затрудняет прогнозирование, так как появление новых компонент в процессе зачастую объясняется новыми факторами, которые отсутствовали при построении первоначальной модели и, следовательно, не могли быть учтены. Поэтому необходимо перейти к рассмотрению некоторых «интегральных» характеристик, которые можно рассчитать для любого распределения вида (1.7), независимо от конкретных значений параметров. В качестве таких величин можно использовать моменты различных порядков, явный вид которых для конечных нормальных сдвиг-масштабных смесей будет получен в данном разделе.

Теорема 3.1. *Моменты случайной величины Z_n с распределением $F(x, k(n), \mathbf{a}_n, \boldsymbol{\sigma}_n, \mathbf{p}_n)$, получаемые в процессе последовательных шагов в методе скользящего разделения смесей, имеют вид:*

– *математическое ожидание:*

$$\mathbb{E}Z_n = \sum_{i=1}^{k(n)} p_i(n) a_i(n); \tag{3.1}$$

– *дисперсия:*

$$\mathbb{D}Z_n = \sum_{i=1}^{k(n)} p_i(n) \left(a_i(n) - \sum_{i=1}^{k(n)} p_i(n) a_i(n) \right)^2 + \sum_{i=1}^{k(n)} p_i(n) \sigma_i^2(n); \tag{3.2}$$

– *коэффициент асимметрии:*

$$\begin{aligned} \gamma_{Z_n} = \Bigg[& \sum_{i=1}^{k(n)} p_i(n) \left(a_i^3(n) + 3 a_i \sigma_i^2(n) \right) - \left(\sum_{i=1}^{k(n)} p_i(n) a_i(n) \right) \times \\ & \times \left(3 \sum_{i=1}^{k(n)} p_i(n) \left(a_i(n) - \sum_{i=1}^{k(n)} p_i(n) a_i(n) \right)^2 + \right. \\ & \left. + 3 \sum_{i=1}^{k(n)} p_i(n) \sigma_i^2(n) - \left(\sum_{i=1}^{k(n)} p_i(n) a_i(n) \right)^2 \right) \Bigg] \times \\ & \times \left[\sum_{i=1}^{k(n)} p_i(n) \left(a_i(n) - \sum_{i=1}^{k(n)} p_i(n) a_i(n) \right)^2 + \sum_{i=1}^{k(n)} p_i(n) \sigma_i^2(n) \right]^{-3/2}; \end{aligned} \tag{3.3}$$

– *коэффициент эксцесса:*

$$\kappa_{Z_n} = \Bigg[\sum_{i=1}^{k(n)} p_i(n)\left(a_i^4(n) + 6a_i^2\sigma_i^2(n) + 3\sigma_i^4(n)\right) -$$
$$-3\Bigg(\sum_{i=1}^{k(n)} p_i(n)a_i(n)\Bigg)^4 - 4\Bigg(\sum_{i=1}^{k(n)} p_i(n)a_i(n)\Bigg)\Bigg(\sum_{i=1}^{k(n)} p_i(n)\Big(a_i^3(n) + 3a_i\sigma_i^2(n)\Big)\Bigg) +$$
$$+6\Bigg(\sum_{i=1}^{k(n)} p_i(n)a_i(n)\Bigg)^2\Bigg(\sum_{i=1}^{k(n)} p_i(n)\Big(a_i^2(n) + \sigma_i^2(n)\Big)\Bigg)\Bigg]\times$$
$$\times\Bigg[\sum_{i=1}^{k(n)} p_i(n)\Big(a_i(n) - \sum_{i=1}^{k(n)} p_i(n)a_i(n)\Big)^2 + \sum_{i=1}^{k(n)} p_i(n)\sigma_i^2(n)\Bigg]^{-2} - 3. \qquad (3.4)$$

Доказательство. Известно, что для начальных моментов случайной величины X с нормальным распределением с параметрами a и σ^2 (то есть $X \sim N(a, \sigma^2)$) справедливы следующие соотношения:

$$\mathbb{E}X^m = \begin{cases} a^2 + \sigma^2, & m = 2; \\ a^3 + 3a\sigma^2, & m = 3; \\ a^4 + 6a^2\sigma^2 + 3\sigma^4, & m = 4. \end{cases} \qquad (3.5)$$

Для начальных моментов случайной величины Z_n с функцией распределения $F(x, k(n), \mathbf{a}_n, \boldsymbol{\sigma}_n, \mathbf{p}_n)$ имеем:

$$\mathbb{E}Z_n^m = \sum_{i=1}^{k(n)} \frac{p_i(n)}{\sigma_i(n)\sqrt{2\pi}} \int_{-\infty}^{+\infty} z^m \exp\Big\{-\frac{(z - a_i(n))^2}{2\sigma_i^2(n)}\Big\}\, dz = \sum_{i=1}^{k(n)} p_i(n)\mathbb{E}X_i^m,$$

где $X_i \sim N(a_i(n), \sigma_i^2(n))$, то есть справедлив следующий аналог выражений (3.5):

$$\mathbb{E}Z_n^m = \begin{cases} \sum\limits_{i=1}^{k(n)} p_i(n)a_i(n), & m = 1; \\ \sum\limits_{i=1}^{k(n)} p_i(n)\left(a_i^2(n) + \sigma_i^2(n)\right), & m = 2; \\ \sum\limits_{i=1}^{k(n)} p_i(n)\left(a_i^3(n) + 3a_i\sigma_i^2(n)\right), & m = 3; \\ \sum\limits_{i=1}^{k(n)} p_i(n)\left(a_i^4(n) + 6a_i^2\sigma_i^2(n) + 3\sigma_i^4(n)\right), & m = 4. \end{cases} \qquad (3.6)$$

Воспользуемся этими выражениями для получения явных формул для дисперсии, а также коэффициентов асимметрии и эксцесса, зависящих только от параметров распределения, а именно величин $p_i(n)$, $a_i(n)$ и $\sigma_i(n)$.

Из первой строки в формуле (3.6) непосредственно следует представление (3.1). Дисперсия случайной величины Z_nс функцией распределения $F(x, k(n), \mathbf{a}_n, \boldsymbol{\sigma}_n, \mathbf{p}_n)$ имеет вид:

$$\mathbb{D}Z_n = \mathbb{E}Z_n^2 - (\mathbb{E}Z_n)^2 = \int\limits_{-\infty}^{+\infty} z^2\, dF_Z(z,t) - \left(\sum_{i=1}^{k(n)} p_i(n)a_i(n)\right)^2 =$$

$$= \sum_{i=1}^{k(n)} \frac{p_i(n)}{\sigma_i(n)\sqrt{2\pi}} \int\limits_{-\infty}^{+\infty} z^2 \exp\left\{-\frac{(z-a_i(n))^2}{2\sigma_i^2(n)}\right\} dz - \left(\sum_{i=1}^{k(n)} p_i(n)a_i(n)\right)^2 =$$

$$= \sum_{i=1}^{k(n)} p_i(n)a_i^2(n) - \left(\sum_{i=1}^{k(n)} p_i(n)a_i(n)\right)^2 + \sum_{i=1}^{k(n)} p_i(n)\sigma_i^2(n) =$$

$$= \sum_{i=1}^{k(n)} p_i(n)\left(a_i(n) - \sum_{i=1}^{k(n)} p_i(n)a_i(n)\right)^2 + \sum_{i=1}^{k(n)} p_i(n)\sigma_i^2(n).$$

Заключительное выражение в этой формуле легко получить путем раскрытия квадрата в первой сумме и приведения подобных слагаемых (например, именно в таком виде дисперсия приводится в книге [88]).

Общий вид коэффициента асимметрии случайной величины Z_n задается следующим выражением:

$$\gamma_{Z_n} = \frac{\mathbb{E}\left(Z_n - \mathbb{E}Z_n\right)^3}{\left(\mathbb{D}Z_n\right)^{3/2}}.$$

Выпишем отдельно выражение для числителя данной дроби:

$$\mathbb{E}\left(Z_n - \mathbb{E}Z_n\right)^3 = \mathbb{E}Z_n^3 - 3\mathbb{E}Z_n \cdot \mathbb{E}Z_n^2 + 3\left(\mathbb{E}Z_n\right)^3 - \left(\mathbb{E}Z_n\right)^3 =$$
$$= \mathbb{E}Z_n^3 - 3\mathbb{E}Z_n \cdot \left(\mathbb{E}Z_n^2 - \left(\mathbb{E}Z_n\right)^2\right) - \left(\mathbb{E}Z_n\right)^3 =$$
$$= \mathbb{E}Z_n^3 - 3\mathbb{E}Z_n \cdot \mathbb{D}Z_n - \left(\mathbb{E}Z_n\right)^3.$$

Тогда коэффициент асимметрии может быть представлен в виде

$$\gamma_{Z_n} = \frac{\mathbb{E}Z_n^3 - 3\mathbb{E}Z_n \cdot \mathbb{D}Z_n - \left(\mathbb{E}Z_n\right)^3}{\left(\mathbb{D}Z_n\right)^{3/2}}. \tag{3.7}$$

Воспользуемся формулами (3.6) и (3.2). Имеем

$$\mathbb{E}\left(Z_n - \mathbb{E}Z_n\right)^3 = \sum_{i=1}^{k(n)} p_i(n)\left(a_i^3(n) + 3a_i\sigma_i^2(n)\right) - 3\mathbb{E}Z_n \cdot \mathbb{D}Z_n - \left(\mathbb{E}Z_n\right)^3 =$$

$$= \sum_{i=1}^{k(n)} p_i(n)\left(a_i^3(n) + 3a_i\sigma_i^2(n)\right) - \left(\sum_{i=1}^{k(n)} p_i(n)a_i(n)\right) \times$$

$$\times \left(3\sum_{i=1}^{k(n)} p_i(n)\left(a_i(n) - \sum_{i=1}^{k(n)} p_i(n)a_i(n)\right)^2 + \right.$$

$$\left. + 3\sum_{i=1}^{k(n)} p_i(n)\sigma_i^2(n) - \left(\sum_{i=1}^{k(n)} p_i(n)a_i(n)\right)^2\right).$$

Откуда, подставляя значение (3.2) в знаменатель дроби (3.7), получим для коэффициента асимметрии представление (3.3).

Общий вид коэффициента эксцесса случайной величины Z_n задается следующим выражением:

$$\kappa_{Z_n} = \frac{\mathbb{E}\left(Z_n - \mathbb{E}Z_n\right)^4}{\left(\mathbb{D}Z_n\right)^2} - 3.$$

Для числителя в указанной выше дроби имеем:

$$\mathbb{E}\left(Z_n^4 - 4Z_n^3 \cdot \mathbb{E}Z_n + 6Z_n^2 \cdot \left(\mathbb{E}Z_n\right)^2 - 4Z_n \cdot \left(\mathbb{E}Z_n\right)^3 + \left(\mathbb{E}Z_n\right)^4\right) =$$

$$= \mathbb{E}Z_n^4 - 3\left(\mathbb{E}Z_n\right)^4 - 4\mathbb{E}Z_n \cdot \mathbb{E}Z_n^3 + 6\left(\mathbb{E}Z_n\right)^2 \cdot \mathbb{E}Z_n^2.$$

Тогда коэффициент эксцесса может быть представлен в виде

$$\kappa_{Z_n} = \frac{\mathbb{E}Z_n^4 - 4\mathbb{E}Z_n \cdot \mathbb{E}Z_n^3 + 6\left(\mathbb{E}Z_n\right)^2 \cdot \mathbb{E}Z_n^2 - 3\left(\mathbb{E}Z_n\right)^4}{\left(\mathbb{D}Z_n\right)^2} - 3. \tag{3.8}$$

Воспользуемся формулами (3.6) и (3.1). Имеем

$$\mathbb{E}\left(Z_n - \mathbb{E}Z_n\right)^4 = \sum_{i=1}^{k(n)} p_i(n)\left(a_i^4(n) + 6a_i^2\sigma_i^2(n) + 3\sigma_i^4(n)\right) -$$

$$-4\left(\sum_{i=1}^{k(n)} p_i(n)a_i(n)\right)\left(\sum_{i=1}^{k(n)} p_i(n)\left(a_i^3(n) + 3a_i\sigma_i^2(n)\right)\right) +$$

$$+6\left(\sum_{i=1}^{k(n)} p_i(n)a_i(n)\right)^2\left(\sum_{i=1}^{k(n)} p_i(n)\left(a_i^2(n) + \sigma_i^2(n)\right)\right) - 3\left(\sum_{i=1}^{k(n)} p_i(n)a_i(n)\right)^4.$$

Откуда, подставляя значение (3.2) в знаменатель дроби (3.4), получим для коэффициента эксцесса представление (3.4).

□

Формулы (3.1)–(3.4) могут быть использованы для непосредственного определения моментных характеристик по величинам $p_i(n)$, $a_i(n)$ и $\sigma_i(n)$ и не требуют знания значения момента предыдущего порядка. Для упрощения программной реализации разумно воспользоваться формулами вида (3.7) и (3.8). Кроме того, многие современные вычислительные системы оптимизированы для выполнения матричных вычислений. Поэтому теорема 3.1 может быть сформулирована в эквивалентной форме.

Теорема 3.2. *Моменты случайной величины Z_n с распределением $F(x, k(n), \mathbf{a}_n, \boldsymbol{\sigma}_n, \mathbf{p}_n)$ для использования в методе скользящего разделения смесей в матричной записи имеют следующий вид:*

– *математическое ожидание:*

$$\mathbb{E}Z_n = \mathbf{p}_n\,\mathbf{a}_n^T; \tag{3.9}$$

– *дисперсия:*

$$\mathbb{D}Z_n = \mathbf{p}_n\left(D_{\mathbf{a}_n}\,\mathbf{a}_n^T + D_{\boldsymbol{\sigma}_n}\,\boldsymbol{\sigma}_n^T\right) - (\mathbf{p}_n\,\mathbf{a}_n^T)^2; \tag{3.10}$$

– *коэффициент ассиметрии:*

$$\begin{aligned}\gamma_{Z_n} = {}& \frac{\mathbf{p}_n\,D_{\mathbf{a}_n}^2\,\mathbf{a}_n^T + 3\,\mathbf{p}_n\,D_{\mathbf{a}_n}\,D_{\boldsymbol{\sigma}_n}\,\boldsymbol{\sigma}_n^T + 2\,(\mathbf{p}_n\,\mathbf{a}_n^T)^2}{(\mathbf{p}_n\,(D_{\mathbf{a}_n}\,\mathbf{a}_n^T + D_{\boldsymbol{\sigma}_n}\,\boldsymbol{\sigma}_n^T) - (\mathbf{p}_n\,\mathbf{a}_n^T)^2)^{3/2}} - \\ & -3\cdot\frac{\mathbf{p}_n\,\mathbf{a}_n^T\,\mathbf{p}_n\,D_{\mathbf{a}_n}\,\mathbf{a}_n^T + \mathbf{p}_n\,\mathbf{a}_n^T\,\mathbf{p}_n\,D_{\boldsymbol{\sigma}_n}\,\boldsymbol{\sigma}_n^T}{(\mathbf{p}_n\,(D_{\mathbf{a}_n}\,\mathbf{a}_n^T + D_{\boldsymbol{\sigma}_n}\,\boldsymbol{\sigma}_n^T) - (\mathbf{p}_n\,\mathbf{a}_n^T)^2)^{3/2}};\end{aligned} \tag{3.11}$$

– *коэффициент эксцесса:*

$$\begin{aligned}\kappa_{Z_n} = {}& \frac{\mathbf{p}_n\left(D_{\mathbf{a}_n}^3\,\mathbf{a}_n^T + 6\,D_{\boldsymbol{\sigma}_n}^2\,D_{\mathbf{a}_n}\,\mathbf{a}_n^T + 3\,D_{\boldsymbol{\sigma}_n}^3\,\boldsymbol{\sigma}_n^T\right)}{(\mathbf{p}_n\,(D_{\mathbf{a}_n}\,\mathbf{a}_n^T + D_{\boldsymbol{\sigma}_n}\,\boldsymbol{\sigma}_n^T) - (\mathbf{p}_n\,\mathbf{a}_n^T)^2)^2} - \\ & -\frac{4\,\mathbb{E}Z_n\,\mathbf{p}_n\,D_{\mathbf{a}_n}\left(D_{\mathbf{a}_n}\,\mathbf{a}_n^T + 3\,D_{\boldsymbol{\sigma}_n}\,\boldsymbol{\sigma}_n^T\right)}{(\mathbf{p}_n\,(D_{\mathbf{a}_n}\,\mathbf{a}_n^T + D_{\boldsymbol{\sigma}_n}\,\boldsymbol{\sigma}_n^T) - (\mathbf{p}_n\,\mathbf{a}_n^T)^2)^2} + \\ & +\frac{6\,(\mathbb{E}Z_n)^2\,\mathbf{p}_n\left(D_{\mathbf{a}_n}\,\mathbf{a}_n^T + D_{\boldsymbol{\sigma}_n}\,\boldsymbol{\sigma}_n^T\right) - 3\,(\mathbb{E}Z_n)^4}{(\mathbf{p}_n\,(D_{\mathbf{a}_n}\,\mathbf{a}_n^T + D_{\boldsymbol{\sigma}_n}\,\boldsymbol{\sigma}_n^T) - (\mathbf{p}_n\,\mathbf{a}_n^T)^2)^2} - 3,\end{aligned} \tag{3.12}$$

где

$$\mathbf{p}_n = \left(p_1, \ldots, p_{k(n)}\right), \quad \mathbf{a}_n = \left(a_1, \ldots, a_{k(n)}\right), \quad \boldsymbol{\sigma}_n = \left(\sigma_1, \ldots, \sigma_{k(n)}\right),$$
$$D_{\mathbf{a}_n} = diag\left\{a_1, \ldots, a_{k(n)}\right\}, \quad D_{\boldsymbol{\sigma}_n} = diag\left\{\sigma_1, \ldots, \sigma_{k(n)}\right\},$$

и через $diag\{\ldots\}$ *обозначены диагональные матрицы с соответствующими элементами.*

Доказательство. Достаточно воспользоваться следующим матричным представлением выражений (3.6):

$$\mathbb{E}Z_n^m = \begin{cases} \mathbf{p}_n \, \mathbf{a}_n^T, & m = 1; \\ \mathbf{p}_n \left(D_{\mathbf{a}_n} \cdot \mathbf{a}_n^T + D_{\boldsymbol{\sigma}_n} \cdot \boldsymbol{\sigma}_n^T \right), & m = 2; \\ \mathbf{p}_n \cdot D_{\mathbf{a}_n} \left(D_{\mathbf{a}_n} \cdot \mathbf{a}_n^T + 3 \cdot D_{\boldsymbol{\sigma}_n} \cdot \boldsymbol{\sigma}_n^T \right), & m = 3; \\ \mathbf{p}_n \left(D_{\mathbf{a}_n}^3 \cdot \mathbf{a}_n^T + 6 \cdot D_{\boldsymbol{\sigma}_n}^2 \cdot D_{\mathbf{a}_n} \cdot \mathbf{a}_n^T + 3 \cdot D_{\boldsymbol{\sigma}_n}^3 \cdot \boldsymbol{\sigma}_n^T \right), & m = 4, \end{cases}$$

которые и ведут к формулам (3.9)–(3.12). □

Замечание 3.1. Значение величины (3.9) был использовано для сокращения вида выражения (3.12). Также можно поступить с формулой (3.11), используя значение (3.10).

Замечание 3.2. Преимущество в скорости с использованием матричных вычислений эмпирически продемонстрировано в работе [222] на примере классического EM-алгоритма.

3.2 Метод адаптивного выделения смешанного нормального сигнала на фоне смешанного гауссовского шума

При регистрации сигналов в большинстве реальных задач помимо содержательной части возникает случайный шум, характеристики которого чаще всего неизвестны. При этом в ряде исследований за счет организации эксперимента возможно заранее получить некоторый набор наблюдений, которые описывают только шум, и связаны как с работой детектирующих приборов, так и с особенностями поведения наблюдаемого объекта – и за счет этого учесть его влияние в процессе дальнейшей работы. Очевидно, что такие модификации получаемой выборки не связаны непосредственно с проводимым экспериментом, однако влияют на его результаты. Данная проблема характерна для широкого спектра исследовательских задач, в том числе в физических экспериментах [146], медицинских приложениях [131, 236, 324], при анализе сигналов с негауссовским шумом [134, 263, 272], предобработке изображений [317].

Во многих реальных задачах обработки сигналов распределение шума существенно отличается стандартного нормального (“белого”). Так, в статье [330] рассмотрен пример анализа климатических временных рядов с учетом “красного” шума. Более того, зачастую шум оказывается существенно негауссовским (см., например, [159]). Подробнее этот вопрос, включая примеры использования конечных смесей нормальных законов для описания шума, рассмотрен в статье [391]. В ней же приводится статистический критерий на основе отношения правдоподобия для определения гауссовского сигнала в смешанном гауссовском шуме.

В данном разделе будут рассмотрены вопросы определения параметров распределения полезного сигнала в предположении, что параметры распределения шума были оценены предварительно. Для полезного сигнала будет использован адаптивный подход на основе метода скользящего разделения смесей. Также предложен прикладной подход к обнаружению момента появления полезного сигнала в наблюдениях.

3.2.1 Постановка задачи

Предположим, что исходные наблюдения представляют собой реализацию случайной величины $Z = X + Y$, где X соответствует полезному сигналу (информации) в данных, а Y – шум измерительного оборудования, обусловленный набором факторов, в том числе случайного характера. Будем предполагать, что случайные величины X и Y являются независимыми. Дополнительно предположим, что до начала эксперимента есть возможность получить набор наблюдений, являющихся реализацией только случайной величины Y. Таким образом, рассматривается модель аддитивного шума для исходных данных.

Пусть распределение случайной величины Y может быть представлено в виде конечной смеси нормальных законов с неизвестными параметрами. Тогда, используя, например, одну из модификаций EM-алгоритма (классическую, являющуюся одним из самых широко распространенных инструментов интеллектуального анализа данных [423], или сеточную [234] версии) можно получить оценки максимального правдоподобия соответствующих параметров и считать данное распределение заданным, то есть с учетом представлений (1.7) и (1.8)

$$Y \sim F(x, \widetilde{k}, \widetilde{\mathbf{a}}, \widetilde{\boldsymbol{\sigma}}, \widetilde{\mathbf{p}}). \tag{3.13}$$

При этом все величины в выражении (3.13) считаем известными.

Предположим, что распределение случайной величины X также может быть описано конечной смесью нормальных распределений (1.7) с ограничениями (1.8), то есть

$$X \sim F(x, k, \mathbf{a}, \boldsymbol{\sigma}, \mathbf{p}). \tag{3.14}$$

Все величины в выражении (3.14), включая и число компонент k, считаем неизвестными. Возникает задача восстановления данного распределения (то есть оценивания неизвестных параметров смеси) с помощью метода скользящего разделения смесей [88] для реализаций случайной величины Z. В следующем разделе подробно рассмотрим несколько подходов к решению данной задачи. Всюду далее будем предполагать, что оценки максимального правдоподобия неизвестных параметров распределения найдены с помощью какой-либо модификации EM-алгоритма.

3.2.2 Анализ полезного сигнала с учетом предварительных оценок для шума

Основываясь на понятии свертки как распределении суммы двух независимых случайных величин, а также факте, что распределение суммы двух независимых нормальных случайных величин также является нормальным с параметрами, получаемыми суммированием исходных, легко показать, что в указанных выше предположениях распределение случайной величины Z имеет вид:

$$Z \sim F(x, k \cdot \widetilde{k}, \widehat{\mathbf{a}}, \widehat{\boldsymbol{\sigma}}, \widehat{\mathbf{p}}), \tag{3.15}$$

причем для параметров в выражении (3.15) справедливы следующие представления:

$$\widehat{p}_{(r-1)\widetilde{k}+j} = p_r \widetilde{p}_j, \quad \widehat{a}_{(r-1)\widetilde{k}+j} = a_r + \widetilde{a}_j, \quad \widehat{\sigma}^2_{(r-1)\widetilde{k}+j} = \sigma_r^2 + \widetilde{\sigma}_j^2 \tag{3.16}$$

сразу для всех индексов $r = \overline{1,k}$ и $j = \overline{1,\widetilde{k}}$. Также возможно проверить данные соотношения с использованием аппарата характеристических функций, как продемонстрировано в книге [88].

В приведенных выше формулах параметр k считается известным, однако дополнительные предположения в выражении (3.14) не требуются. Действительно, количество компонент в смеси (3.15) (скажем, $m = k \cdot \widetilde{k}$) может считаться неизвестным – и подлежать автоматическому определению, как это и принято в методе скользящего разделения смесей. Достаточно потребовать, чтобы $\widetilde{k}$ было делителем числа m. Тогда параметр

k также может быть определен в автоматическом режиме. Кроме того, необходимые «дополнительные» компоненты могут быть получены искусственным добавлением слагаемых с нулевыми весами.Тогда указанное предположение о делимости может и не выполняться.

Относительно параметров p_r, a_r и σ_r^2, $r = \overline{1,k}$, система (3.16) является переопределенной, так как количество уравнений превышает число переменных ($j = \overline{1,\widetilde{k}}$):

$$p_{r,j} = \widehat{p}_{(r-1)\widetilde{k}+j} \cdot \widetilde{p}_j^{-1}, \quad a_{r,j} = \widehat{a}_{(r-1)\widetilde{k}+j} - \widetilde{a}_j, \quad \sigma_{r,j}^2 = \widehat{\sigma}_{(r-1)\widetilde{k}+j}^2 - \widetilde{\sigma}_j^2. \tag{3.17}$$

Классическим подходом для поиска приближенного решения переопределенной системы линейных алгебраических уравнений является использование метода наименьших квадратов (МНК).

Теорема 3.3. *Оценки параметров неизвестного распределения случайной величины X (3.14) на основе МНК могут быть записаны в следующем виде:*

$$p_r = \widetilde{k}^{-1} \sum_{j=1}^{\widetilde{k}} \widehat{p}_{(r-1)\widetilde{k}+j} \cdot \widetilde{p}_j^{-1}, \quad a_r = \widetilde{k}^{-1} \sum_{j=1}^{\widetilde{k}} \left(\widehat{a}_{(r-1)\widetilde{k}+j} - \widetilde{a}_j \right), \tag{3.18}$$

$$\sigma_r^2 = \widetilde{k}^{-1} \sum_{j=1}^{\widetilde{k}} \left(\widehat{\sigma}_{(r-1)\widetilde{k}+j}^2 - \widetilde{\sigma}_j^2 \right). \tag{3.19}$$

Доказательство. Для системы вида $\mathcal{A}\mathbf{x} = \mathbf{b}$, где $\mathcal{A} \in \mathbb{R}^{m \times n}$, $\mathbf{x} \in \mathbb{R}^n$ и $\mathbf{b} \in \mathbb{R}^m$, решение на основе МНК может быть записано как

$$\mathbf{x} = (\mathcal{A}^T \mathcal{A})^{-1} \mathcal{A}^T \mathbf{b}. \tag{3.20}$$

Будем рассматривать как отдельную систему выражения для каждого из параметров в (3.17), положив в формуле (3.20) $m = \widetilde{k}$, $n = 1$. Тогда матрица $\mathcal{A} = \mathbf{1}_{\widetilde{k} \times 1}$ (единичный вектор соответствующего размера) и

$$(\mathcal{A}^T \mathcal{A})^{-1} \mathcal{A}^T = \widetilde{k}^{-1} \mathbf{1}_{1 \times \widetilde{k}}.$$

Подставляя конкретный вид векторов $\mathbf{b}$ в выражение (3.20), получим формулы (3.18) и (3.19). □

Замечание 3.3. В распределении (3.13) разумно использовать умеренные значения для параметра $\widetilde{k}$ (например, 1–2), так как в реальных данных результирующее число компонент редко превышает 4–6, а выбор

больших значений для $\widetilde{k}$ может привести к необходимости аппроксимации смесями с заметным количеством компонент (в силу мультипликативности в распределении (3.15)). Это снижает точность и повышает нагрузку на вычислительные ресурсы.

Замечание 3.4. В методе скользящего разделения смесей выражения вида (3.18) и (3.19) принято записывать (и оценивать) в зависимости от времени. При этом оценку шума необходимо делать сразу по всему ряду, не используя скользящее окно. Данное предложение является достаточно естественным, но также и сокращает время работы численных методов. Таким образом, формулы (3.18) и (3.19) можно представить в следующем виде ($r = \overline{1,k}$):

$$p_r(t) = \widetilde{k}^{-1} \sum_{j=1}^{\widetilde{k}} \widehat{p}_{(r-1)\widetilde{k}+j}(t) \cdot \widetilde{p}_j^{-1}, \quad a_r(t) = \widetilde{k}^{-1} \sum_{j=1}^{\widetilde{k}} \left(\widehat{a}_{(r-1)\widetilde{k}+j}(t) - \widetilde{a}_j \right),$$

$$\sigma_r^2(t) = \widetilde{k}^{-1} \sum_{j=1}^{\widetilde{k}} \left(\widehat{\sigma}^2_{(r-1)\widetilde{k}+j}(t) - \widetilde{\sigma}_j^2 \right).$$

В данном случае параметр t может быть ассоциирован как с астрономическим временем, так и с положением (номером) скользящего окна в СРС-методе. Оба этих варианта являются равнозначными.

Замечание 3.5. Для корректной аппроксимации (3.13) требуется выборка достаточного объема, не зависящая от полезного сигнала. Указанное обстоятельство должно приниматься во внимание при проведении экспериментов и сборе данных.

Получим матричные аналоги формул (3.18) и (3.19). Введем следующие обозначения ($r = \overline{1,k}$):

$$\widetilde{A} = \widetilde{\mathbf{a}}\, \mathbf{1}_{\widetilde{k}\times 1}, \quad \widetilde{\Sigma} = \widetilde{\boldsymbol{\sigma}}\, \mathbf{1}_{\widetilde{k}\times 1}, \quad \mathcal{E} = \bigoplus_{r=1}^{k} \mathbf{1}_{\widetilde{k}\times 1}, \tag{3.21}$$

$$\widetilde{\mathbf{a}} = (a_1, \ldots, a_{\widetilde{k}}), \quad \widetilde{\boldsymbol{\sigma}} = (\sigma_1^2, \ldots, \sigma_{\widetilde{k}}^2), \quad \widehat{\mathbf{p}}_r = (\widehat{p}_{(r-1)\widetilde{k}+1}, \ldots, \widehat{p}_{r\widetilde{k}}), \tag{3.22}$$

$$\widehat{\mathbf{a}}_r = (a_{(r-1)\widetilde{k}+1}, \ldots, a_{r\widetilde{k}}), \quad \widehat{\boldsymbol{\sigma}}_r = (\sigma^2_{(r-1)\widetilde{k}+1}, \ldots, \sigma^2_{r\widetilde{k}}), \tag{3.23}$$

$$\widetilde{\mathbf{p}}_r^{-1} = (\widetilde{p}_1^{-1}, \ldots, \widetilde{p}_{\widetilde{k}}^{-1}). \tag{3.24}$$

Здесь $\bigoplus$ обозначает прямую сумму соответствующих матриц [138], таким образом, $\mathcal{E}$ в формуле (3.21) имеет блочно-диагональную структуру (элементы – векторы из единиц размера $\widetilde{k}$).

Теорема 3.4. *В обозначениях* (3.21)–(3.24) *оценки параметров неизвестного распределения случайной величины* X (3.14) *на основе МНК могут быть записаны в следующем виде:*

$$\mathbf{p} = \widetilde{k}^{-1}\left[\left(\widetilde{\mathbf{p}}_1^{-1}\,\widetilde{\mathbf{p}}_2^{-1}\,\cdots\,\widetilde{\mathbf{p}}_k^{-1}\right)\circ\widehat{\mathbf{p}}\right]\,\mathcal{E},\quad \mathbf{a} = \widetilde{k}^{-1}\left(\widehat{\mathbf{a}}\,\mathcal{E} - \widetilde{A}\,\mathbf{1}_{1\times k}\right), \tag{3.25}$$

$$\boldsymbol{\Sigma} = \widetilde{k}^{-1}\left(\widehat{\boldsymbol{\sigma}}\,\mathcal{E} - \widetilde{\Sigma}\,\mathbf{1}_{1\times k}\right), \tag{3.26}$$

где $\mathbf{p} = (p_1, \ldots, p_k)$, $\mathbf{a} = (a_1, \ldots, a_k)$, $\boldsymbol{\Sigma} = (\sigma_1, \ldots, \sigma_k)$, $\widehat{\mathbf{p}} = (\widehat{\mathbf{p}}_1\ \cdots\ \widehat{\mathbf{p}}_k)$, $\widehat{\mathbf{a}} = (\widehat{\mathbf{a}}_1\ \cdots\ \widehat{\mathbf{a}}_k)$, $\widehat{\boldsymbol{\sigma}} = (\widehat{\boldsymbol{\sigma}}_1\ \cdots\ \widehat{\boldsymbol{\sigma}}_k)$, *а символ* $\circ$ *обозначает произведение Адамара [190, 264, 396].*

Справедливость данного утверждения следует непосредственно из записи формул (3.18) и (3.19) в матричной форме, в которых операция суммирования заменяется умножением вектора-строки на соответствующий по размеру единичный вектор-столбец.

Выражения (3.25), (3.26) могут быть переписаны с привязкой к номеру окна/времени (аналогично замечанию 3.4), однако для упрощения представления в теореме 3.4 формулы приводятся для каждого окна СРС-метода. В общем виде всем векторам можно добавить еще один индекс t, а для соответствующих компонентов указать зависимость от времени. Сами выражения от этого принципиально не изменятся.

3.2.3 Алгоритм адаптивного определения параметров распределения полезного сигнала

Итак, процедура определения параметров распределения полезного сигнала с учетом предварительных оценок для шума может быть представлена в виде последовательности следующих шагов:

1. Определение части выборки, содержащей только шум.
2. Определение параметров распределения шума (3.13) $\widetilde{p}_j$, $\widetilde{a}_j$, $\widetilde{\sigma}_j$, $j = \overline{1, \widetilde{k}}$, с помощью EM-алгоритма по части выборки, свободной от полезного сигнала.
3. Определение части выборки, содержащей как сигнал, так и шум.
4. Определение параметров распределения сигнала с шумом (3.15) $\widehat{p}_j$, $\widehat{a}_j$, $\widehat{\sigma}_j$, $j = \overline{1, k\cdot\widetilde{k}}$, с помощью EM-алгоритма в режиме скользящего окна (для каждого фиксированного положения).
5. Определение параметров распределения «чистого» сигнала (3.14) p_j, a_j, σ_j, $j = \overline{1, k}$, с помощью EM-алгоритма в режиме скользящего окна на основе теорем 3.3 или 3.4 (для каждого фиксированного положения).

Данная процедура должна продолжаться до момента исчерпания выборки. Описанный алгоритм (см. также рисунок 3.1) может быть реализован на любом удобном пользователю языке программирования.

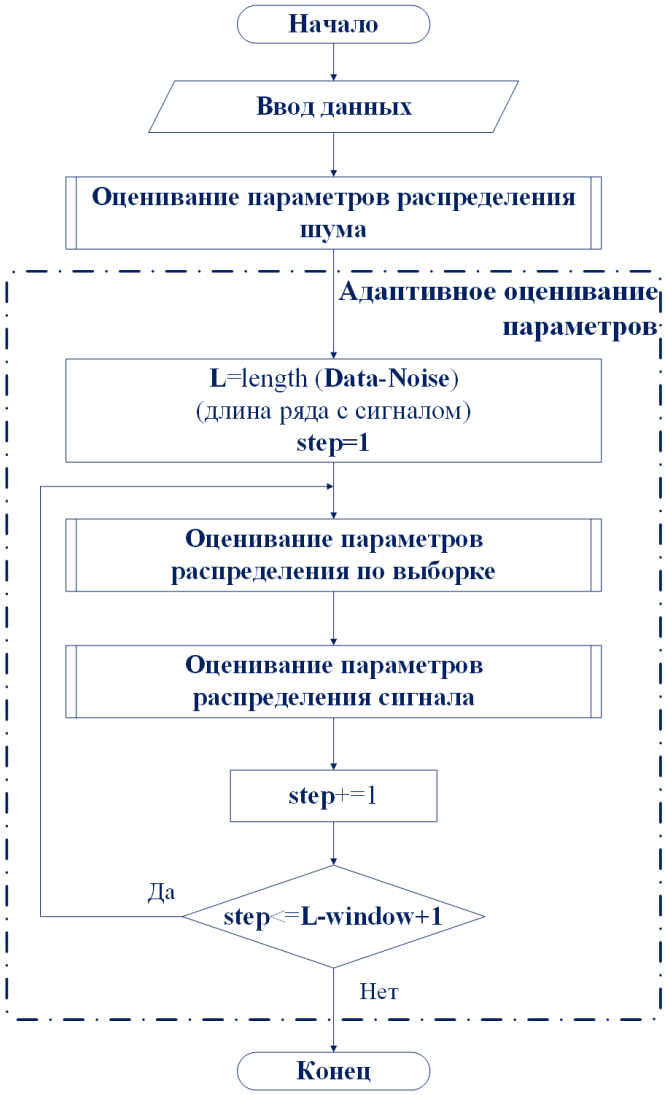

Рис. 3.1. Алгоритм адаптивного определения параметров распределения сигнала с помощью СРС-метода

Здесь **Data** обозначает исходные данные; **Noise** – часть ряда, содержащую только шум; переменная **step** используется для подсчета шагов в СРС-методе. В блоках скрыты реализации процедур определения оценок максимального правдоподобия с помощью алгоритмов ЕМ-типа.

В следующих разделах будет продемонстрирована эффективность предложенной процедуры определения параметров полезного сигнала при условии наличия шума для различных соотношений между их параметрами на примере 24 модельных выборок, охватывающих большинство возможных реальных сценариев. Также обсуждаются вопросы прикладного подхода к обнаружению момента появления полезного сигнала в наблюдениях.

3.2.4 Генерация тестовых выборок

При генерации тестовых выборок необходимо учитывать, что распределение шума (3.13) удобнее описывать небольшим числом компонент в смеси (то есть $\widetilde{k}$ равняется 1–2), так как в реальных данных результирующее число компонент редко превышает 4–6, а данный параметр непосредственно влияет на число компонент в распределении (3.15). Выбор больших значений $\widetilde{k}$ может вести к заметному снижению вычислительной точности, а также значимому увеличению необходимого для расчетов времени. Процедура генерации выборок представлена в алгоритме 3.1.

Алгоритм 3.1. Генерация тестовых выборок

```
1:  function SIMULATION(L_X, L_Y)
2:      // Y_PURE – выборка для оценки параметров шума;
3:      max_sigma = 0.0;
4:      for (i = 0, i < L_X, i++) do
5:          far_enough ← false;
6:          while not (far_enough) do
7:              [mean, sigma] ← RANDUNIFORM( );
8:              max_σ ← MAX(σ, max_σ);
9:              far_enough ← true; j ← 1;
10:             for (j < i) do
11:                 if (ABS(mean - Y_means(j)) < MAX(Y_σ(j),max_σ)) then
12:                     far_enough ← false;
13:                 j++;
14:     // Функция генерации выборок с заданными параметрами
15:     [X, Y, Y_PURE]←MIXTURESIMULATION(X_Params, Y_Params, L_X, L_Y);
16:     Z ← X + Y;
17:     return [X, Y, Y_PURE, Z];
```

Число компонент для сигнала при симуляции выбиралось как $k = \overline{1,4}$, для шума – $\widetilde{k} = \{1, 2\}$. Веса для всех случаев задавались случайным вектор с компонентами из диапазона $[0.1, 0.9]$ и дальнейшей нормировкой для выполнения условий из формул (3.14) и (3.13). Дисперсии для шума выбирались случайно из отрезка $[0.3, 1.5]$, для сигнала – из отрезка $[0.1, 3.0]$.

Параметры математического ожидания для шума генерировались следующим образом. Сначала выбиралось случайное число из отрезка $[-5, 5]$. Если из уже сгенерированных к данному моменту компонент

шума не находилось такой компоненты, расстояние до среднего которой было бы меньше, чем максимальное из уже сгенерированных отклонений, то это число назначалось средним и цикл продолжался. Если оно не удовлетворяло этому условию, то операция повторялась, до тех пор, пока подходящее (достаточно удаленное от других компонент) значение не будет найдено.

Симуляция параметров для сигнала проводилась аналогичным образом, но при этом в зависимости от параметра *положение шума относительно сигнала* менялся отрезок, из которого выбирается случайное число для генерации среднего каждой компоненты. При близком положении шума относительно сигнала использовались значения из диапазона $[-10, 10]$, для промежуточной ситуации – из отрезка $[5, 30]$, для далеко расположенных – из $[15, 100]$.

Таблица 3.1. Параметры распределений сигнала для модельных выборок

No.	**k**	**Параметры сигнала X**
1	1	[1,0]; [6,76]; [1,28]
2	1	[1,0]; [12,51]; [0,15]
3	1	[1,0]; [89,2]; [2,91]
4	1	[1,0]; [5,08]; [1,47]
5	1	[1,0]; [15,83]; [0,19]
6	1	[1,0]; [44,87]; [1,24]
7	2	[0,29; 0,71]; [9,07; 5,81]; [2,4; 0,72]
8	2	[0,35; 0,65]; [19,93; 5,75]; [0,93; 2,19]
9	2	[0,5; 0,5]; [63,65; 91,93]; [2,7; 1,11]
10	2	[0,56; 0,44]; [7,39; −9,58]; [1,95; 0,43]
11	2	[0,26; 0,74]; [25,94; 29,77]; [0,14; 1,21]
12	2	[0,43; 0,57]; [54,86; 77,32]; [1,96; 2,3]
13	3	[0,11; 0,42; 0,47]; [7,71; 0,76; −4,51]; [1,17; 2,43; 1,67]
14	3	[0,39; 0,5; 0,11]; [23,65; 16,1; 10,15]; [2,91; 1,09; 1,38]
15	3	[0,35; 0,36; 0,3]; [72,89; 63,51; 48,83]; [2,75; 2,13; 1,59]
16	3	[0,31; 0,25; 0,44]; [−5,74; 7,63; −0,01]; [0,29; 2,26; 1,03]
17	3	[0,29; 0,52; 0,19]; [7,38; 16,09; 24,01]; [0,94; 2,92; 2,44]
18	3	[0,13; 0,45; 0,43]; [42,42; 68,55; 73,92]; [0,17; 0,29; 0,67]
19	4	[0,25; 0,06; 0,34; 0,35]; [−4,71; 6,38; 0,77; 3,17]; [2,38; 1,77; 1,27; 1,95]
20	4	[0,37; 0,36; 0,14; 0,12]; [6,9; 24,97; 18,74; 28,15]; [2,58; 2,47; 2,59; 0,22]
21	4	[0,28; 0,05; 0,42; 0,24]; [69,55; 85,85; 91,88; 59,18]; [2,11; 1,8; 2,42; 0,97]
22	4	[0,29; 0,19; 0,35; 0,17]; [0,42; −4,15; −9,09; 6,41]; [2,58; 1,69; 0,3; 0,57]
23	4	[0,29; 0,19; 0,28; 0,24]; [17,6; 7,84; 12,11; 26,12]; [0,88; 0,28; 1,71; 1,87]
24	4	[0,29; 0,2; 0,27; 0,23]; [38,63; 48,02; 73,88; 62,73]; [0,43; 0,2; 0,3; 0,85]

В таблицах 3.1 и 3.2 приведены параметры смешанных распреде-

лений, использованные для симуляции тестовых выборок: для аналога «экспериментальных» значений выбран объем наблюдений $L_Z = 1500$, а для оценки шума – размер $L_Y = 10000$. Параметры распределений генерируются случайным образом, при этом задействована процедура, которая позволяет моделировать как близкие параметры для сигнала и шума (выборка 4), так и существенно отличающиеся (выборка 11).

Таблица 3.2. Параметры распределений шума для модельных выборок

No.	$\widetilde{\mathbf{k}}$	**Параметры шума Y**
1	1	[1,0]; [1,59]; [0,5]
2	1	[1,0]; [−4,68]; [0,84]
3	1	[1,0]; [2,62]; [1,08]
4	2	[0,41; 0,59]; [−1,95; 4,26]; [1,26; 1,28]
5	2	[0,21; 0,79]; [−3,09; −0,9]; [0,33; 1,41]
6	2	[0,51; 0,49]; [−4,45; −1,49]; [1,1; 1,34]
7	1	[1,0]; [−1,65]; [1,17]
8	1	[1,0]; [−4,98]; [1,18]
9	1	[1,0]; [0,49]; [0,86]
10	2	[0,65; 0,35]; [−3,59; 2,53]; [1,46; 0,89]
11	2	[0,73; 0,27]; [2,5; 4,35]; [0,95; 0,87]
12	2	[0,56; 0,44]; [−1,31; 0,29]; [0,39; 1,31]
13	1	[1,0]; [−4,46]; [0,97]
14	1	[1,0]; [4,41]; [0,69]
15	1	[1,0]; [−1,31]; [0,99]
16	2	[0,33; 0,67]; [−2,42; −4,25]; [0,4; 0,67]
17	2	[0,58; 0,42]; [−2,76; 0,21]; [0,43; 1,16]
18	2	[0,57; 0,43]; [−0,45; −3,75]; [1,42; 1,18]
19	1	[1,0]; [−2,74]; [0,87]
20	1	[1,0]; [−3,11]; [0,68]
21	1	[1,0]; [3,13]; [0,42]
22	2	[0,31; 0,69]; [2,99; −0,85]; [1,04; 0,79]
23	2	[0,39; 0,61]; [−3,96; −2,17]; [1,1; 0,79]
24	2	[0,45; 0,55]; [0,48; −4,98]; [0,56; 0,36]

Результаты симуляции (распределения тестовых выборок) представлены на рисунках 3.2 и 3.3. Очевидно, что в данном случае охватывается широкий спектр возможных соотношений между параметрами сигнала и шума как с точки зрения соотношения между математическими ожида-

ниями, так и для дисперсий. Также рассмотрены различные комбинации для компонент распределения случайных величин X и Y.

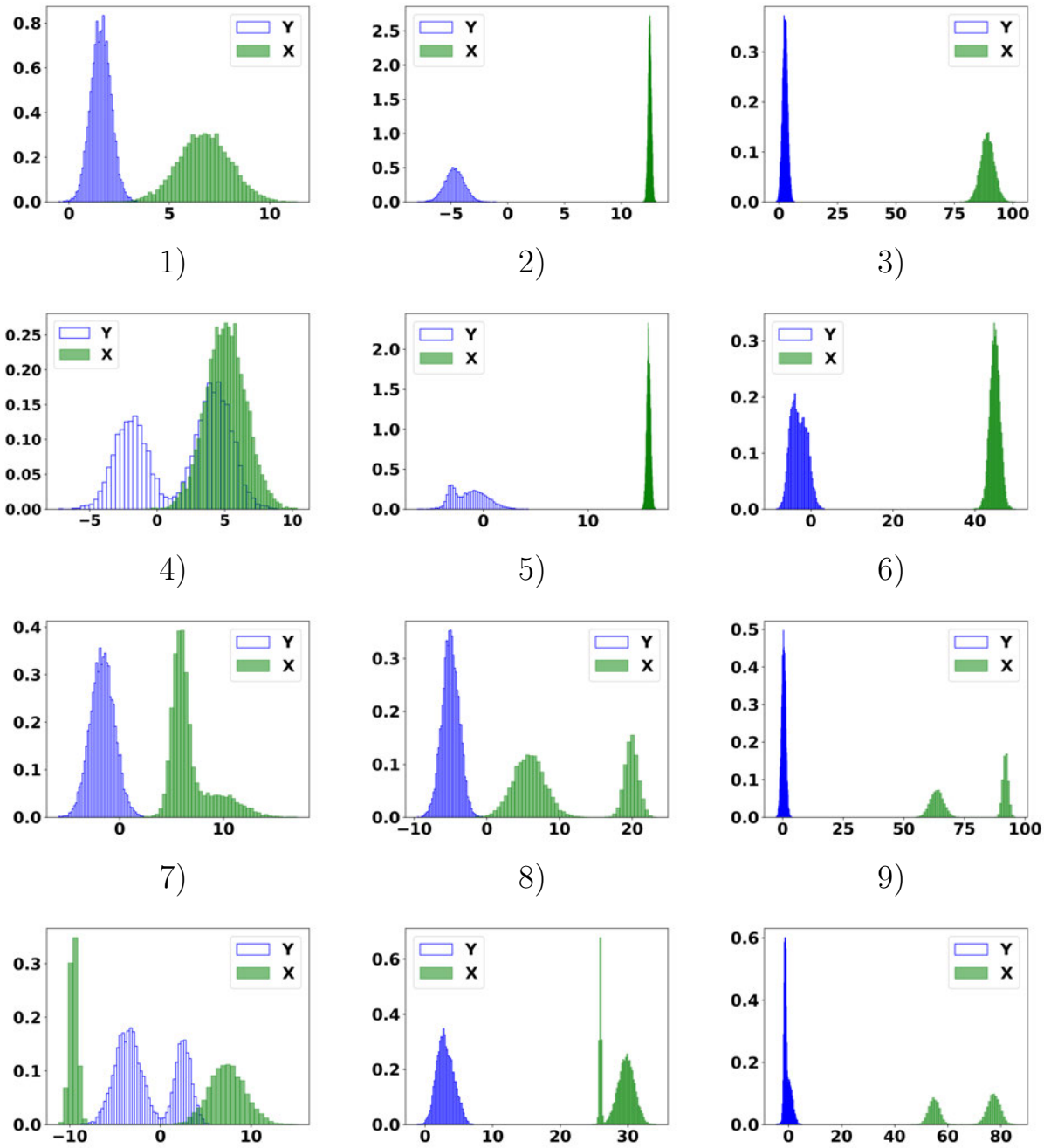

Рис. 3.2. Тестовые данные для сигнала и шума (выборки $1-12$)

3.2.5 Обнаружение момента разладки

Момент точного начала регистрации полезного сигнала на практике заранее обычно неизвестен (например, в силу наличия задержки в отклике системы или регистрирующего оборудования). Поэтому возникает необходимость решения классической задачи о разладке [348, 352],

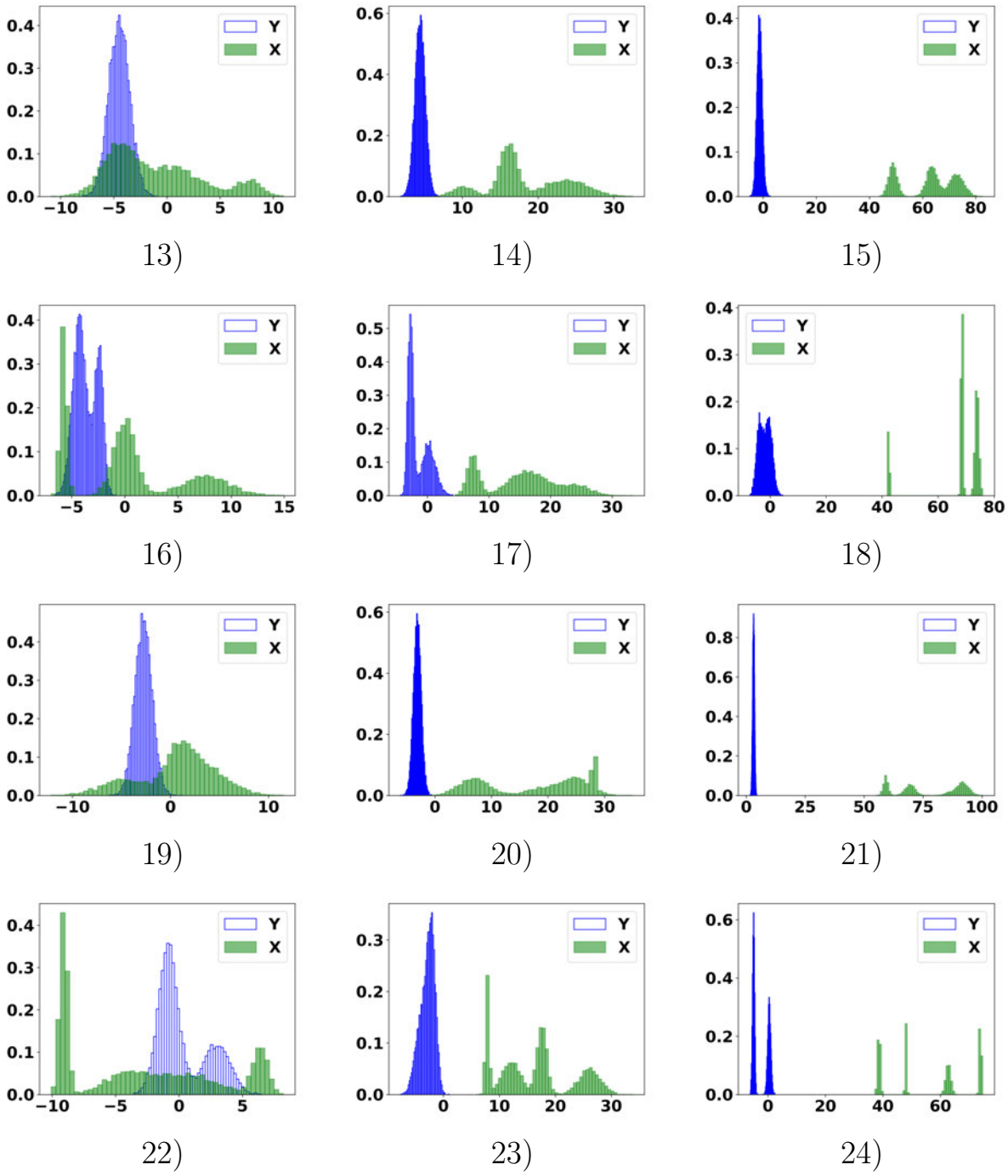

Рис. 3.3. Тестовые данные для сигнала и шума (выборки 13 – 24)

то есть об определении момента существенного изменения распределения данных в выборке. Для ее практического решения был использован метод скользящего тестирования однородности на основе критериев Колмогорова-Смирнова и Колмогорова.

Итак, первый метод заключался в выборе двух подряд идущих непересекающихся окон (выборок) ширины от 75 до 100 (в зависимости от настроек) каждое. Для проверки однородности использовался критерий

Колмогорова-Смирнова. Второй метод основан на использовании окон большего размера $(250 - 500)$ и критерия однородности Колмогорова. При этом во втором случае использовалась независимая выборка для предоценивания параметров шума для сравнения в рамках критерия с эмпирической функцией распределения текущего окна. Скользящее окно позволяет считать попадающую в нее выборку однородной, что необходимо для применения указанных статистических процедур. Кроме того, применялся двунаправленный проход по выборке: в прямом и обратном направлениях. В качестве момента разладки выбиралось среднее значения решений для каждого из проходов. Это в конечном итоге позволило уменьшить ошибку детектирования (задержку) по сравнению с использованием однопроходного метода.

По итогам предварительного тестирования обоих методов было установлено, что подход на основе критерия однородности Колмогорова-Смирнова предпочтительнее в силу его более высокой точности, а также меньшей средней ошибке детектирования при однонаправленном (прежде всего, прямом) проходе по сравнению с алгоритмом на основе критерия Колмогорова.

На практике содержательная часть сигнала может быть существенно неоднородной, поэтому корректность применения двунаправленого метода может нарушаться. Таким образом, качество одностороннего решения является ключевым. Также необходимо отметить, что с вычислительная эффективность данного метода примерно на два порядка выше аналогичной процедуры на основе критерия Колмогорова. Данная процедура представлена в алгоритме 3.2. На рисунке 3.4 продемонстрировано применение данного алгоритма к каждой из тестовых выборок.

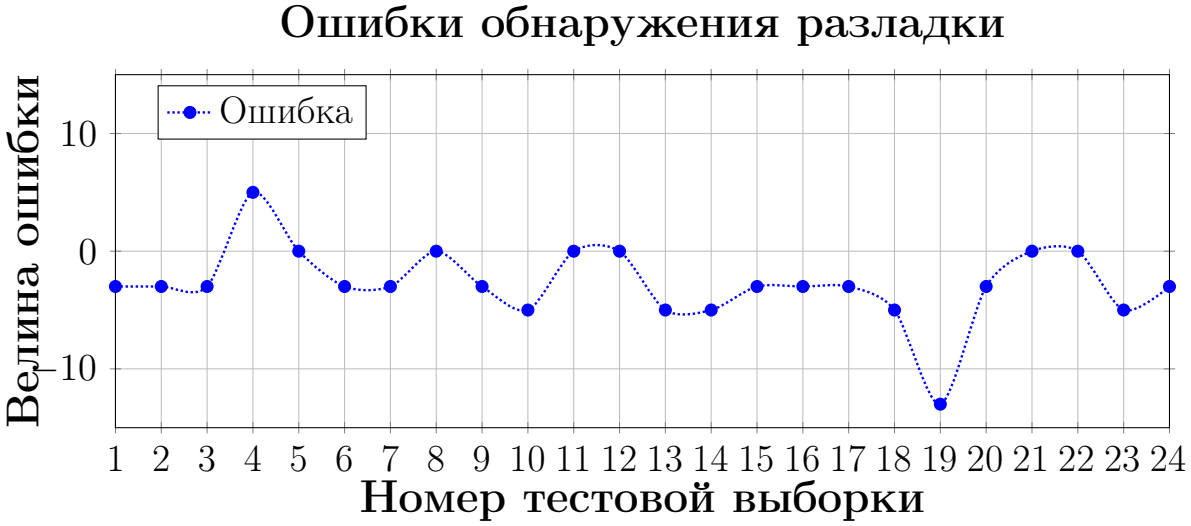

Рис. 3.4. Величина ошибки обнаружения разладки для тестовых выборок

Алгоритм 3.2. Детектирование момента разладки с использованием критерия однородности Колмогорова-Смирнова

```
1: function ChangePoint(Data, direct, window=100, step=5, α = 10^-5)
2:     flag ← (direct=='fwd');     // флаг прямого и обратного проходов
3:     L←Length(Data) - 2·window+1;
4:     i←(1-flag)·L;
5:     while (0⩽i⩽L) do          // Скользящая проверка однородности
6:         [Sample1, Sample2] ← SubSamples(Data, window, i);
7:         if (KSTest(Sample1, Sample2)<α) then
8:             P←i;
9:             break;
10:        else
11:            i ← i+(2·flag-1)·step;
12:            if ((i=0) or (i=L)) then
13:                break;
14:    return P;
   ...
15: for all TimeSeries do // Все тестовые выборки из таблиц 3.1 и 3.2
16:    Data←TimeSeries(i);
17:    Point ← 1/2 · (ChangeP(Data, 'fwd') + ChangeP(Data, 'back'));
```

Положительное значение показывает, насколько метод запаздывает в детектировании момента разладки, а отрицательное – насколько «опережает». Наибольшая ошибка была получена для выборки номер 19 – на 13 элементов раньше реальной позиции начала разладки. Стоит выделить выборки с номерами $5, 8, 11, 12, 21$ и 22, для которых двухпроходный метод позволил диагностировать момент изменения распределения полностью корректно.

3.2.6 Реализация метода оценивания неизвестных параметров

Номера компонент в формулах (3.15) для простоты записи имеют некоторую специальную блочную структуру. А именно, предполагается, что первые $\widetilde{k}$ компонент распределения случайной величины Z соответствуют первой компоненте сигнала, следующие подряд идущие $\widetilde{k}$ членов – второй и так далее. Очевидно, что при реализации вычислительного

алгоритма данное предположение сложно реализовать, поскольку оно требует знание структуры распределения. Для преодоления указанной проблемы при реализации адаптивного метода оценивания была предложена следующая процедура. Значения в формулах (3.18) и (3.19) вычисляются для всех возможных перестановок компонент распределения случайной величины Z, при этом в качестве «корректного» порядка используется такая, при которой слагаемые в каждой из сумм наиболее близки между собой (в данном случае используется стандартная метрика ℓ^2). Данная процедура представлена в алгоритме 3.3.

Алгоритм 3.3. Определение неизвестных параметров

```
1:  function GetX_Params(L_X, L_Y)
2:      EM_distance ← 1.1;                  // точность оценивания параметров Z
3:      while (EM_distance > 1.0) do
4:          // Если оценка недостаточно точна, перегенерация выборки
5:          [X, Y, Y_PURE, Z] ← Simulation(L_X, L_Y)     // Алгоритм 3.1
6:          // Z_params_real – реальные параметры зашумленного сигнала
7:          EM_distance ← RMSE(Z_params_est, Z_params_real)
8:      // Z_params_est – оценки параметров зашумленного сигнала
9:      Z_params_est ← ApplyGMM(Z);
10:     // Y_params_est – оценки параметров шума
11:     Y_params_est ← ApplyGMM(Y_PURE);
12:     // Перебор всех сочетаний компонент в (3.15)
13:     for all Z_Permutations do
14:         δ ← 0; r ←1;
15:         for (r ⩽ L_X) do
16:             j←1;
17:             for (j ⩽ L_Y) do
18:                 // Оценивание параметров по формулам (3.18) и (3.19)
19:                 Params←ParamsEstimation(Z,Y);
    // На основе суммы квадратов разностей подряд идущих пар
    // элементов векторов средних, дисперсий и весов
20:                 δ += Distance(Params);
21:                 j++;
22:             r++;
23:     return Params;
```

Необходимо отметить, что использование величины $\mathrm{EM}_{distance}$ подхо-

дит только для случая модельных данных, когда известно распределение случайной величины Z. Для реальных рядов соответствующая процедура в алгоритме 3.3 должна быть исключена.

Оценки неизвестных параметров распределения (3.15), также как и предоценки параметров распределения шума (3.13), в алгоритме 3.3 определяются с помощью некоторой модификации EM-алгоритма.

3.2.7 Статистический эксперимент

Процедура, описанная в алгоритме 3.3, реализована на языке программирования `Python 3.7.3` с использованием библиотек `scikit-learn`, `NumPy` и `pandas`. Использована реализация EM-алгоритма из `scikit-learn` со следующими настройками: 15 запусков для каждой выборки с максимальным количеством итераций, равным 1500, и точностью 10^{-8}.

Таблица 3.3. Оцененные параметры сигнала и ошибки аппроксимации математических ожиданий (Err_{Exp}), дисперсий (Err_{Var}) и весов (Err_W)

Оцененные параметры сигнала X	$\mathbf{Err_{Exp}}$	$\mathbf{Err_{Var}}$	$\mathbf{Err_W}$
[1,0]; [6,79]; [1,31]	0,034	0,026	0
[1,0]; [12,53]; [0,0]	0,013	0,149	0
[1,0]; [89,12]; [2,95]	0,081	0,042	0
[1,0]; [5,04]; [1,52]	0,042	0,054	0
[1,0]; [15,83]; [0,23]	0,005	0,040	0
[1,0]; [45,15]; [1,19]	0,279	0,046	0
[0,3; 0,7]; [9,06; 5,89]; [2,31; 0,76]	0,053	0,071	0,008
[0,37; 0,63]; [19,9; 5,7]; [0,76; 2,14]	0,037	0,128	0,020
[0,5; 0,5]; [91,91; 63,55]; [1,07; 2,67]	0,071	0,040	0,003
[0,44; 0,56]; [−9,6; 7,51]; [0,43; 1,95]	0,089	0,004	0,007
[0,78; 0,22]; [29,74; 26,07]; [1,16; 0,24]	0,094	0,078	0,036
[0,4; 0,6]; [55,36; 77,33]; [1,86; 2,34]	0,357	0,078	0,027
[0,35; 0,53; 0,12]; [1,16; −4,25; 7,72]; [2,23; 1,86; 1,26]	0,279	0,171	0,052
[0,51; 0,39; 0,11]; [16,13; 23,59; 10,0]; [1,07; 2,85; 1,3]	0,098	0,062	0,005
[0,38; 0,27; 0,35]; [63,52; 48,78; 72,7]; [2,08; 1,59; 2,82]	0,114	0,049	0,018
[0,24; 0,45; 0,31]; [8,03; 0,05; −5,71]; [1,86; 1,22; 0,23]	0,236	0,261	0,005
[0,37; 0,26; 0,36]; [7,39; 23,21; 16,3]; [1,24; 2,55; 1,1]	0,477	1,067	0,110
[0,41; 0,13; 0,47]; [74,08; 42,86; 68,63]; [0,0; 0,27; 0,0]	0,274	0,426	0,017
[0,07; 0,23; 0,65; 0,04]; [0,86; −4,6; 2,24; 7,27]; [0,0; 2,38; 2,23; 0,84]	0,647	0,802	0,206
[0,36; 0,38; 0,13; 0,13]; [24,66; 6,92; 28,43; 18,68]; [2,58; 2,7; 0,09; 2,44]	0,212	0,128	0,008
[0,3; 0,39; 0,23; 0,07]; [69,53; 91,78; 59,17; 86,09]; [2,12; 2,6; 1,01; 2,03]	0,127	0,147	0,020
[0,47; 0,33; 0,02; 0,17]; [−8,88; −2,34; 0,14; 6,42]; [0,53; 0,58; 0,0; 0,54]	0,925	1,409	0,162
[0,25; 0,28; 0,2; 0,26]; [7,45; 17,15; 12,56; 26,74]; [0,56; 0,49; 0,45; 1,8]	0,487	0,676	0,054
[0,28; 0,28; 0,19; 0,25]; [73,84; 38,61; 48,02; 62,8]; [0,36; 0,42; 0,11; 0,75]	0,042	0,075	0,015

В таблице 3.3 приведены результаты полученных оценок для полезно-

го сигнала (истинные значения приведены в таблице 3.1, третий столбец), а также величины ошибок для каждого из параметров (рассчитывается на основе метрики RMSE).

В виде графиков результаты представлены на рисунке 3.5. В абсолютном большинстве случаев ошибка оценки для каждого из параметров не превосходит 1 (исходные наблюдения не нормировались). Исключение составляют дисперсии выборок с номерами 17 и 22. Во всех остальных случаях ошибка является весьма умеренной.

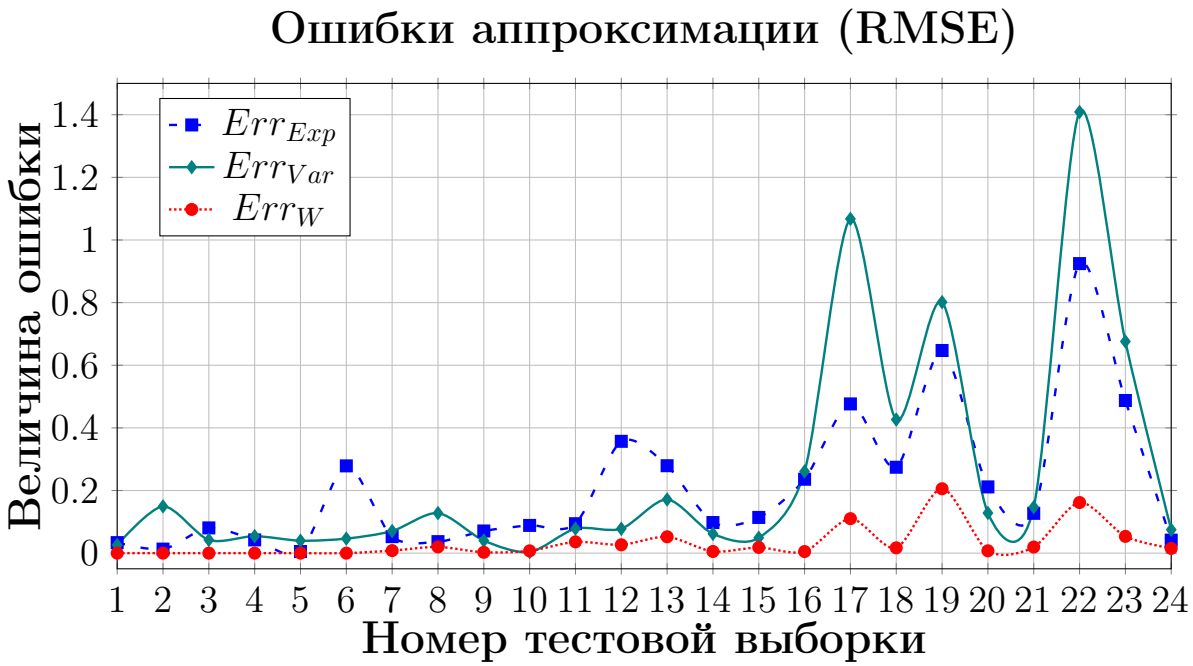

Рис. 3.5. Величины вычислительных ошибок для математических ожиданий (Err_{Exp}), дисперсий (Err_{Var}) и весов (Err_W)

При этом тестовые данные не подвергались предварительной нормировке. Полученные результаты означают, что метод может быть корректно использован для анализа реальных данных, какими бы ни были параметры распределений полезного сигнала и шума в нем. При этом, безусловно, предполагается, что в экспериментах используется достаточно точное измерительное оборудование для того, чтобы шум был не слишком «сильным» по сравнению с сигналом.

Необходимо отметить, что корректность оценок сигнала неразрывным образом связана с качеством работы вычислительной процедуры получения оценок максимального правдоподобия, то есть со сходимостью EM-алгоритма. Поэтому для минимизации влияния данного метода на результаты оценки предложенного адаптивного метода использовалась дополнительная проверка на близость параметров распределения модельной случайной величины и соответствующих аппроксимаций

(также в терминах стандартной метрики ℓ_2). На практике можно организовывать множество однотипных расчетов и сравнивать их результаты, либо выбирать в качестве итогового усредненное по всем запускам значение. В рамках экспериментальных исследований было установлено, что увеличение числа компонент в аппроксимирующем распределении шума существенно снижает точность итоговых расчетов, при этом метод менее чувствителен к увеличению числа компонент в распределении содержательного сигнала.

3.3 Метод определения связности локальных компонент смесей

В данном разделе предложены процедуры, позволяющие автоматизировать процесс выделения связанных компонент в методе скользящего разделения смесей. Развитие подобных методов востребовано для анализа явлений в физике турбулентной плазмы и океанологии (см. разделы 5.2 и 6.5) с целью определения количества формирующих процессов, а также оценивания распределений параметров стохастических дифференциальных уравнений, являющихся моделями соответствующих явлений.

В процессе шагов СРС-метода выделяются компоненты волатильности, которые эволюционируют при сдвигах скользящего окна. При этом достаточно сложно судить о том, какая из оценок параметров соответствует тому или иному значению на предыдущем шаге. Обычно подобная взаимосвязь определяется визуально и, очевидно, является достаточно субъективной. Ниже предложен подход к автоматизации решения данной задачи на основе комбинации жадного алгоритма и методов кластеризации k- или c-средних [189, 196, 326, 392]. При этом на первом этапе определяется число кластеров для методов машинного обучения, которые используются непосредственно для выявления связанных компонент волатильности.

Обозначим через $I^{(n)}$ набор индексов (номеров) компонент для шага с номером n СРС-метода, то есть $I^{(n)} = \{1, 2, \ldots, k^{(n)}\}$, а через $J^{(n+1)} = \{1, 2, \ldots, k^{(n+1)}\}$ аналогичный набор для шага $(n+1)$. Через I_0 и J_0 обозначим множество индексов из первого и второго наборов соответственно, для которых удалось найти ближайшую компоненту. Первоначально полагаем, что $I_0 = \emptyset$ и $J_0 = \emptyset$. Для каждого фиксированного

$J \in J^{(n+1)} \setminus J_0$ находим наиболее близкий номер I в смысле решения следующей оптимизационной задачи:

$$I = \underset{i \in I^{(n)} \setminus I_0}{\arg\min} \left(\left| a_i^{(n)} - a_J^{(n+1)} \right|^p + \left| \sigma_i^{(n)} - \sigma_J^{(n+1)} \right|^p \right)^{1/p}. \tag{3.27}$$

В этом случае минимизируемое в правой части выражение представляет собой ℓ^p-норму ($p = 1, 2, \ldots$) соответствующего вектора разностей координат в пространстве точек (a, σ).

Полагая после этого $I_0 = I_0 \cup I$ и $J_0 = J_0 \cup J$, необходимо повторить процедуру заново. Возможны следующие случаи:

1. $\left|I^{(n)}\right| = \left|J^{(n+1)}\right|$, то есть $k^{(n)} = k^{(n+1)}$. В этом случае оба набора будут исчерпаны одновременно;
2. $\left|I^{(n)}\right| < \left|J^{(n+1)}\right|$, то есть $k^{(n)} < k^{(n+1)}$. В этом случае процедура останавливается, когда исчерпан набор $I^{(n)} \setminus I_0$. Оставшиеся в $J^{(n+1)} \setminus J_0$ элементы формируют новые компоненты, которые появились только на $(n+1)$-ом шаге.

Случай $\left|I^{(n)}\right| > \left|J^{(n+1)}\right|$, то есть $k^{(n)} > k^{(n+1)}$, не рассматривается, поскольку основная цель данной процедуры – определение числа компонент за весь период эволюции процесса, поэтому уменьшаться оно не может, даже если на каком-то шаге произошло сокращение локального значения для числа компонент. Отметим, что указанная процедура, очевидно, является жадной. При этом, поскольку ее конечная цель состоит в определении числа кластеров для следующего шага, данная особенность не представляется критической.

Для точного определения числа компонент необходимо задавать некоторый допустимый порог близости $\varepsilon(\mathbf{a}, \boldsymbol{\sigma})$:

$$\left(\left| a_I^{(n)} - a_J^{(n+1)} \right|^p + \left| \sigma_I^{(n)} - \sigma_J^{(n+1)} \right|^p + \left| p_I^{(n)} - p_J^{(n+1)} \right|^p \right)^{1/p} < \varepsilon(\mathbf{a}, \boldsymbol{\sigma}). \tag{3.28}$$

Такая проверка нужна для того, чтобы корректно определять ситуацию, при которой компоненты с номерами I и J на n-м и $(n+1)$-м шагах считались одинаковыми, и не было необходимости создавать новую в рамках жадного алгоритма. Реализация метода определения числа компонент приведена в алгоритме 3.4.

Предполагается, что данный алгоритм применяется для компонент с положительными весами (нулевые значения соответствуют случаю уменьшению их числа на каком-либо шаге). Кроме того, для всех допустимых значений $i \neq j$ и n должны существовать такие $\delta_a > 0$

Алгоритм 3.4. Динамическое определение числа локальных компонент

```
1:  function NumGreedy(Params, I^(n), J^(n+1))
2:      I_0 ← ∅, J_0 ← ∅, Comps← ∅;                    // Инициализация
3:      repeat                  // Продолжающиеся или новые компоненты
4:          // Оптимизация выражения (3.27) с учетом условия (3.28)
5:          [I, J]←FindI(Params, J^(n+1) \ J_0, I^(n) \ I_0);
6:          if I ≠ ∅ then      // Найдена предшествующая J компонента
7:              I_0 ← I_0 ∪ I;
8:              J_0 ← J_0 ∪ J;
9:          else                                  // Добавление новой компоненты
10:             J_0 ← J_0 ∪ J, Comps ← AddNewComp(Params, J);
11:     until (J^(n+1) \ J_0 ≠ ∅)
12:     return Comps;
```

и $\delta_\sigma > 0$, что выполнено хотя бы одно из условий $\left|a_i^{(n)} - a_j^{(n)}\right| > \delta_a$, $\left|\sigma_i^{(n)} - \sigma_j^{(n)}\right| > \delta_\sigma$. Если же они оба нарушены, то необходимо объединять эти компоненты в одну с соответствующей корректировкой (суммированием) весов. То есть предполагается, что все компоненты различны, – это гарантирует корректность применения жадного алгоритма.

Алгоритм 3.4 используется в качестве важной составной части метода формирования матрицы связности. Она представляет собой вспомогательную структуру, в которую на каждом шаге скользящего окна сохраняется актуальное состояние всех выделенных к текущему моменту компонент. Сначала ко всему ряду применяется метод EM-типа для определения числа компонент на каждом шаге (как было отмечено выше, оно не может убывать). Затем в двухмерном пространстве $(\mathbf{a}, \boldsymbol{\sigma})$ используется один из методов кластеризации с полученным жадным алгоритмом числом кластеров. Веса не учитываются, так как вклад компоненты в смесь может изменяться, при этом параметры – математическое ожидание и дисперсия варьируются не слишком сильно и тогда считается, что это та же самая компонента. Соответствующая процедура представлена в алгоритме 3.5.

В разделах 5.2.2 и 6.5.3 будут приведены примеры его применения для определения числа формирующих процессов в плазме и статистического оценивания параметров в уравнении Ланжевена.

Очевидно, данная процедура может быть использована и в случае смесей многомерных распределений. При этом в формулу оптимиза-

Алгоритм 3.5. Определение компонент связности в СРС-методе

```
1: function MSMComponents(Data, options)
2:     Params ←EMs(Data, options.EM)                    // СРС-метод
3:     // Инициализация числом ненулевых компонент на первом шаге
4:     Comps^(1)←Params.k^(1);
5:     for (n = 1:length(Params)-1) do
6:         Comps^(n+1) ← NumGreedy(Params, I^(n), J^(n+1));
7:     // Метки для каждого набора параметров, кластеризация
8:     Labels ← Clust(Params, max(Comps), options.ClustAlg);
9:     // Матрица связности для компонент СРС-метода
10:    HistMatrConnect← Connectivity(Params, Labels);
11:    PlotComp(HistMatrConnect);      // Визуализация результатов
12:    return HistMatrConnect;
```

ции (3.27) должны быть добавлены все параметры соответствующего распределения (с соответствующей модификацией условия (3.28)), а затем может быть проведена кластеризация в пространстве переменных новой размерности.

3.4 Детектирование событий на основе анализа динамической компоненты

В разделе 3.2 для определения момента разладки был использован двухпроходный метод проверки однородности, который позволил достаточно точно определять точку изменения распределения в данных. В этом разделе аналогичный подход будет развит в целом для детектирования событий с использованием построенной в рамках СРС-метода динамической компоненты волатильности.При этом существенным образом используется отмеченная в разделе 2.1 возможность соотнести данный объект с локальными трендами процесса.

3.4.1 Методология

Идея данной процедуры заключается в следующем. Сначала с помощью СРС-метода определяются динамическая и диффузионная компоненты волатильности (см. раздел 2.1) для приращений исходного ряда. При этом в качестве метода получения оценок максимального правдопо-

добия используется сглаженная модификация EM-алгоритма. Основная ее особенность – использование на каждом следующем шаге CPC-метода в качестве начального приближения оценок, полученных на текущем этапе. Это позволяет строить максимально «гладкую» кривую для динамической и диффузионной компонент, избегая появления дополнительных оценок. Данный метод позволяет выявить основную компоненту с максимальным весом.

Для определения порогового значения, которое будет использовано для выявления событий, можно воспользоваться разумным предположением о том, что в начале и конце наблюдаемой выборки есть области достаточно большого размера, которые могут использоваться для настройки метода. При этом можно ожидать, что их распределение не будет унимодальным, а значит, для разностей (приращений) корректно использовать аппроксимацию конечными нормальными смесями вида (1.7) со стандартными ограничениями (1.8). Это позволяет учесть ситуацию, когда нельзя корректно воспользоваться правилом w сигма [113], как в случае унимодального нормального распределения с параметрами a и σ:

$$\mathbb{P}\left(|X - a| < w\sigma\right) = \sqrt{\frac{2}{\pi}} \int\limits_{0}^{w} e^{-t^2/2}\, dt,$$

которое позволило бы задавать достаточный уровень доверия и определять нужное значение величины w. Также границы могут быть выбраны эмпирически с помощью дополнительной информации об эксперименте или с помощью статистического подхода, основанного на использовании критерия χ^2.

Тогда возможные точки движения определяются пересечением установленного порога данными. Очевидно, что в данном подходе, также как и в случае проверки однородности в разделе 3.2, будет возникать временная задержка в определении произошедшего события. Проход в прямом и обратном направлении и сравнение результатов позволяют минимизировать возникающую ошибку. А именно, осуществляется группировка вероятных точек, в которые произошли события, определенные при проходах в обоих направлениях. Представляется целесообразным классифицировать группы с помощью метрики, связанной с размером окна. Например, все возможные точки, которые находятся внутри интервала, кратного размеру окна, должны быть классифицированы как единая группа. Тогда временем события признается среднее значение по группе. Описанная процедура представлена в виде псевдокода в алгоритме 3.6.

Алгоритм 3.6. Двухэтапный метод детектирования событий в данных на основе анализа динамической компоненты

```
1: function MovementsDetection(Data, window)
2:     // Data – исходные данные
3:     // window – ширина окна для CPC-метода
4:     DiffData ←diff(Data);                    // Переход к приращениям
5:     // CPC-метод со сглаженным EM-алгоритмом
6:     [DynComp, DiffComp] ← EMs(DiffData, window, 'smoothedEM');
7:     // Поиск событий в прямом и обратном направлениях
8:     MovForw ← Forward(DynComp, DiffComp);
9:     MovBack ← Backward(DynComp, DiffComp);
10:    Movements ← Detect(MovForw, MovBack, window);
11:    return Movements;
```

3.4.2 Пример: детектирование моментов активности головного мозга с использованием миограммы

В данном разделе рассмотрим пример применения описанной выше методологии для решения прикладной задачи неинвазивного определения областей активности в головном мозге с использованием данных миограммы – запись электрических сигналов, полученных в результате регистрации мышечных сокращений. Подобное направление исследований может быть весьма полезным для развития нейрокомпьютерных интерфейсов [162, 207, 214], а также для клинических исследований (подробнее см., например, введение в статье [236]).

Одним из наиболее популярных экспериментальных методов исследования активности мозга является так называемый метод вызванных потенциалов [199, 333]: испытуемый несколько раз совершает движение пальцем, что приводит к возникновению активности в мозге, которая регистрируется в эксперименте. При этом ключевой проблемой [79] становится обнаружение точек на миограмме, которые соответствуют началу движений, для синхронизации с данными магнитоэнцефалограммы (МЭГ), используемой для локализации различных областей мозга. Данный подход позволяет более точно определять датчики, наиболее близко расположенные с нужными областям активности.

Описанная ситуация является частным случаем так называемой обратной задачи нахождения источника сигнала по характеристикам поля, генерируемого источником. Одно из самых простых решений – найти

канал с лучшим откликом путем усреднения частей сигналов МЭГ по моментам начала движений, а затем использовать его для улучшения отношения сигнал/шум. Шум в этом случае представляет собой суперпозицию физического, создаваемого датчиками, усилителями, аналого-цифровыми преобразованиями, внешними сигналами, сетевыми помехами и т. п., и физиологического шума, отражающего фоновую активность мозга. При этом начальные точки не могут быть корректно определены по сигналам МЭГ из-за высокой доли шума в канале (0, 95 и более). Поэтому для локализации моментов начала движения используется миограмме. Кроме того, в эксперименте может быть задействовано дополнительное техническое решение (акселерометр, фотоэлемент) для дополнительной корректировки факта постукивания по поверхности добровольцем.

Для решения задачи определения моментов начала движений воспользуемся статистическим методом, описанным в предыдущем разделе (см. алгоритм 3.6). Динамическая компонента миограммы анализируется в прямом и обратном направлении, затем с применением критерия χ^2 ищутся экстремальные значения, которые и соответствуют моментам начала движения.

Миограмма используется в качестве исходных данных, при этом для исключения влияния трендов осуществляется переход к приращениям. В качестве размеров окна *window* (см. алгоритм 3.6) проверялись различные значения, например, 20, 30, 50 элементов. Ниже продемонстрированы результаты для последнего из них. Выше было отмечено, что распределение новой выборки – динамической компоненты, – полученной сглаженным EM-алгоритмом может не быть унимодальным. Данный факт продемонстрирован на рисунке 3.6 для подвыборки, соответствующей периоду времени, когда регистрирующее оборудование уже включено, а испытуемый еще не приступал к выполнению движений.

На рисунках 3.7 и 3.8 и нанесены момента начала движений, полученные методом оконной дисперсии [79] (вертикальные красные пунктирные линии) и алгоритмом 3.6 на основе скользящего разделения смесей (зеленые треугольники). Ось абсцисс соответствует времени эксперимента в миллисекундах, ось ординат демонстрирует соответствующие значения компонент. На верхних графиках на каждом из рисунков изображаются динамические, а на нижних – диффузионные компоненты. Рисунок 3.7 соответствует прямому проходу (см. функцию `Forward` алгоритма 3.6), а рисунок 3.8 – обратному (см. функцию `Backward`).

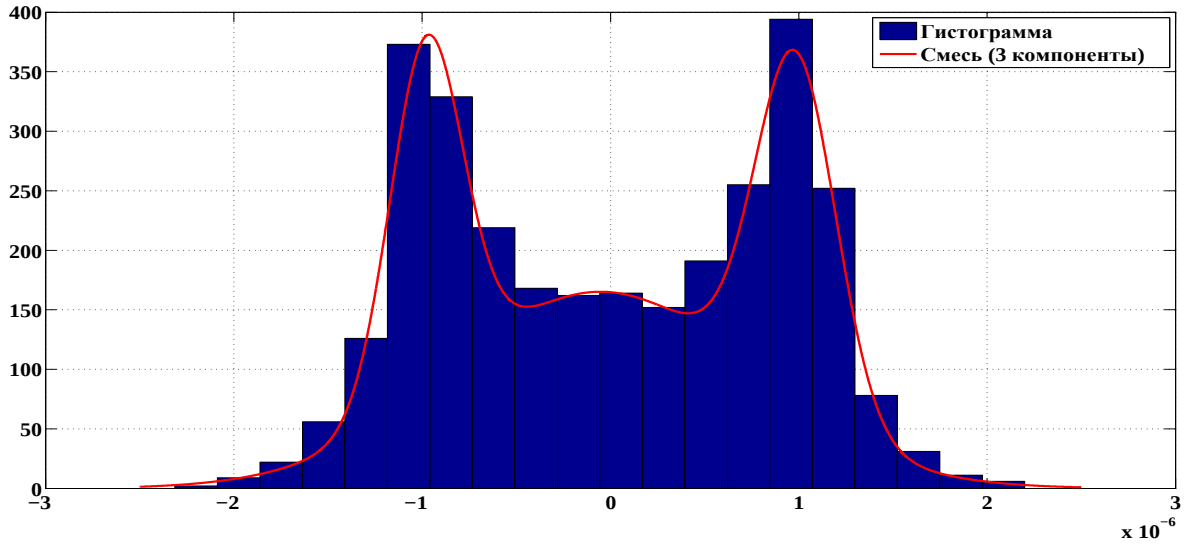

Рис. 3.6. Гистограмма для динамической компоненты волатильности для этапа до начала эксперимента с аппроксимирующей трехкомпонентной нормальной смесью

Можно выделить очевидные отличия в графиках для прямого и обратного проходов. А именно, форма и значения диффузионных компонент практически совпадают, за исключением начала и конца графиков, поскольку соответствующая область используется для настройки сглаженного EM-алгоритма при прямом и обратном проходах, соответственно. Аналогичный эффект наблюдается и для динамических компонент, которые, фактически, являются зеркальным отражением друг друга (в силу того, что это – математические ожидания). Полученные точки начала движений нанесены на график миограммы (см. рисунок 3.9). Отмечены истинные моменты начала движения (вертикальные красные пунктирные линии), а также решения алгоритма 3.6 (зеленые треугольники).

Очевидно, метод достаточно точно выявляет все действительно имевшие место события. По сравнению с методом оконной дисперсии, для которой была доступна дополнительная информация с фотоэлемента, регистрирующего движения пальца добровольца, метод скользящего разделения выявляет одно лишнее событие (в районе отметки 6200 мс, второй треугольник слева). Таким образом, точность предложенного метода достаточно высока и без использования дополнительной информации. Методы на основе сеточных алгоритмов для оценивания параметров обобщенных гиперболических или дисперсионно-сдвиговых распределений [236] для исходных данных позволяют получить близкие результаты, однако являются более сложными с точки зрения алгоритмической реализации.

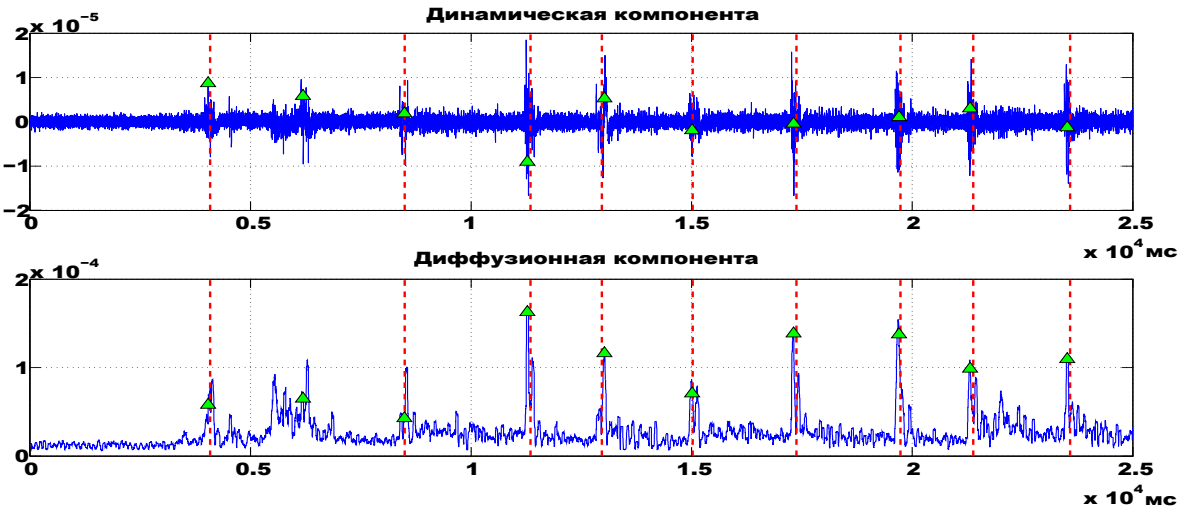

Рис. 3.7. Скользящее детектирование событий: прямое направление

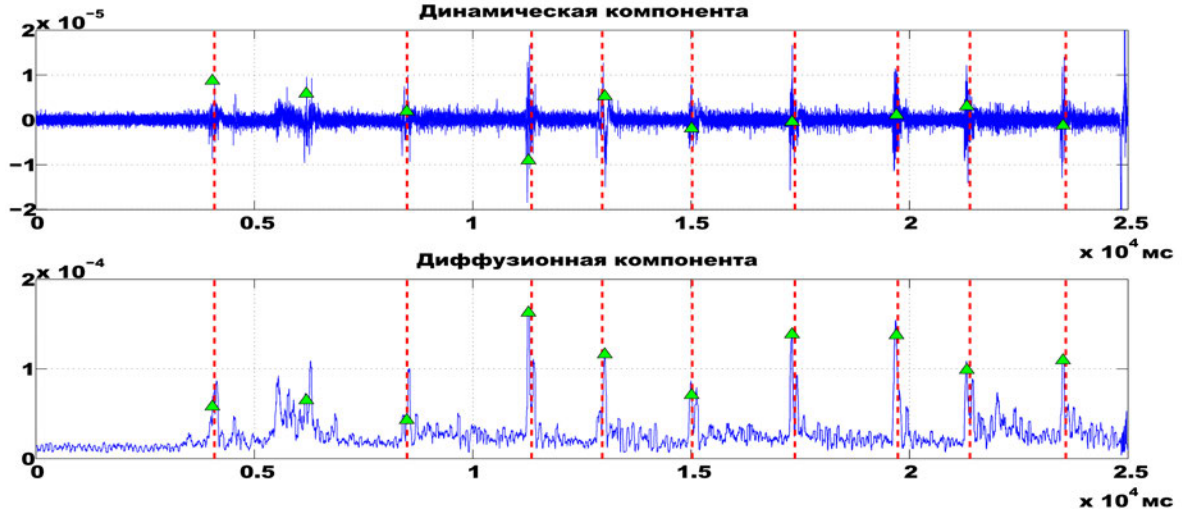

Рис. 3.8. Скользящее детектирования событий: обратное направление

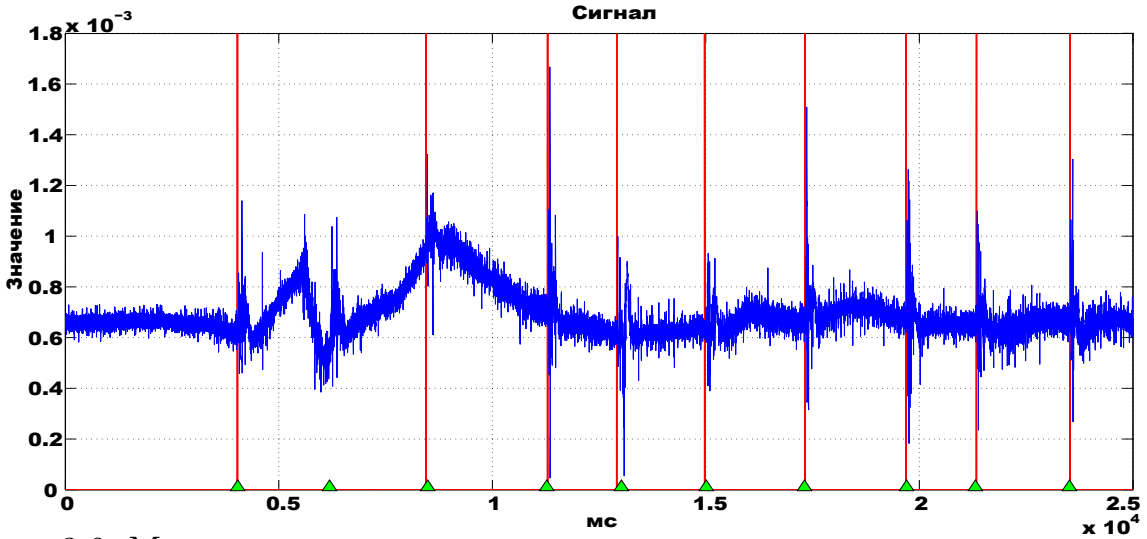

Рис. 3.9. Миограмма с известными моментами начала движения и решениями алгоритма

3.5 Метод искусственного зашумления для улучшения результатов СРС-анализа

Во введении упоминались результаты статьи [346], в которой продемонстрировано ускорение классического EM-алгоритм на 10–15% с помощью внесение искусственного шума в данные. В этом разделе будет предложен подход для повышения точности работы метода скользящего разделения конечных нормальных смесей за счет введения дополнительной аддитивной «шумовой» компоненты, имеющей нормальное распределение с заданными параметрами, а также некоторые подходы для их автоматического определения. Отметим, что предполагается независимость исходных данных и зашумляющей компоненты.

3.5.1 Методология

Предположим, что все наблюдения X_j в выборке имеют распределение типа конечных нормальных смесей вида (1.7) со стандартными ограничениями (1.8). Аддитивное зашумление означает замену исходных данных по следующему правилу:

$$\widetilde{X}_j = X_j + \varepsilon_j, \tag{3.29}$$

где $j = \overline{1, N}$, где N – объем выборки, $\varepsilon_j \sim \mathcal{N}(0, \sigma^2)$ – зашумляющая случайная величина с нормальным распределением с нулевым средним и некоторым среднеквадратическим отклонением σ, величина которого должна выбираться так, чтобы умеренно модифицировать первоначальную стохастическую структуру данных (например, 1% от выборочного среднеквадратического отклонения).

В терминах конечных нормальных смесей, подобное зашумление означает добавление новой компоненты с заданными средним и дисперсией и неизвестным весом. Преобразование выборки вида (3.29) не влияет на первое слагаемое в формуле (1.9) для дисперсии, в то время как второе изменяется следующим образом:

$$\sum_{i=1}^{k} p_i(\sigma_i^2 + \sigma^2) = \sum_{i=1}^{k} p_i\sigma_i^2 + \sigma^2,$$

так как с учетом независимости случайных величин X_j и ε_j можно использовать соотношение $\mathbb{D}(X_j + \varepsilon_j) = \sigma_j^2 + \sigma^2$. Подобное преобразование позволяет достаточно просто удалять дополнительную компоненту

из результатов СРС-анализа, поскольку требуется модификация только диффузионной компоненты (см. раздел 2.1), в то время как динамическая остается без изменений. Для выполнения условия для весов в (1.8) требуется их перенормировка (см. формулу (3.30) далее).

В то же время, поскольку СРС-метод основан на процедуре итерационных вычислений с помощью EM-алгоритма, то даже нулевое математическое ожидание может оцениваться не тривиальными значениями. Поэтому на практике требуется на каждом шаге t (положении окна) определить компоненту, параметры которой в некоторой метрике, например в ℓ^1, близки к заданным значениям $(0, \sigma)$ и удалить ее, пересчитав веса содержательных компонент по формуле:

$$p_i^{(t)} = \widetilde{p}_i^{(t)} \left(1 - \widetilde{p}^{(t)}\right)^{-1}, \tag{3.30}$$

где $\widetilde{p}^{(t)}$ и $\widetilde{p}_i^{(t)}$ – оцененные веса зашумляющей и содержательных компонент на шаге t, $i = \overline{1,k}$. При этом сумма всех $\widetilde{p}_i^{(t)}$ (3.30) на каждом шаге, очевидно, равна 1. Описанная процедура приведена в алгоритме 3.7.

Алгоритм 3.7. Улучшение результатов СРС-метода за счет искусственного зашумления данных

```
1:  function ImprovedMSM(options)
2:      // Исходные данные и параметры шума
3:      [Data, Noise]←Input( );
4:      // Искусственное зашумление данных на основе формулы (3.29)
5:      NData←Noising(Data, Noise);
6:      // СРС-оценки для зашумленных данных
7:      NParams←EMs(NData, options.MSM, options.EM);
8:      // СРС-оценки для исходных данных на основе формулы (3.30)
9:      Params←Denoising(NParams);
10:     return Params;
```

Ниже рассмотрен ряд модельных примеров применения алгоритма 3.7 в нескольких предметных областях (метеорология, разработка программного обеспечения).

3.5.2 Модельный пример: суточные объемы осадков

В данном разделе рассмотрим применение СРС-метода на примере суточных объемов осадков в Элисте. Особенностью подобных наблюдений является то, что они являются неотрицательными, поэтому прямое

применение моделей типа конечных нормальных смесей с бесконечным носителем может быть не вполне корректным.

На рисунке 3.10 приведена часть исследуемого ряда. Красной сплошной линией изображаются исходные наблюдения, пунктирной голубой – модифицированная в соответствии с процедурой 3.7 и формулой (3.29) выборка. Отметим, что в данном случае использовано значение σ, равное 0,01 от выборочного среднеквадратического отклонения.

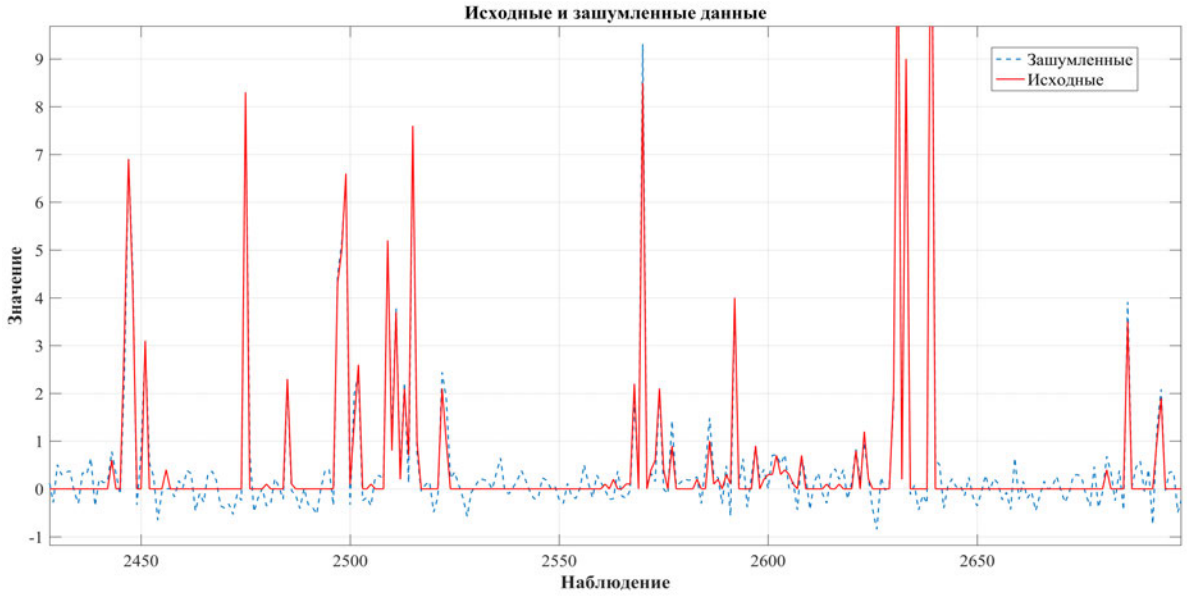

Рис. 3.10. Исходные и зашумленные данные для объемов осадков

Характерные пики в положительной полуплоскости не были потеряны, и в то же время появились отрицательные наблюдения, которые позволяют корректно использовать модели типа (1.7). На рисунке 3.11 приведены гистограммы для исходных (слева) и зашумленных (справа) данных.

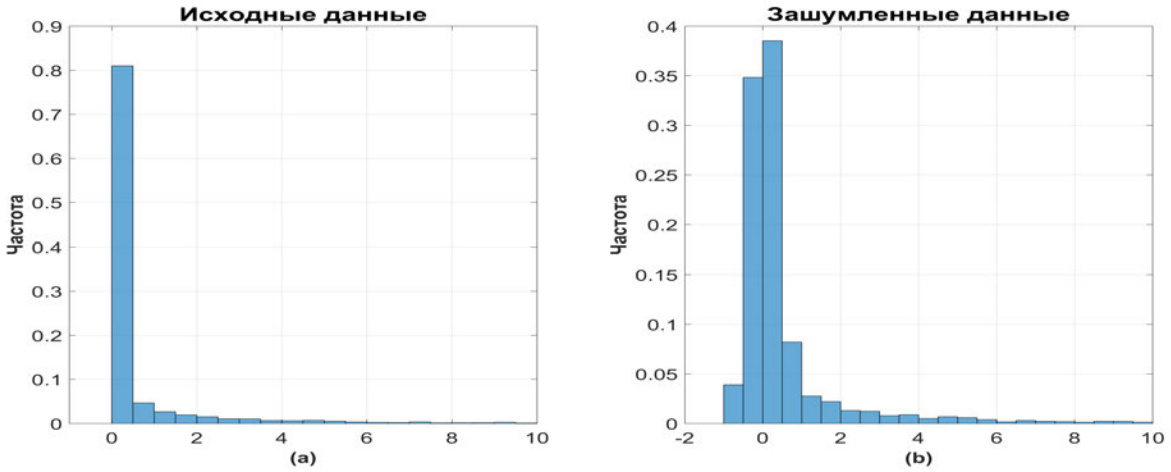

Рис. 3.11. Гистограммы для исходных и зашумленных данных

Очевидно, распределение изменилось, однако параметры шума являются не чрезмерными, и общий характер новой выборки можно считать близким к исходной.

Рассмотрим визуальное представление результатов формального СРС-анализа для неотрицательных исходных данных (см. рисунки 3.12 и 3.13).

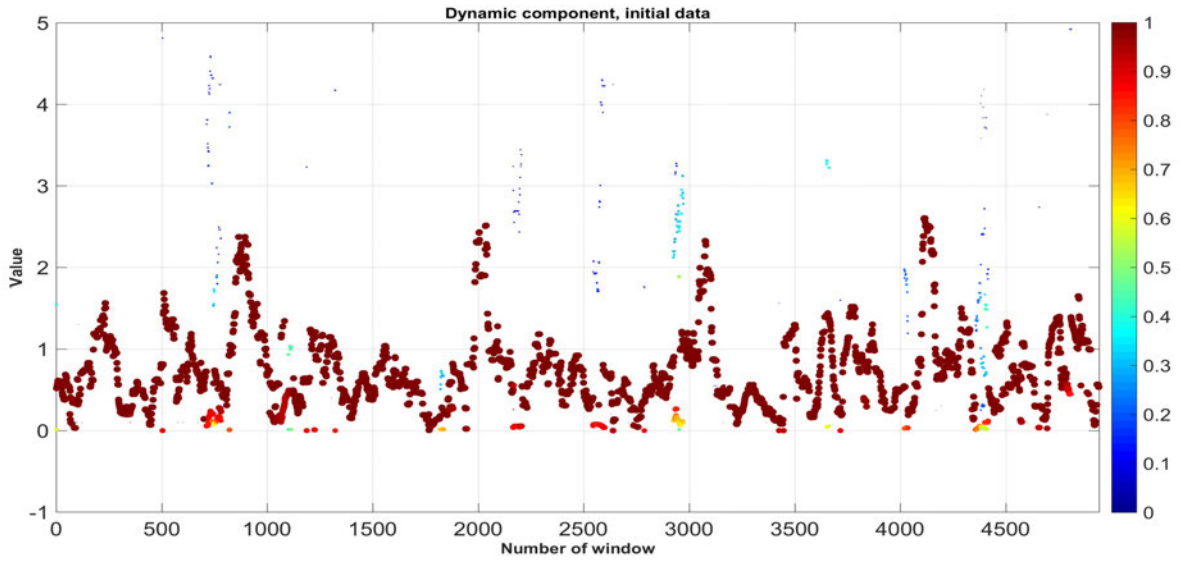

Рис. 3.12. Динамическая компонента, исходные данные

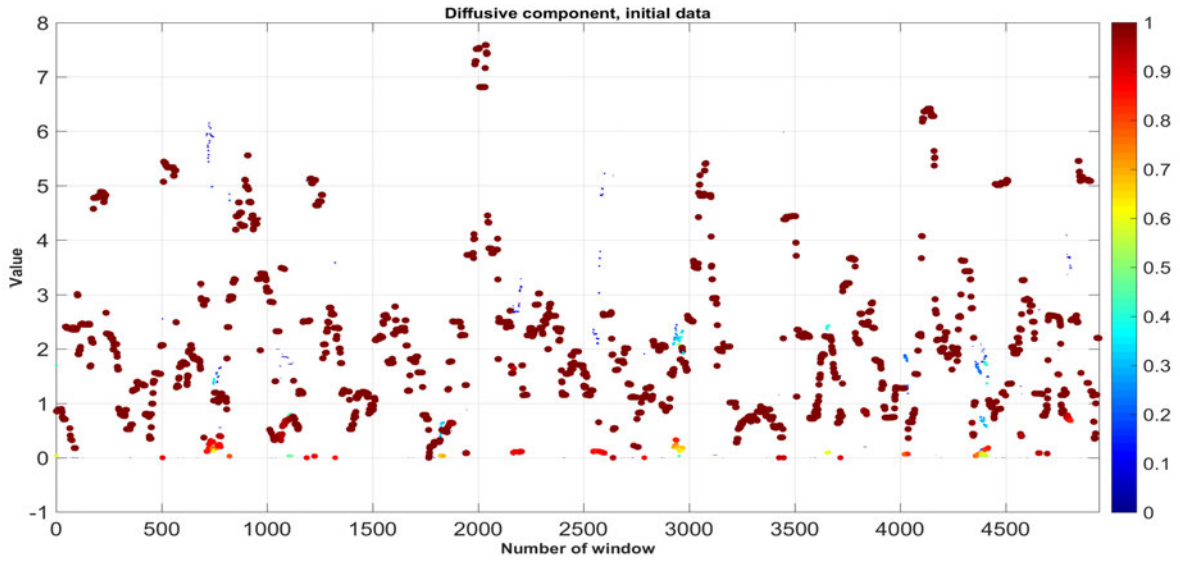

Рис. 3.13. Диффузионная компонента, исходные данные

По оси абсцисс отложены номера t, определяющие положение скользящего окна, а по оси ординат – математические ожидания и среднеквадратические отклонения, соответственно. Цветовая шкала справа показывает веса компонент. Кроме того, веса также задаются размером точек –

чем ближе к 1, тем больше данная точка, и наоборот. На графиках выделяется одна компонента с весом, близким к единице, отдельные отклонения от нее могу быть проинтерпретированы как шумы, возникающие в результате вычислительных погрешностей EM-алгоритма. Ситуация представляется тривиальной – и структурные составляющие процесса выделить невозможно.

Рассмотрим теперь результаты CPC-анализа для модифицированной выборки (см. рисунки 3.14 и 3.15), при этом шумовая компонента не удалена и веса содержательных компонент не пересчитаны согласно формуле (3.30). Данный шаг соответствует процедуре `EMs` алгоритма 3.7.

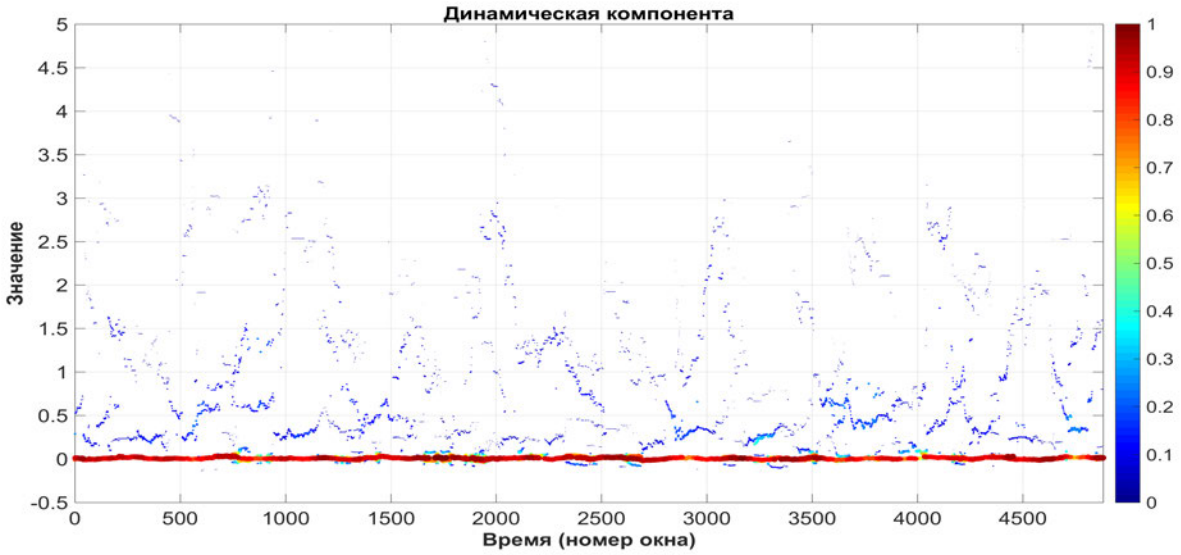

Рис. 3.14. Динамическая компонента модифицированной выборки

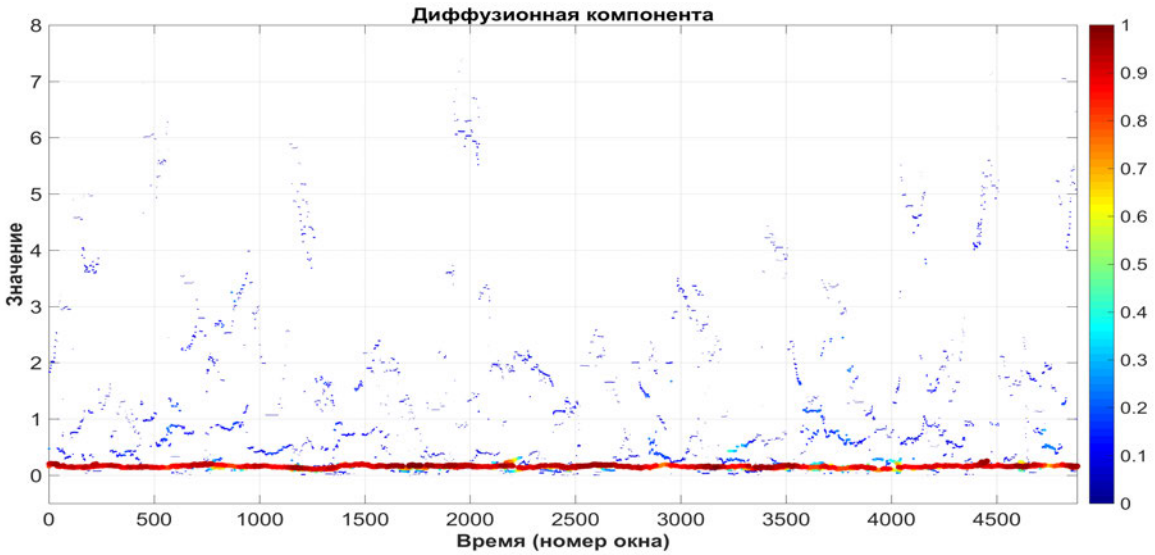

Рис. 3.15. Диффузионная компонента модифицированной выборки

На графиках четко видна искусственно внесенная шумовая компонента (яркая красная кривая). Как было отмечено ранее, ее математическое ожидание не оценивается тождественным нулем, точно также, как и ее среднеквадратическое отклонение флуктуирует около заданного значения σ. Очевидно, при удалении данная компонента должна быть идентифицирована и исключена. На графиках 3.16 и 3.17 представлены результаты исключения лишней компоненты.

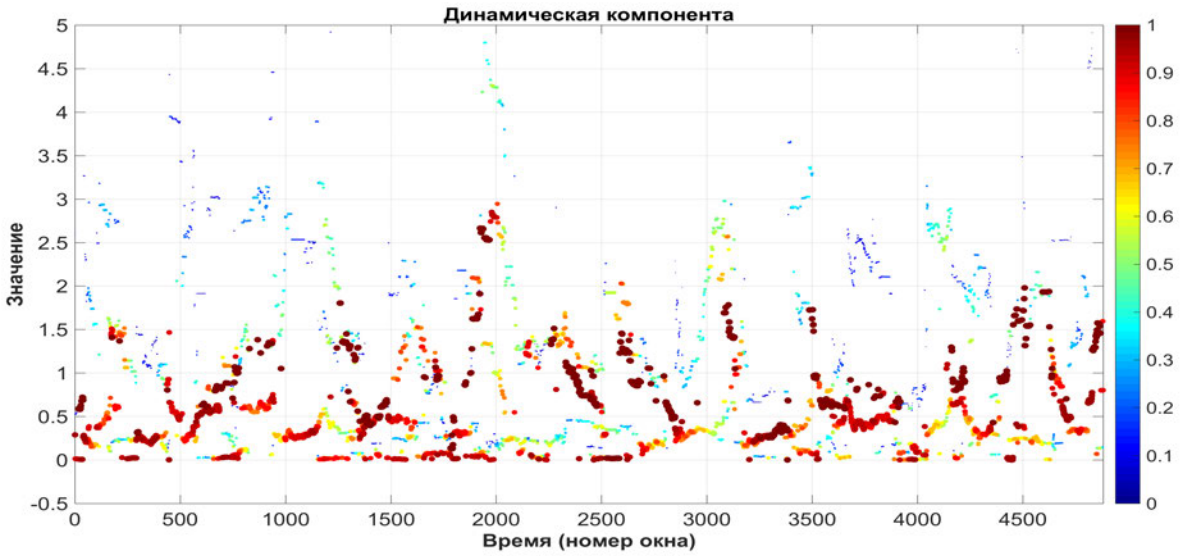

Рис. 3.16. Динамическая компонента исходных данных после удаления внесенного шума

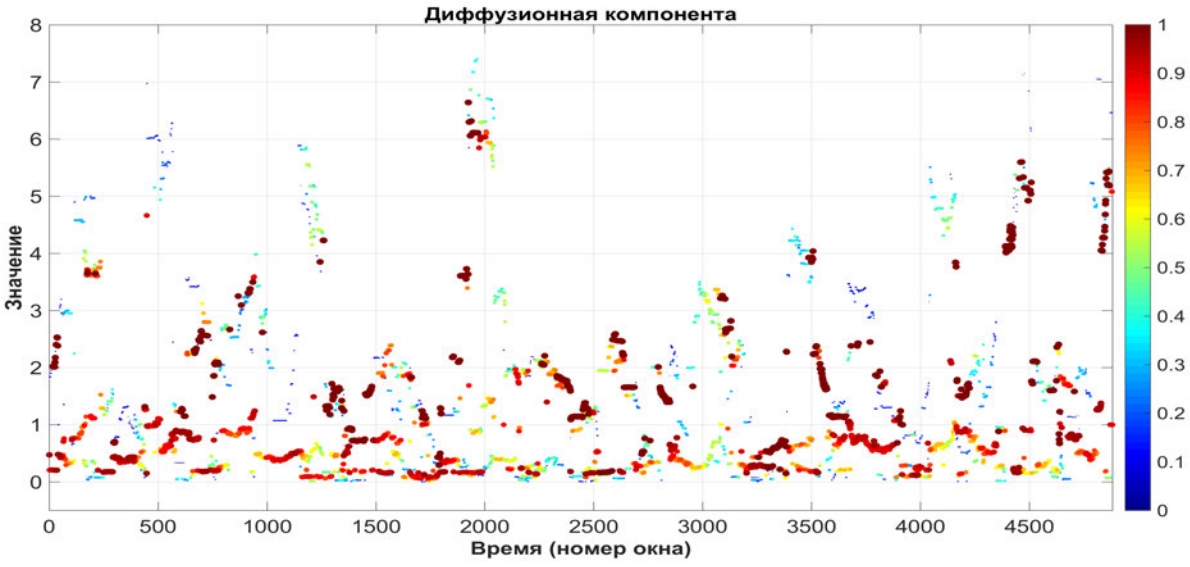

Рис. 3.17. Диффузионная компонента исходных данных после удаления внесенного шума

При этом также удаляются компоненты с отрицательными матема-

тическими ожиданиями, так как они отсутствуют у исходных данных, с пересчетом соответствующих весов. Данный шаг соответствует процедуре `Denoising` алгоритма 3.7. Во всех случаях в рамках процедуры `EMs` использовались трехкомпонентная смесь, скользящее окно размером 120 наблюдений и значение точности аппроксимации (2.5), равное 10^{-8}.

Алгоритм 3.7 может быть успешно использован и для анализа приращений (разностей) исходного ряда. Это продемонстрировано на рисунках 3.18–3.21.

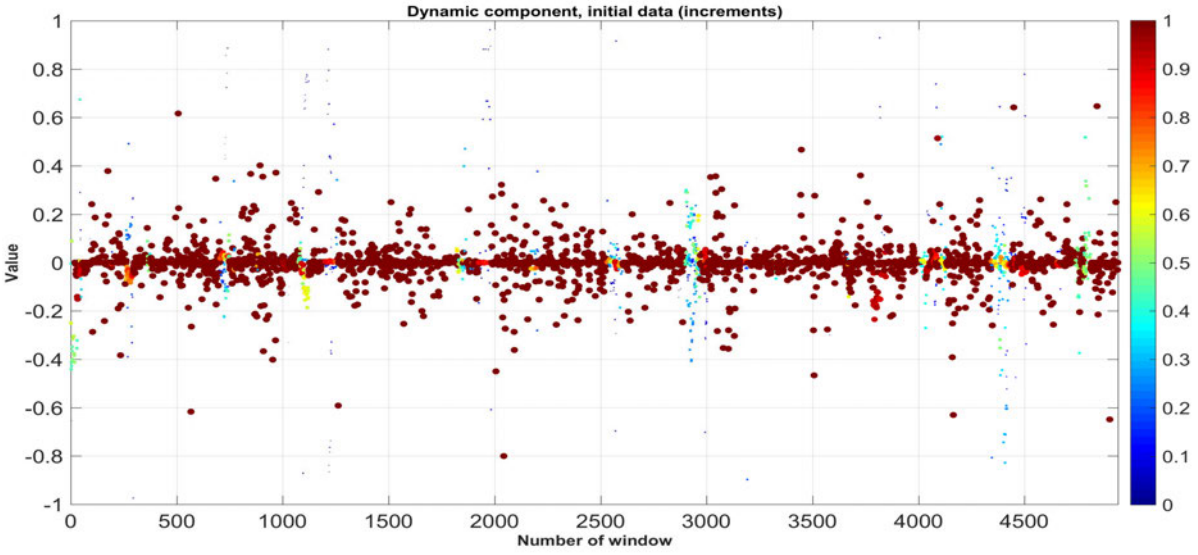

Рис. 3.18. Динамическая компонента для приращений исходных данных

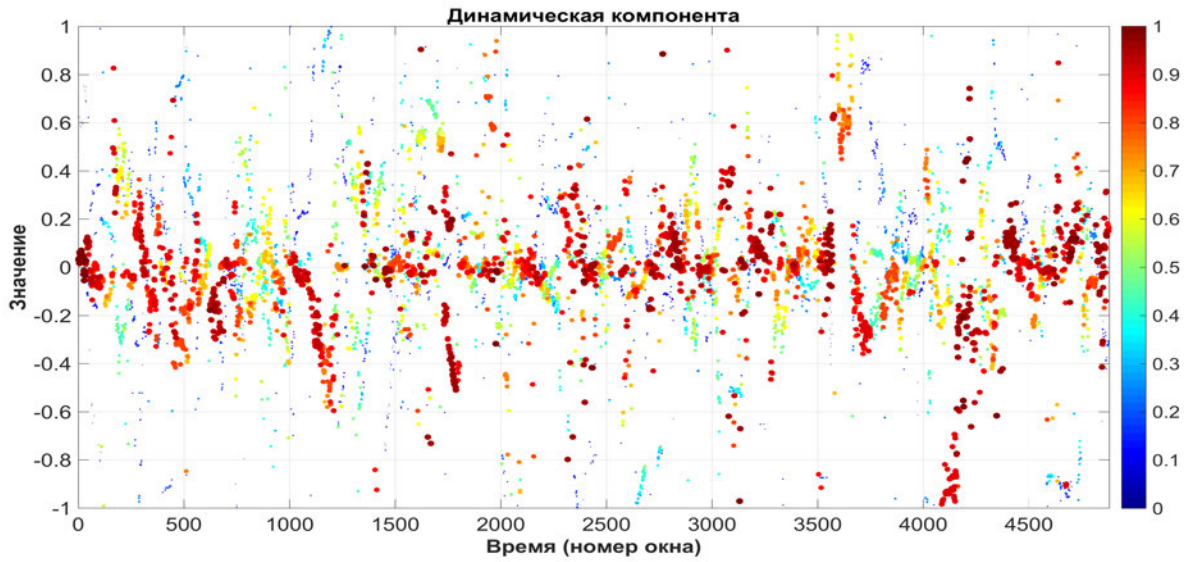

Рис. 3.19. Динамическая компонента для приращений после модификации исходных данных

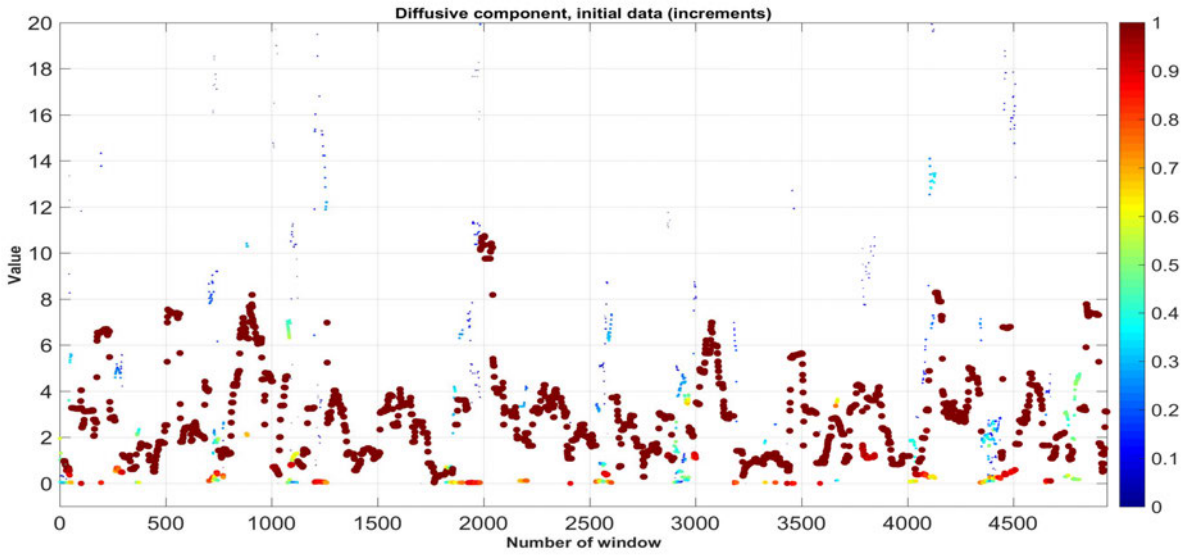

Рис. 3.20. Диффузионная компонента для приращений исходных данных

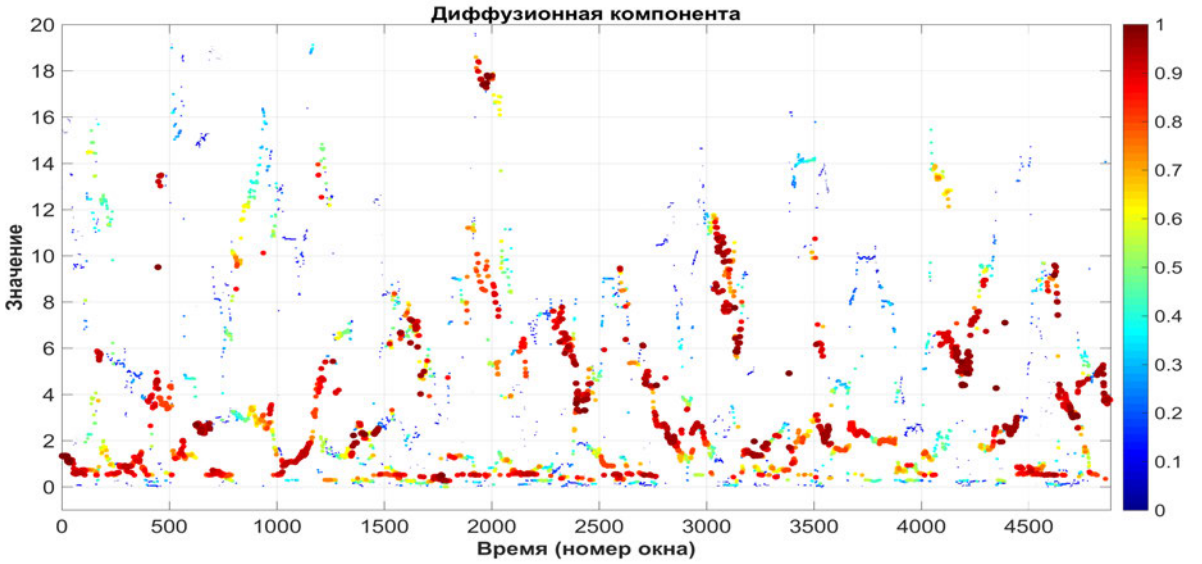

Рис. 3.21. Диффузионная компонента для приращений после модификации исходных данных

Очевидно, что результаты для модифицированной выборки несколько отличаются от того, что продемонстрировано для исходного ряда. Необходимо отметить, что в случае анализа первоначальных наблюдений это может быть связано с корректностью использования аппроксимаций с бесконечным носителем. В обоих случаях влияние на данный процесс оказывает выбор параметров зашумляющих случайных величин, поэтому при практическом применении необходимо уделять значительное внимание данному вопросу. Некоторые возможные подходы для решения данной задачи, основанные на результатах, полученных в главе 2, будут рассмотрены в разделе 3.5.4

3.5.3 Модельный пример: оценка производительности программного кода

Также указанный подход может оказаться полезным при анализе различных характеристик информационных систем, например, данных профилировщика, на основе методологии скользящего разделения смесей. Действительно, процесс разработки современного программного обеспечения требует внедрения различных интеллектуальных инструментов проектирования [205] для достижения надлежащего уровня качества конечного продукта. В частности, необходимо установить, что некоторый процесс выполняется в заданных временных рамках (эта проблема связана с проблемой времени выполнения и котролем накладных расходов [345]), например, с помощью профилировщика (см. рисунок 3.22).

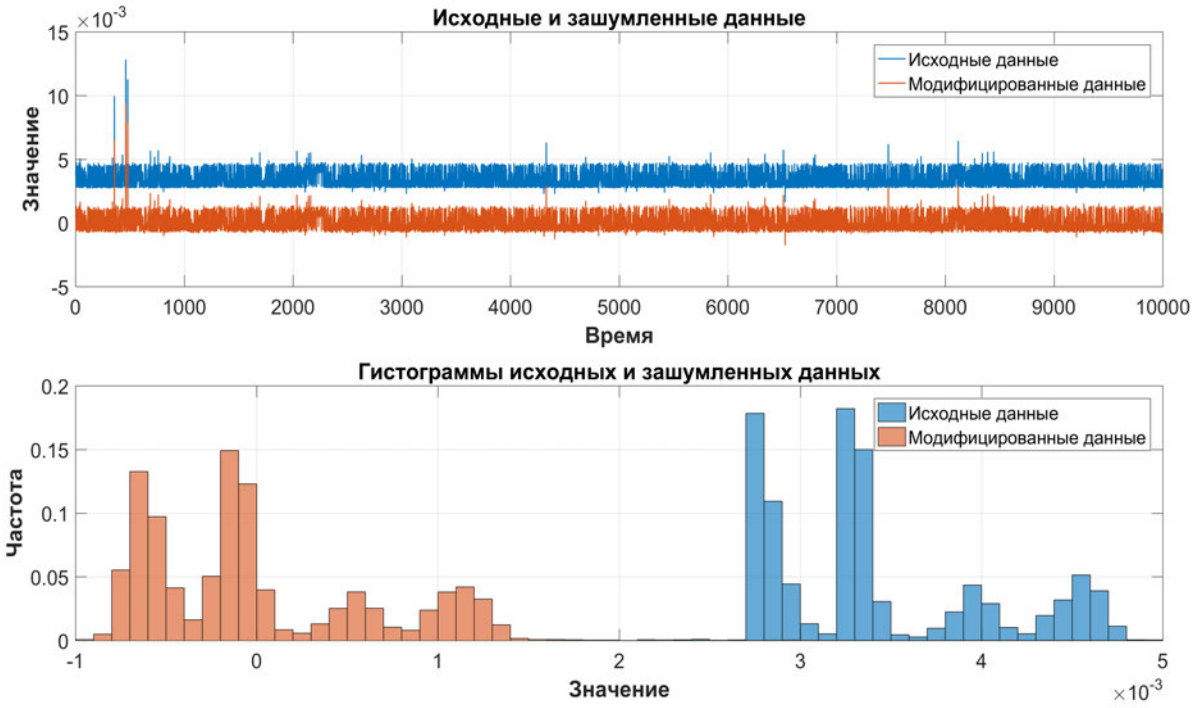

Рис. 3.22. Исходные и модифицированные для CPC-анализа данные профилировщика

Возникающие накладные расходы зачастую имеют вероятностный характер, поскольку определяются средой окружения программного решения, ошибками в кэша и т. д. Поэтому на этапе проектирования необходимо учитывать стохастических факторы, которые могут влиять на результаты профилирования – широко используемого инструмента анализа производительности системы. Подход на основе скользящего разделения конечных нормальных смесей позволяет проанализировать эволюцию неизвестного распределения вероятности времени выполнения для

различных компонентов программы, и затем использовать его для возможного управления накладными расходами.

На рисунке 3.22 (верхний график) приведены значения времени выполнения (по данным профилировщика `MATLAB`) для 10000 независимых запусков функции построения трехмерных поверхностей `surf`. Очевидно, что в данном случае все наблюдения будут строго положительны, поэтому для корректности применения аппроксимации в виде конечной нормальной смеси в данном модельном примере использована методология зашумления, описанная в разделе 3.5.1. Гистограммы на нижнем графике на рисунке 3.22 показывают, как при этом изменилось выборочное распределение наблюдений.

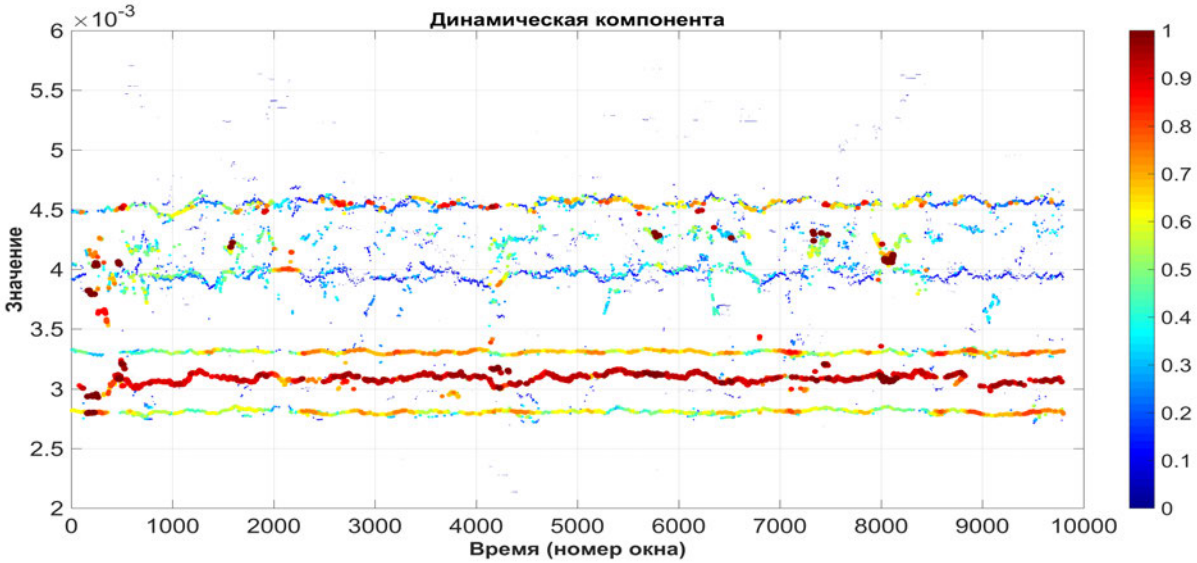

Рис. 3.23. Динамическая компонента после удаления шума

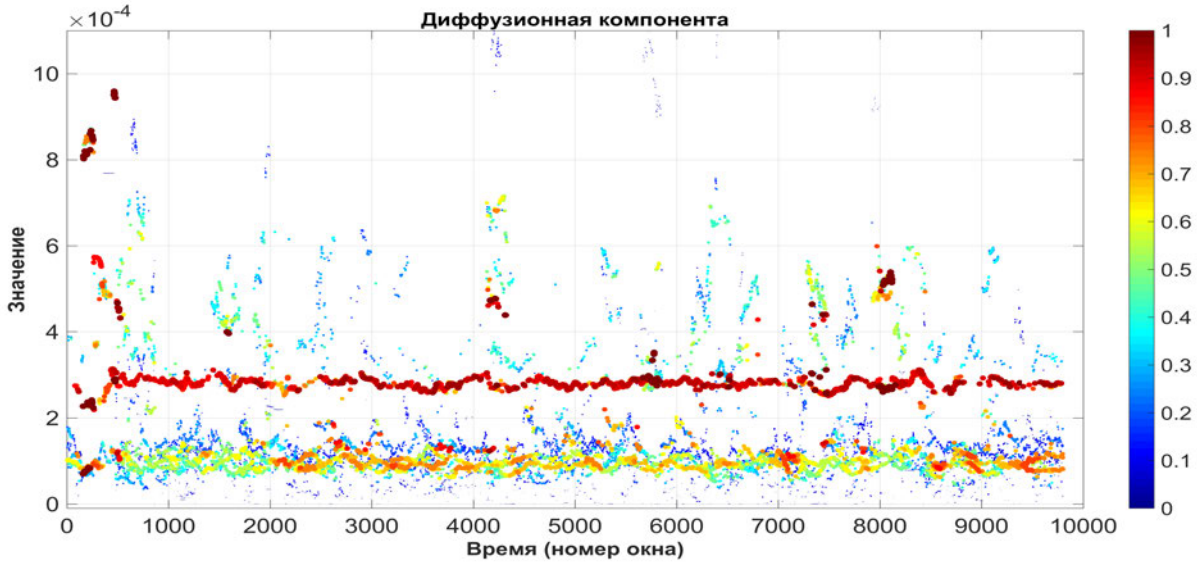

Рис. 3.24. Диффузионная компонента после удаления шума

Действуя согласно алгоритму 3.7, для зашумленных данных применен СРС-метод, получены оценки параметров аппроксимирующей четырехкомпонентной смеси, из которой удалены шумовая составляющая. Полученные графики динамической и диффузионной компонент представлены на рисунках 3.23 и 3.24, соответственно.

Приведенные примеры показывают, что предложенный метод на основе искусственного зашумления (см. алгоритм 3.7) может быть использован для анализа широкого класса информационных систем различной природы.

3.5.4 Подходы к определению параметров шума

Выше продемонстрировано повышение точности работы метода скользящего разделения конечных нормальных смесей за счет введения дополнительной компоненты, имеющей нормальное распределение $\mathcal{N}(0, \sigma^2)$ с математическим ожиданием, равным 0, и стандартным отклонением σ. При этом была отмечена сложность выбора параметра σ для сохранения структуры выборки, близкой к исходной. Для выбора параметров зашумляющего распределения можно воспользоваться теоремой 2.14 из раздела 2.6, положив $k = 1$, $a_j = 0$ для всех $j = 1, 2, \ldots$ и выбирая величину σ как минимизирующую длину доверительного интервала (2.64). Для этого необходимо найти производную функции $f(0, \sigma, \alpha, n)$ (2.66) и численно решить уравнение

$$f'_\sigma(0, \sigma, \alpha, n) \equiv \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}} - e^{-2\pi^2\sigma^2}\left(4\pi\sigma + \frac{1}{2\pi^3\sigma^3} + \frac{1}{\pi\sigma}\right) = 0 \qquad (3.31)$$

относительно неизвестного параметра σ при выбранных значениях величин n и α.

Таблица 3.4. Численные решения уравнений (3.31) и (3.32) относительно параметра σ для некоторых значений n и α

Размер выборки (n)	Уровень $\alpha = 0{,}1$		Уровень $\alpha = 0{,}05$		Уровень $\alpha = 0{,}01$	
	σ_1	σ_2	σ_1	σ_2	σ_1	σ_2
100	0,4302	0,435	0,419	0,425	0,4002	0,408
200	0,452	0,455	0,441	0,445	0,424	0,429
1 000	0,499	0,499	0,489	0,489	0,473	0,475
10 000	0,558	0,556	0,549	0,547	0,536	0,534
10 0000	0,611	0,607	0,603	0,599	0,591	0,588

В качестве альтернативы можно использовать вид доверительного интервала из статьи [411], полученный с помощью неравенства $\mathbb{D}[Z] \leqslant 2\mathbb{D}Z + \frac{1}{2}$, и искать решение уравнения вида

$$\frac{2\sigma z_{1-\frac{\alpha}{2}}}{\sqrt{n(2\sigma^2 + \frac{1}{2})}} - e^{-2\pi^2\sigma^2}\left(4\pi\sigma + \frac{1}{2\pi^3\sigma^3} + \frac{1}{\pi\sigma}\right) = 0. \tag{3.32}$$

Примеры найденных значений σ для типичных размеров выборок в CPC-методе (учитываются как возможная ширина окна, так и общее число наблюдений в анализируемом ряде) приведены в таблице 3.4. σ_1 и σ_2 – приближенные решение уравнений (3.31) и (3.32), соответственно. В качестве метода оптимизации использован `Trust-Region Dogleg` пакета `MATLAB` с настройками по умолчанию.

Глава 4

Логнормальные смеси как модели размеров частиц лунного реголита

Изучение закономерностей, определяющих размеры частиц, образующих лунные и другие планетарные реголиты, имеет очень большое значение при планировании автоматических и пилотируемых миссий для изучения космических тел (Луны, астероидов, планет и их спутников). В условиях малой гравитации и отсутствия плотной атмосферы пылевые структуры приобретают качества, не типичные для земных условий, представляя собой облака заряженных частиц, обладающих высокими абразивными свойствами. Они осаждаются на элементах аппаратуры (например, солнечных батареях) и скафандрах, что быстро выводит их из строя. Таким образом, решение рассматриваемой задачи может оказаться весьма полезным при подготовке новых космических миссий для повышения уровня безопасности и общей успешности подобных проектов.

В данной главе рассмотрена задача моделирования распределений размеров пылевых частиц лунного реголита, возникающих в результате различных воздействий, например при бомбардировках поверхности Луны метеоритами. При таких воздействиях развиваются как взрывные процессы разлета частиц с их дроблением, так и спекание частиц в экзотермических плазмохимических реакциях синтеза [106]. Как показано в книге [116], наблюдаемые статистические распределения размеров осажденных частиц имеют тяжелые степенные хвосты.

Изначально в рамках моделей формирования частиц в некоторых по-

родах рассматривался только процесс дробления. В частности, в работе [108] для рудных месторождений было отмечено, что размеры частиц имеют распределение, близкое к логнормальному, а затем А. Н. Колмогоровым предложена математическая модель процесса дробления частиц, аналитически объясняющая данный факт [84]. Данная модель основана на изучении изменения во времени числа частиц, размер которых не превосходит заданный порог, и справедлива при предположении, что скорость дробления является постоянной. Однако с уменьшением размера частицы интенсивность ее соударений с другими частицами может изменяться. Кроме того, например, для песка в естественной форме в книге [139] продемонстрировано, что характер распределения его частиц не является «чистым» логнормальным, и скорее имеет экспоненциально уменьшающиеся хвосты.

В статьях [365,387] в модель Колмогорова, по сути, вводится рандомизация, отражающая случайности времени, в течение которого наблюдается частица. Если считать, что время «жизни» частицы является экспоненциальной случайной величиной [365], то возникают модели размера на основе двустороннего Парето-логнормального распределения. Если же предположить, что соответствующее время описывается обратным гауссовским распределением [387], то в качестве модели для логарифмов размеров возникают NIG-распределения (normal-inverse Gaussian), являющихся представителями класса обобщенных гиперболических распределений [143]. Известно, что классические гиперболические распределения являются качественными аппроксимациями эмпирических распределений реальных геологических данных [267, 325, 416]. Однако во всех указанных случаях считается, что интенсивность процесса дробления остается постоянной.

В данной главе существенно используются результаты раздела 1.4, для учета случайной интенсивности процесса изменения размера в рамках эволюция отдельной частицы. Они служат математическим обоснованием моделей типа конечных логнормальных смесей в разработанных методах статистической обработки всех образцов лунного реголита, доставленных миссиями «Аполлон-11, 12, 14–17» и «Луна 24». На основе логнормальных смешанных моделей созданы алгоритмы статистического анализа с использованием бутстреп-процедуры и минимизации значений статистики χ^2 для аппроксимации эмпирических распределений размеров ансамблей пылевых частиц, возникающих в лунном реголите. Для всех 317 проб, представленных в каталоге `NASA` [252], будет про-

демонстрировано высокое статистическое согласие модели типа смеси конечных логнормальных законов, которые получаются в рамках применения указанных методов, и реальных данных, причем измеряемых как в классической метрической (микрометры, мкм), так и в принятой в геологии ϕ-шкалах [194]. Кроме того, обсуждаются вопросы выделения кластеров в полученных параметрических наборах, характеризующих аппроксимирующие распределения, для возможного соотнесения с химическим составом проб или иными характеристиками реголита.

4.1 Конечные логнормальные смеси как модели для аппроксимации распределений размеров частиц лунного реголита

Пусть частица лунного реголита в некоторый момент t_0, который для удобства положим равным 0, имеет начальный размер s_0. С течением времени частица подвергается различным воздействиям из-за нагрева, столкновений с метеоритами и т. п. Предположим, что после i-го преобразования размер частицы становится $D_i > 0$ по сравнению с тем, который был после $(i-1)$-го воздействия. Пусть $N(t)$ число преобразований частицы за время t. Тогда размер в момент t определяется выражением $Z(t) = s_0 \prod\limits_{i=1}^{N(t)} D_i$. Откуда следует, что

$$S(t) \equiv \log Z(t) = \log(s_0) + \sum_{i=1}^{N(t)} \log D_i = \mu + \sum_{i=1}^{N(t)} X_i. \tag{4.1}$$

Воспользуемся предельной схемой, описанной в разделе 1.4. Предположим, что вместо последовательности случайных величин $X_1, X_2, \ldots$ рассматривается схема серий $\{X_{n,j}\}_{j \geqslant 1}$ с некоторым формальным параметром $n \in \mathbb{N}$, с независимыми строками необязательно одинаково распределенных случайных величин. Вместо единственного считающего процесса $N(t)$ рассмотрим последовательность $\{N_n(t)\}$, поэтому базовая модель (4.1) преобразуется в $S_n(t) = \mu_n + S_{n,N_n(t)}$. Зафиксировав некоторый конечный момент времени $t = T > 0$ и полагая $N_n = N_n(T)$ для всех $n \geqslant 1$, получим, что $S_n(T) = S_n$ и

$$S_n = \mu_n + S_{n,N_n}. \tag{4.2}$$

Модель (4.2) допускает следующую интерпретацию. Для фиксированного n последовательность $X_{n,1}, \ldots, X_{n,N_n}$ описывает возможную траекторию процесса изменения размера частиц. Исследуется возможность асимптотического приближения распределения размера S_{n,N_n} (1.37) при большом N_n (то есть для большого значения момента времени T или с большой интенсивностью трансформации). Введя вспомогательный (бесконечно большой) параметр n, становится возможным рассмотреть схему, в которой допускается, что при модификации n распределения $X_{n,j}$ также изменяются. То есть вместо рассмотрения одной частицы рассматривается некоторый их ансамбль с длительной предысторией трансформаций. При этом каждая траектория начинается с собственного начального значения $s_{n,0}$, так что $\mu_n = \log s_{n,0}$.

Пусть как и в разделе 1.4 $\mu_{n,j} = \mathbb{E}X_{n,j}$, $\sigma^2_{n,j} = \mathbb{D}X_{n,j}$, причем $0 < \sigma^2_{n,j} < \infty$, $n, j \in \mathbb{N}$. В предшествующих работах рассматривался процесс чистая фрагментация (дробление), и предполагалось, что $0 \leqslant D_j \leqslant 1$, то есть $\mu = \mathbb{E}X_1 = \mathbb{E}\log D_i \leqslant 0$. Равенство μ нулю соответствовало случаю «вырожденной» фрагментации, при которой каждая частица оставалась неизменной. В рамках рассматриваемой схемы параметры $\mu_{n,j}$ могут быть как положительными, так и отрицательными: первый случай соответствует дроблению, второй – спеканию.

Предположение $0 < \sigma^2_{n,j} < \infty$ не является слишком ограничительным. С практической точки зрения это предположение фактически исключает из диапазона рассмотренных ситуаций только те, которые допускают либо мгновенное обнуление размера частицы, либо, наоборот, мгновенное увеличение размера маленькой частицы до очень больших (бесконечно больших) значений. В качестве еще одного аргумента в пользу предположения конечности второго момента случайных величин $X_{n,j}$ отметим, что устойчивые законы являются хорошо известными примерами распределений с очень тяжелыми хвостами, убывающими как степенные функции с малым показателем степени, однако даже их логарифмические моменты конечны.

Из теоремы 1.10 следует, что при весьма общих условиях приближение к распределению логарифма по размеру частиц следует искать в семействе нормальных сдвиг-масштабных смесей, а аппроксимации размеров самих частиц имеют лог-смешанное распределение указанного типа. То есть если Z – размер частицы, то

$$\mathbb{P}(Z < x) = \mathbb{P}(\log Z < \log x) \approx \mathbb{E}\Phi\left(\frac{\log x - V}{U}\right) \tag{4.3}$$

для некоторой пары случайных величин $(U, V) \in \mathcal{W}(Z|X)$ (см. раздел 1.4). Отметим, что распределение, стоящее в правой части (4.3) является сдвиг-масштабной смесью логнормальных законов. Из определения интеграла Лебега следует, что любая непрерывная смесь может быть приближена конечной, поэтому в качестве модели распределения размеров частиц в ϕ-шкале (то есть, фактически, логарифмов исходных размеров) будем использовать конечную смесь нормальных законов вида (1.7) со стандартными ограничениями (1.8), а именно:

$$\mathbb{P}(Z < x) = \sum_{i=1}^{k} p_i \Phi\Big(\frac{\log x - a_i}{\sigma_i}\Big). \tag{4.4}$$

4.2 Аппроксимации с помощью метода статистической симуляции выборок

В каталогах `NASA` [252] данные представлены в табличной форме в виде пар «размер частицы – доля (в процентах) частиц такого размера в просеиваемых образцах». Доступны сведения только о нескольких (как правило, не более десяти) точках роста (выбранных, вообще говоря, бессистемно) эмпирической функции распределения, но не о ее поведении между ними. В работе [252] описано применение интерполяции Стайнемана [393] для их соединения, благодаря которому с использованием макросов пакета `Excel` возможно построение соответствующих гистограмм. В данном разделе в аналогичных целях применяются кусочные кубические полиномы Эрмита [208], так как они позволили получить наиболее близкие кривые по сравнению с представленными в каталоге `NASA`.

Предложенная аппроксимация полиномами Эрмита позволила получить непрерывную эмпирическую функцию распределения (ECDF), а значит, стало возможным использовать метод обратных функций для генерации тестовых выборок. Объем выборки для оценивания параметров составлял 10 000 наблюдений, кроме того, для проверки соответствия приближающей смеси и исходной эмпирической функции распределения генерировалась еще одна независимая выборка объемом 2500 наблюдений (для проверки гипотез с помощью критерию согласия Колмогорова). Данная процедура в целом близка к такому статистическому подходу, как бутстреп. Ее описание приведено в алгоритме 4.1.

Полученные тестовые выборки используются для получения оценок максимального правдоподобия параметров аппроксимирующего смешан-

Алгоритм 4.1. Имитационное моделирование выборок для аппроксимации и статистического теста

```
1:  function GenSamples(ECDF, SampleSize, TestSampleSize)
2:      // Случайные векторы для метода обратных функций
3:      r←Rand(SampleSize);                  // Для оценивания параметров
4:      rTest←Rand(TestSampleSize);          // Для критерия Колмогорова
5:      // Имитационное моделирование, метод обратных функций
6:      for i=1:SampleSize do
7:          Sample(i)←fsolve(ECDF, r(i));
8:          if (i⩽TestSampleSize) then
9:              TestSample(i)←fsolve(ECDF, rTest(i));
10:     return [Sample, TestSample];
```

ного распределения (4.4) с помощью EM-алгоритма для нормальных законов. Примеры применения данной бутстреп-процедуры к реальным пробам лунного реголита представлены на рис. 4.1–4.9.

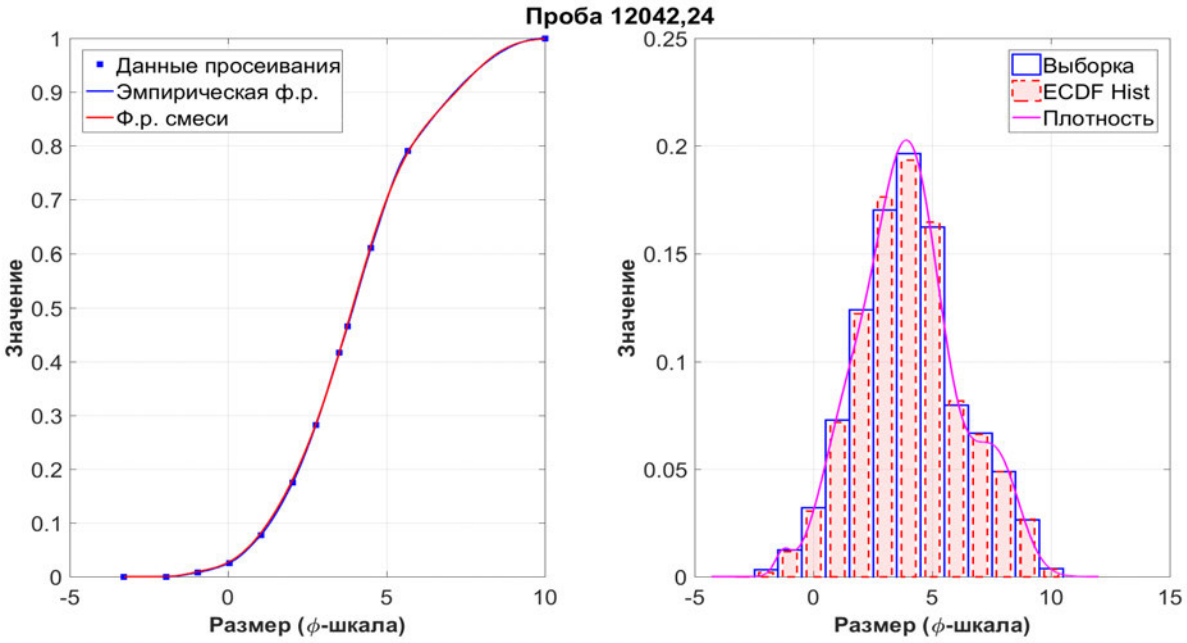

Рис. 4.1. Проба лунного грунта 12042,24 (миссия «Апполон-12»)

На графиках слева представлены исходные данные из таблиц каталога `NASA` (они изображаются квадратами), полученные в процессе просеивания, их интерполяция с помощью полиномов Эрмита (синяя сплошная линия) и аппроксимирующая смесь (сиреневая линия). Видно, что обе кривые практически всюду совпадают. Такая ситуация повторяется для абсолютного большинства анализируемых проб. Отметим, что для числа

компонент k в формуле (4.4) проверялись разные значения. Эмпирически было установлено, что необходимый баланс между качеством аппроксимации и вычислительной сложностью достигается при значении $k = 4$, которое и использовано при обработке всех 317 выборок.

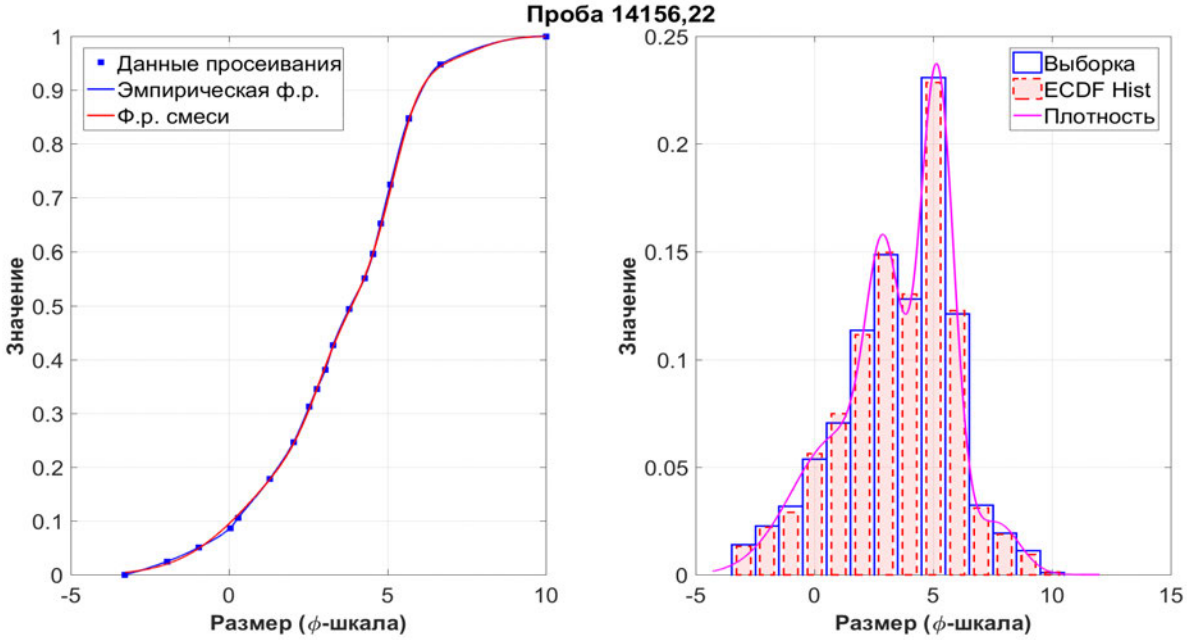

Рис. 4.2. Проба лунного грунта 14156,22 (миссия «Апполон-14»)

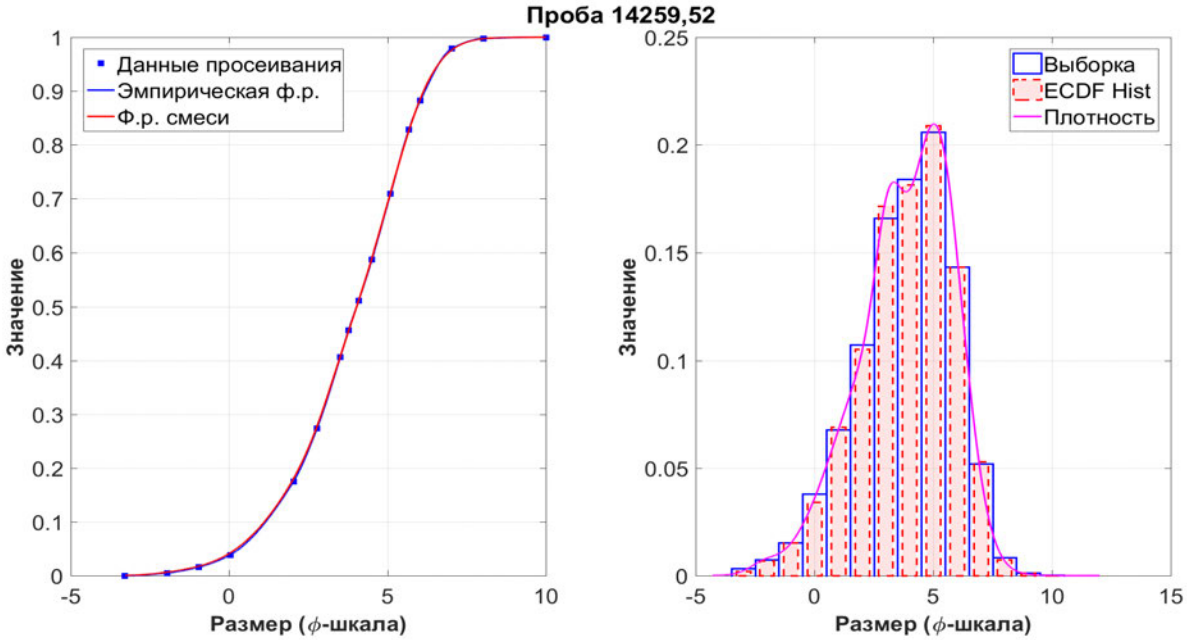

Рис. 4.3. Проба лунного грунта 14259,52 (миссия «Апполон-14»)

На графиках справа на рис. 4.1–4.9 приведены гистограммы для имитационных выборок и просеянных данных (`ECDF Hist`), построенных по

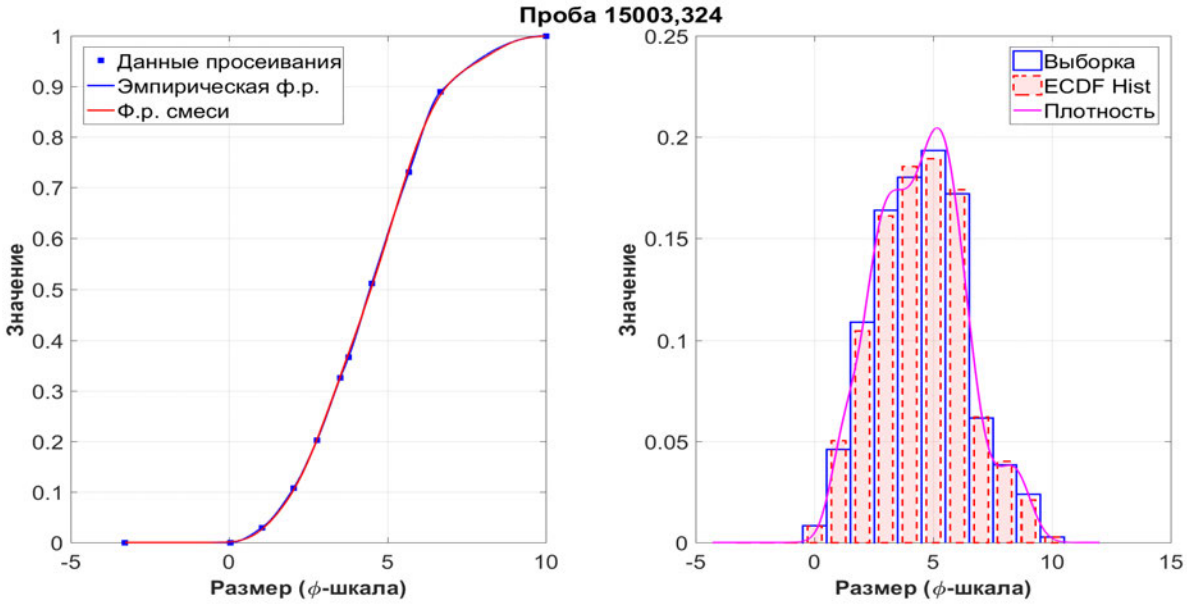

Рис. 4.4. Проба лунного грунта 15003,324 (миссия «Апполон-15»)

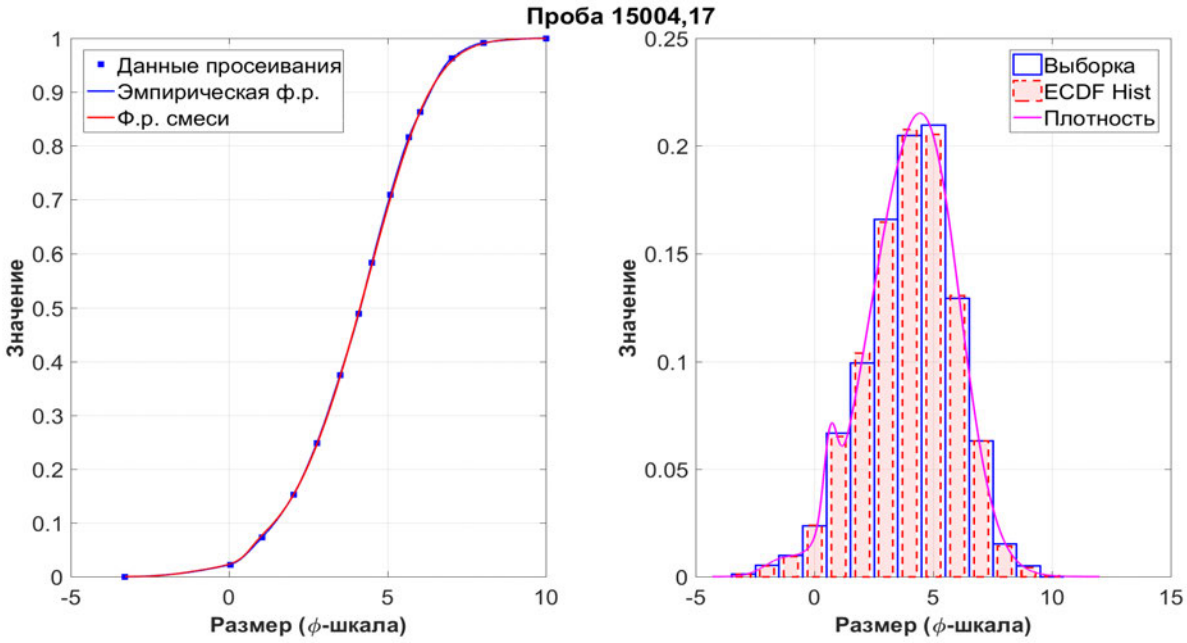

Рис. 4.5. Проба лунного грунта 15004,17 (миссия «Апполон-15»)

эмпирической функции распределения, для конкретной пробы, а также их приближение плотностью смешанного распределения (`Mixture PDF`). Необходимо отметить, что точность аппроксимации определялась именно по сравнению с эмпирической функцией распределения, а данные графики служат лишь для более подробной иллюстрации работы метода. Очевидно, что как функции распределения, так и гистограммы приближаются визуально очень хорошо, даже с учетом различных особенно-

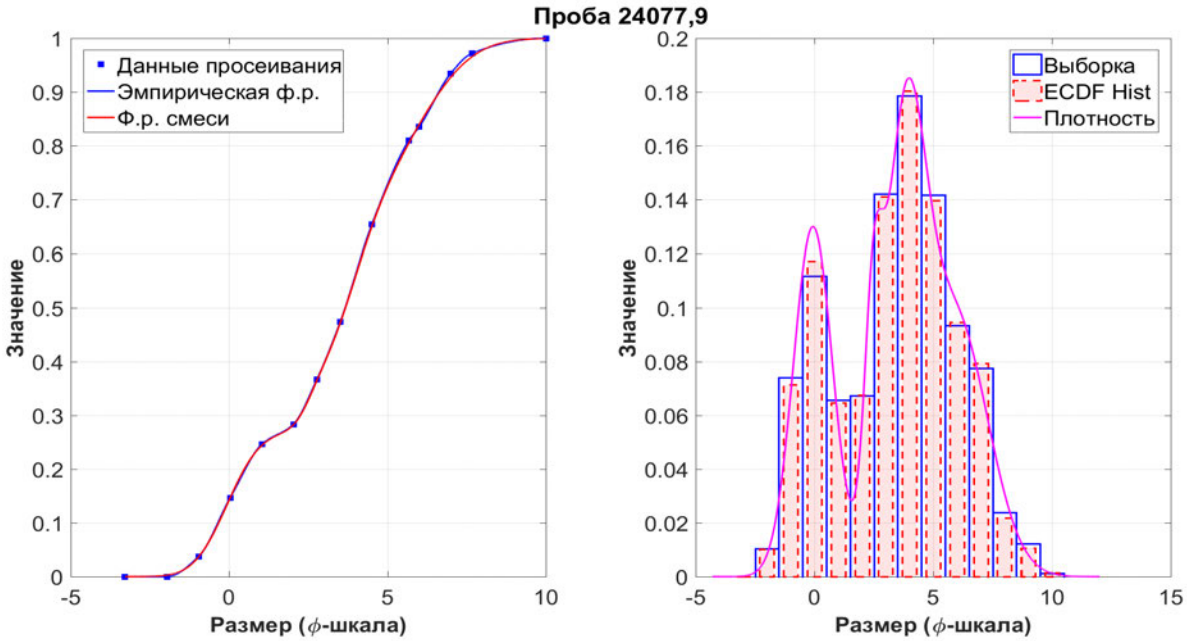

Рис. 4.6. Проба лунного грунта 24077,9 (миссия «Луна-24»)

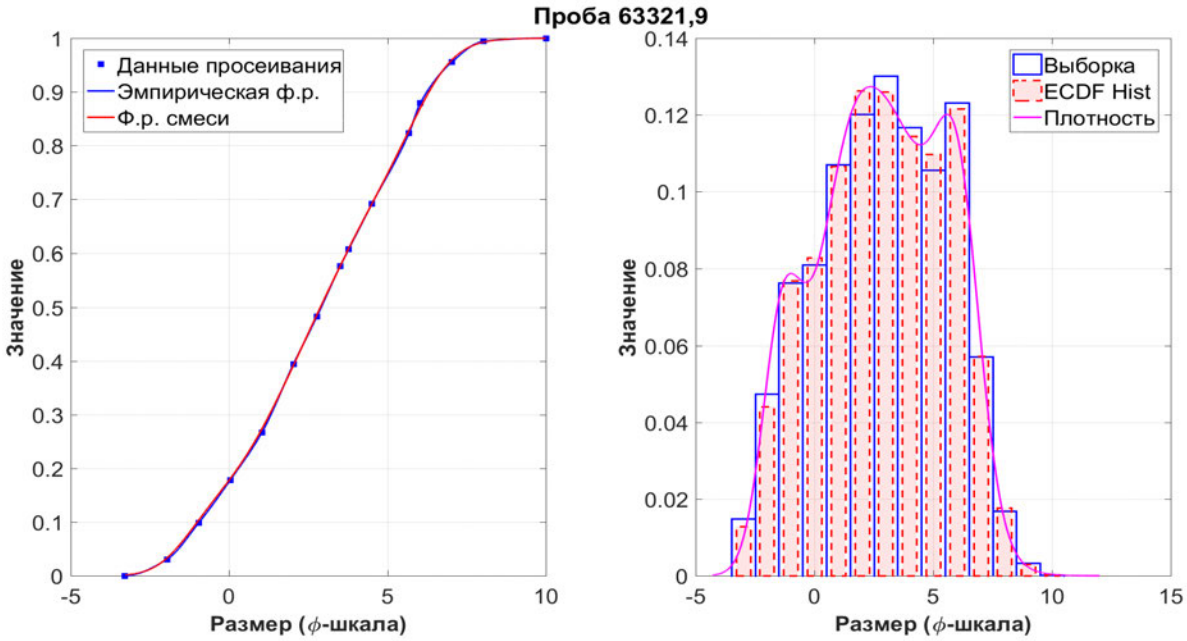

Рис. 4.7. Проба лунного грунта 63321,9 (миссия «Апполон-16»)

стей в них. Во всех случаях форма распределений является существенно негауссовской, поэтому и востребовано использование более сложных смешанных вероятностных моделей.

На рис. 4.10 приведены результаты проверки с помощью критерия однородности Колмогорова и дополнительно симулированных выборок. Для большей наглядности на графиках обозначены стандартные критические уровни 0,01 и 0,05. Первый из них превышен P-значениями

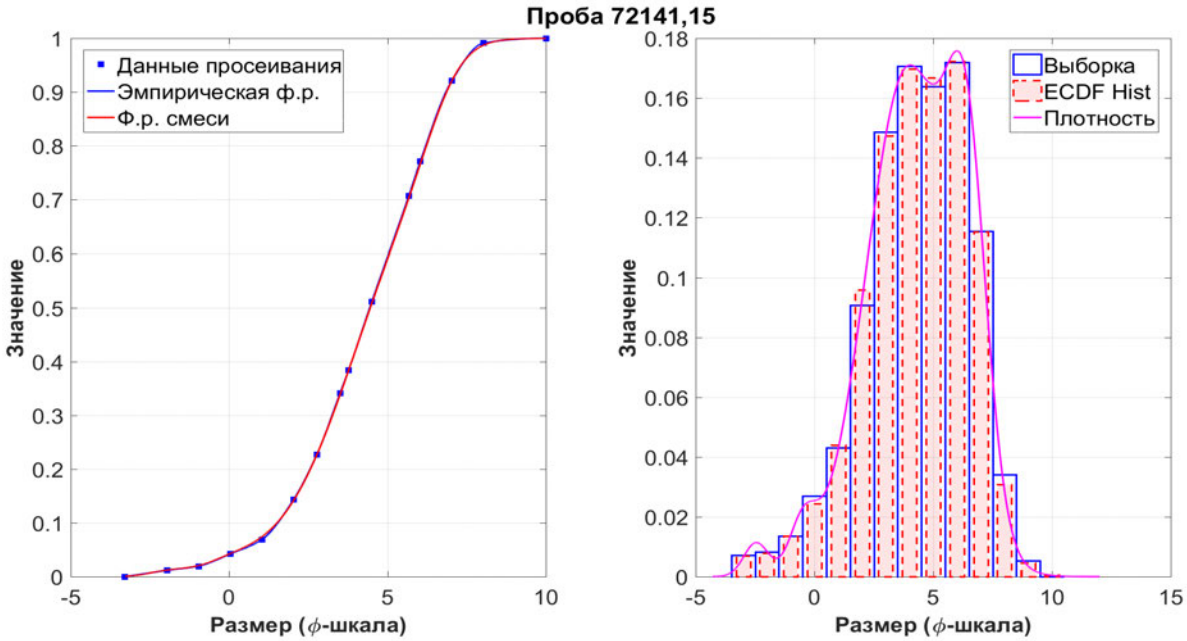

Рис. 4.8. Проба лунного грунта 72141,15 (миссия «Апполон-17»)

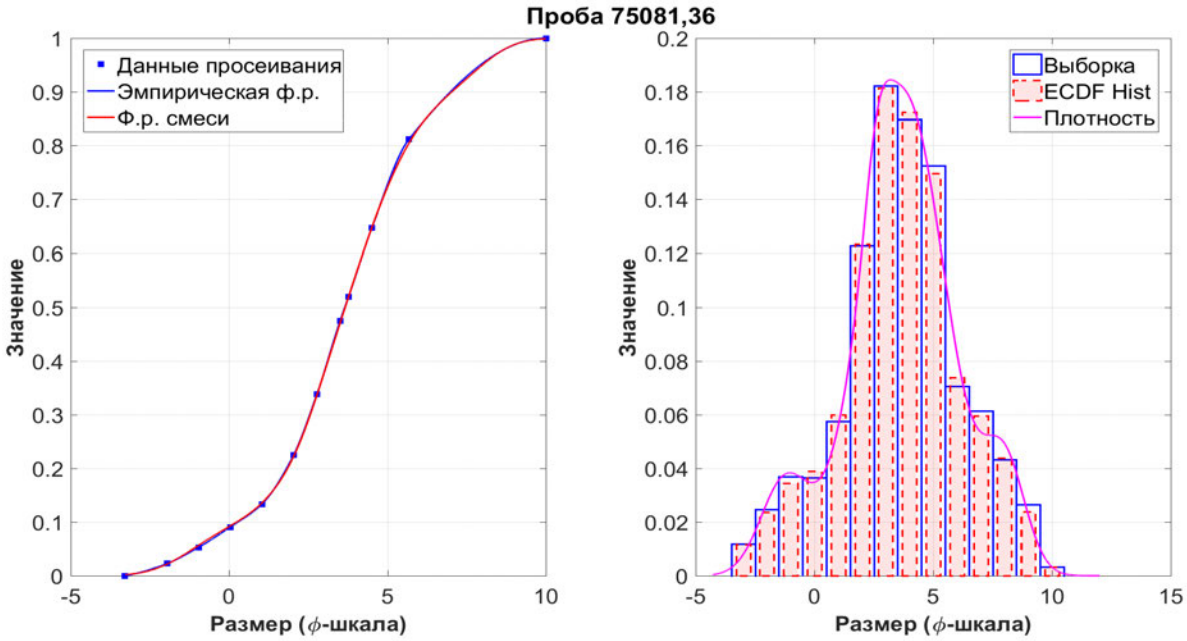

Рис. 4.9. Проба лунного грунта 75081,36 (миссия «Апполон-17»)

для 84,5% выборок (268 из 317), а второй – для 70,7% наборов (224 из 317). Таким образом, для абсолютного большинства проб с помощью бутстреп-метода получены достаточно хорошие результаты аппроксимации. Вместе с тем, их существенное улучшение не достигается и при повторном моделировании выборок для отдельных рядов. Поэтому, несмотря на высокое визуальное соответствие гистограмм для смоделированных выборок и данных из каталога NASA, со статистической точки зре-

ния расхождения между модельным и эмпирическим распределением в ряже случаев является значимым – и требуется развитие методов преодоления данной проблемы. Подход к решению будет предложен в следующем разделе, в котором параметры приближающего распределения будут определяться как наилучшие возможные с точки зрения используемого статистического критерия χ^2.

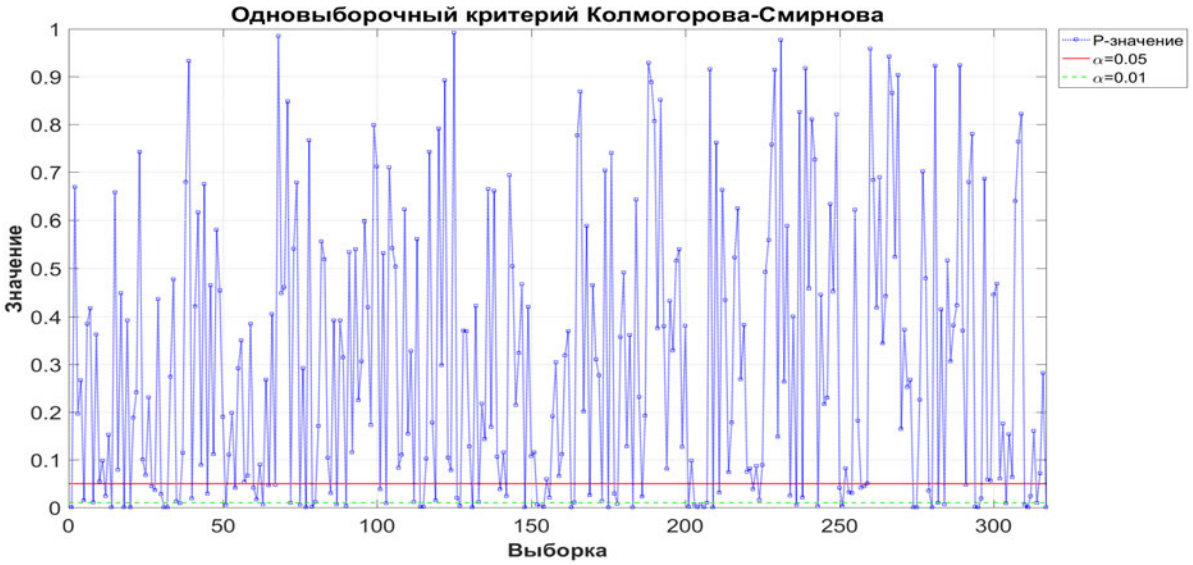

Рис. 4.10. Ошибки аппроксимации (критерий Колмогорова)

4.3 Аппроксимации с помощью метода минимизации статистики χ^2

В предыдущем разделе был продемонстрирован достаточно высокий уровень согласия получаемых с помощью бутстреп-процедуры вероятностных распределений и данных просеивания образцов лунного реголита. Однако данный метод не лишен ряда недостатков. Во-первых, обязательно требуется интерполировать точки – исходные данные из каталога NASA, что может существенным образом повлиять на конечные результаты. Во-вторых, на генерацию выборок и их обработку EM-алгоритмом затрачивается достаточно заметное время, причем оно растет с объемом выборки. В этом разделе будет предложен альтернативный подход, в рамках которого сгруппированные исходные данные также приближаются функциями, имеющими вид конечной смеси логнормальных законов, но без статистической симуляции выборок. Это может повысить точность аппроксимации, а также уменьшить затрачиваемое на расчеты время.

Итак, в данном методе не предполагается интерполяция отдельных точек, представленных в каталоге, а предлагается отыскание парамет-

ров аппроксимирующей функции (4.4) путем минимизации расстояния между приближающей кривой и известными значениями эмпирической функции распределения. В предыдущем разделе было отмечено, что наилучшие результаты получены для четырехкомпонентных смесей, однако в данном методе может использоваться и иное значение параметра k.

Для аппроксимации будет использована функция $\mathcal{F}(x, \boldsymbol{\theta})$ типа (4.4), где $\boldsymbol{\theta} = \{\theta_i\}$, $\theta_i = (a_i, \sigma_i, p_i)$ и для каждого из параметров справедливы ограничения (1.8). Тогда статистика $\chi^2(\boldsymbol{\theta})$ может быть записана в следующем виде:

$$\chi^2(\boldsymbol{\theta}) = \sum_{i=1}^{N} \frac{((y_{i+1} - y_i) - (\mathcal{F}(x_{i+1}, \boldsymbol{\theta}) - \mathcal{F}(x_i, \boldsymbol{\theta})))^2}{\mathcal{F}(x_{i+1}, \boldsymbol{\theta}) - \mathcal{F}(x_i, \boldsymbol{\theta})}, \tag{4.5}$$

где $\mathbf{x}$ и $\mathbf{y}$ – известные наборы значений для соответствующих размеров частиц и их долей в общем числе из каталога `NASA`, а n – число ненулевых разностей $(y_{i+1} - y_i)$. Отметим, что в исходных данных есть повторяющиеся значения y_i, это приходилось учитывать при обработке. В формуле (4.5) предполагается, что все y_i различны.

Для отыскания числовых оценок неизвестных параметров $\boldsymbol{\theta}$ с учетом ограничений (1.8) для аппроксимирующей модели, использовалось решение задачи нелинейного программирования алгоритмом на основе метода внутренней точки [168, 169, 419]. Качество аппроксимации оценивалось путем подстановки значения статистики $\chi^2(\boldsymbol{\theta}_{opt})$ с оптимальным набором параметров $\boldsymbol{\theta}_{opt}$ в функцию распределения χ^2 с $N-1$ степенью свободы и получением соответствующего P-значения. Описание данной процедуры приведено в алгоритме 4.2.

Алгоритм 4.2. Оценивание параметров аппроксимирующей смеси методом минимизации статистики χ^2

```
1: function ChiApprox(x, y)
2:     pEmp←Diff(y);                    // Нулевые значения исключаются
3:     N←Length(pEmp);
4:     χ²(θ) ←Stat(pEmp,x, y, N);
5:     // Задание ограничений (1.8) для минимизации статистики χ²
6:     options←Constraints( );
7:     // Поиск минимума функции (4.5), условная оптимизация
8:     [ChiParams, ChiPval]←fMinCon(x,χ²(θ), N, options);
9:     return [ChiParams, ChiPval];
```

Примеры применения данной процедуры к реальным пробам лунного реголита представлены на рис. 4.11–4.19. На графиках представлены исходные данные из таблиц каталога NASA, их интерполяция с помощью полиномов Эрмита (синяя сплошная линия) и бутстреп-процедуры (красная линия), также аппроксимирующая смесь, полученная с помощью метода минимизации статистики χ^2 (сиреневая линия). Здесь совпадение кривых, полученных двумя методами не является обязательным – внимание уделяется приближению набора исходных точек. На графиках приведены кривые для одинакового числа компонент $k = 4$, при этом для каждого из методов может быть использовано свое значение.

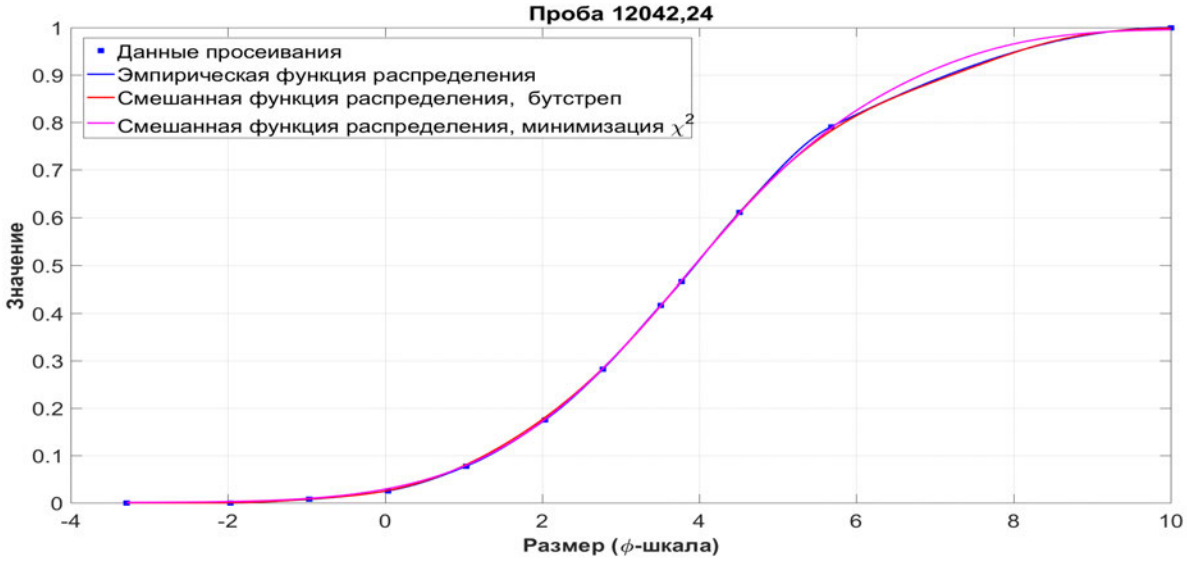

Рис. 4.11. Проба лунного грунта 12042,24 (миссия «Апполон-12»)

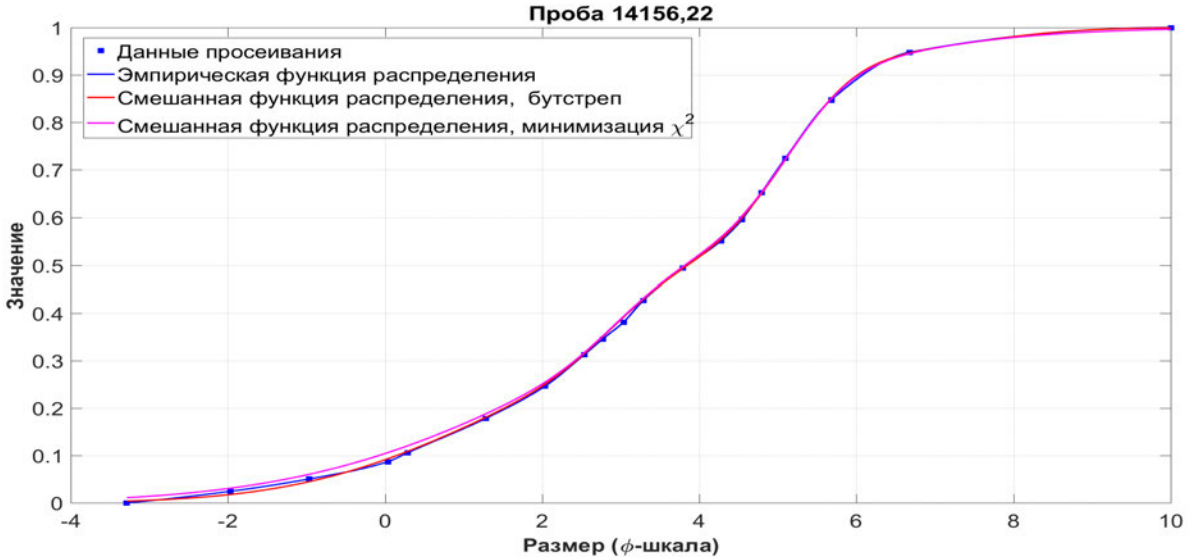

Рис. 4.12. Проба лунного грунта 14156,22 (миссия «Апполон-14»)

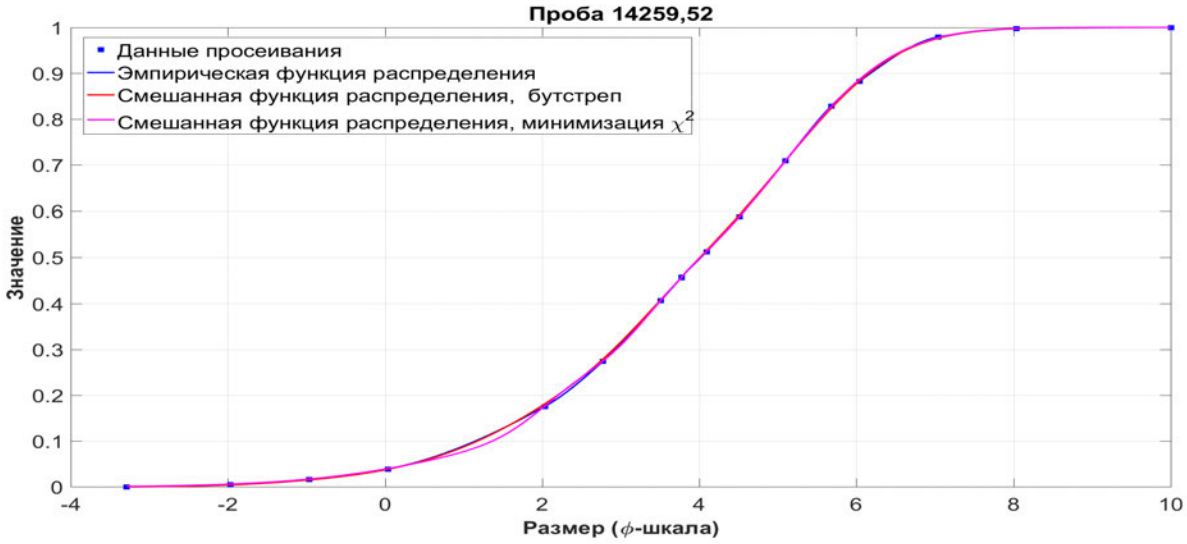

Рис. 4.13. Проба лунного грунта 14259,52 (миссия «Апполон-14»)

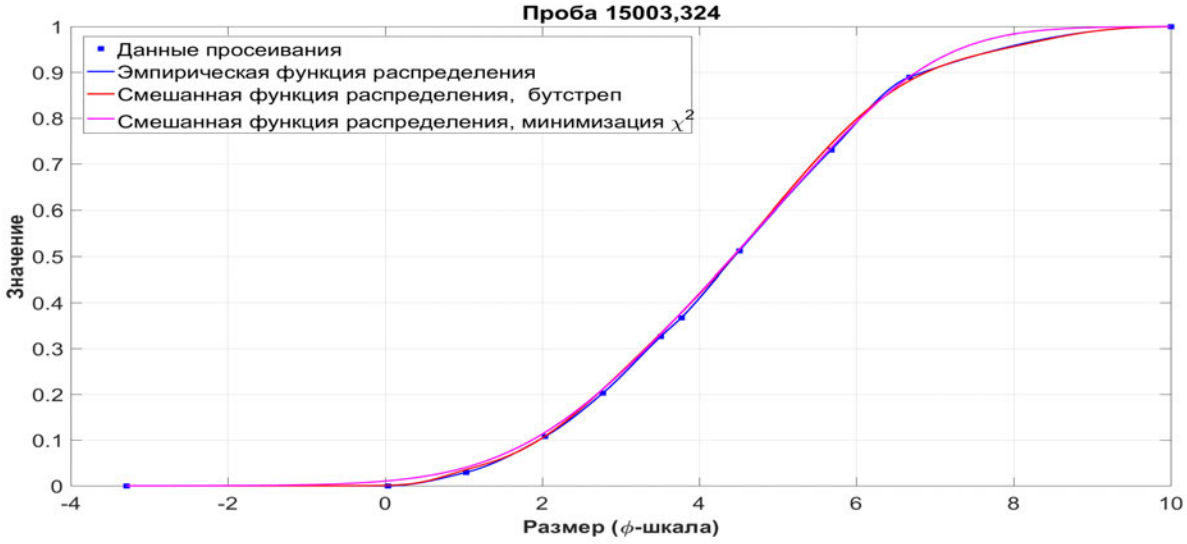

Рис. 4.14. Проба лунного грунта 15003,324 (миссия «Апполон-15»)

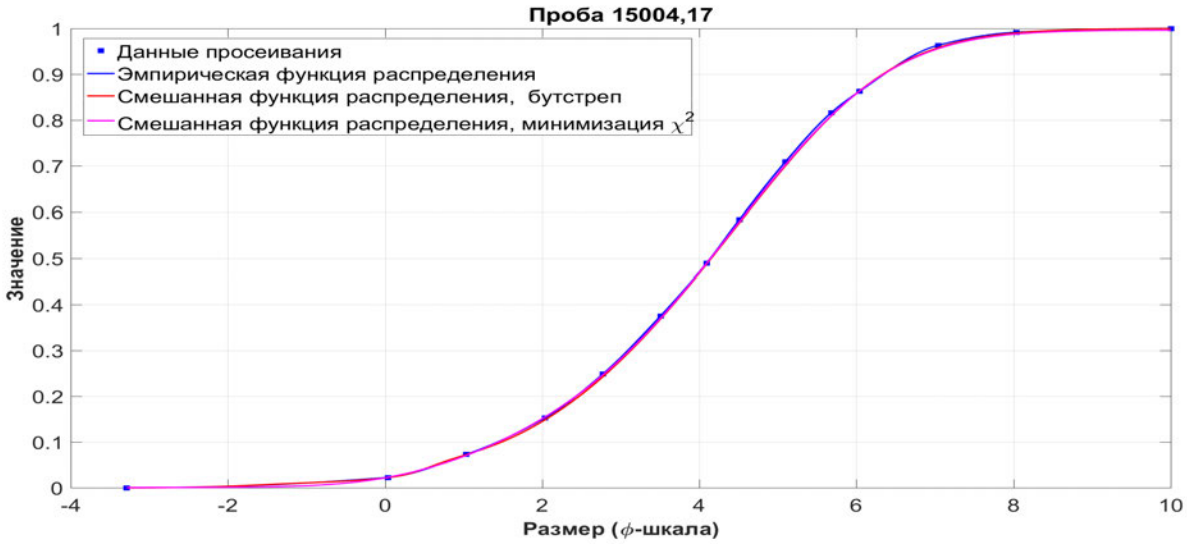

Рис. 4.15. Проба лунного грунта 15004,17 (миссия «Апполон-15»)

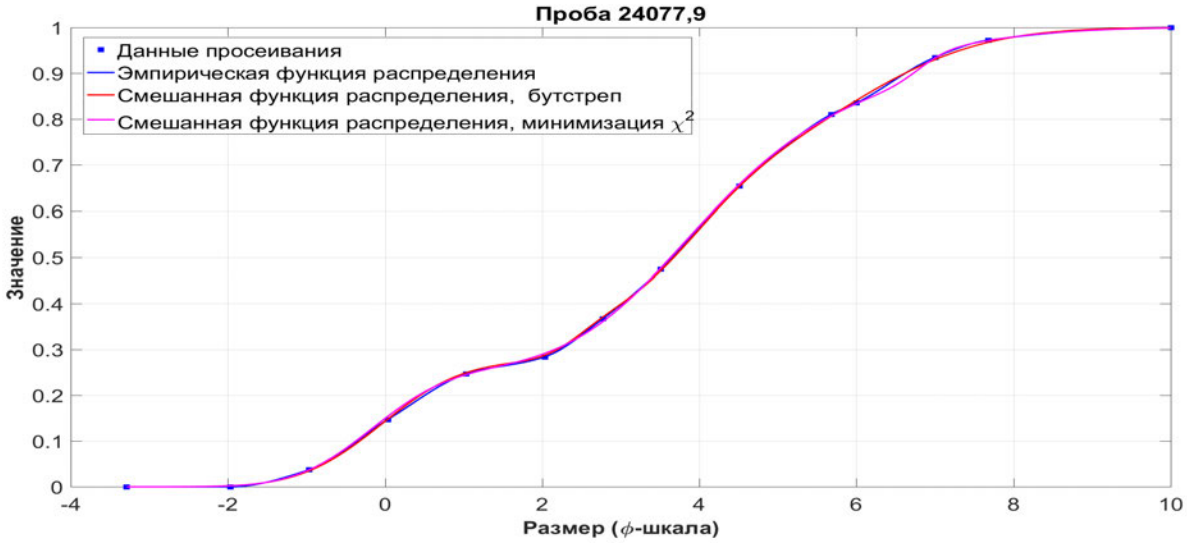

Рис. 4.16. Проба лунного грунта 24077,9 (миссия «Луна-24»)

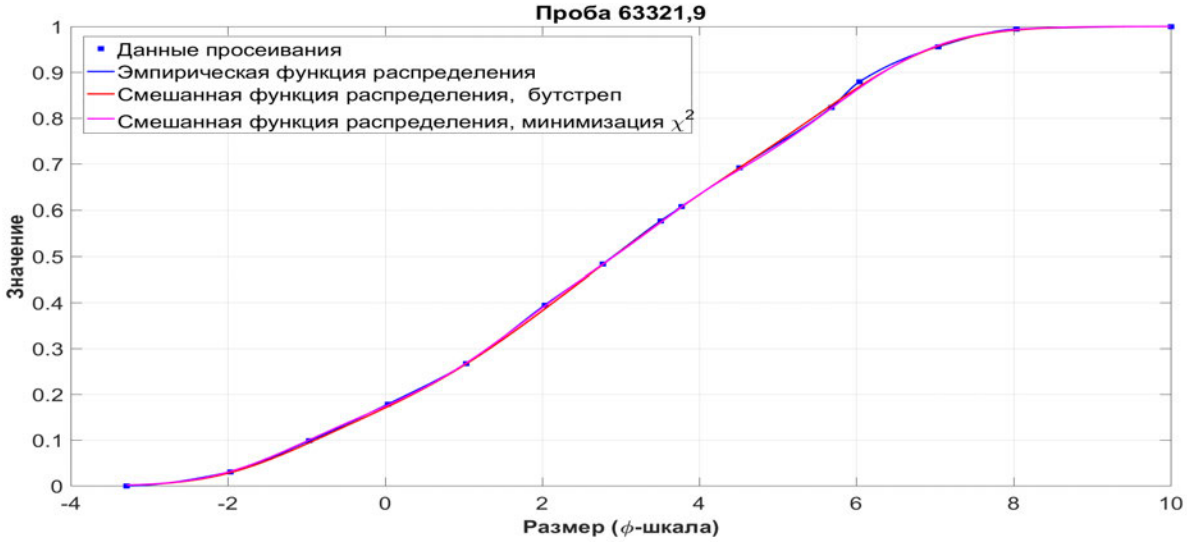

Рис. 4.17. Проба лунного грунта 63321,9 (миссия «Апполон-16»)

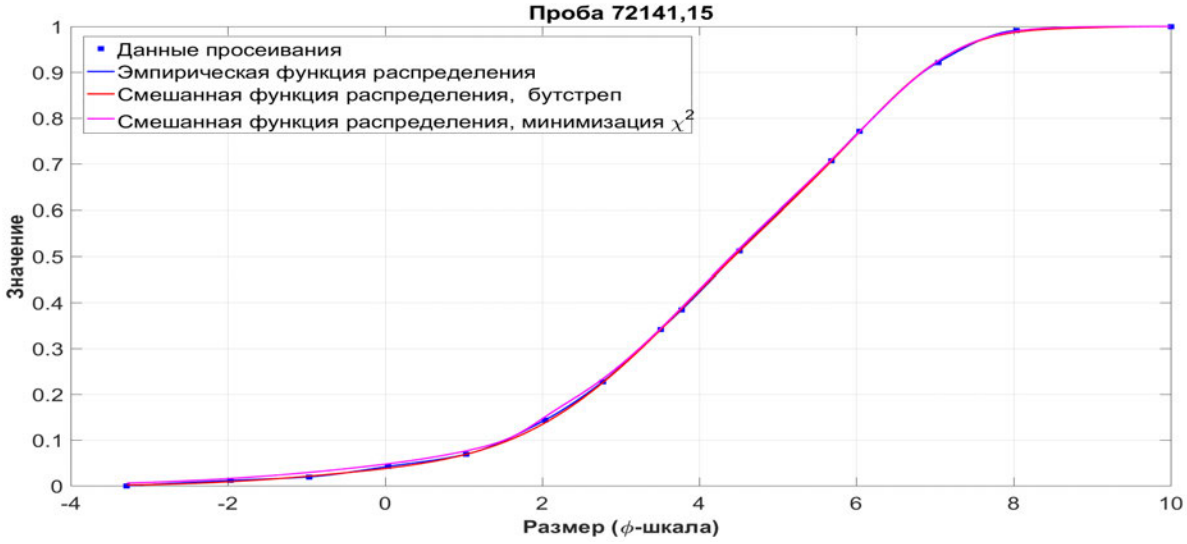

Рис. 4.18. Проба лунного грунта 72141,15 (миссия «Апполон-17»)

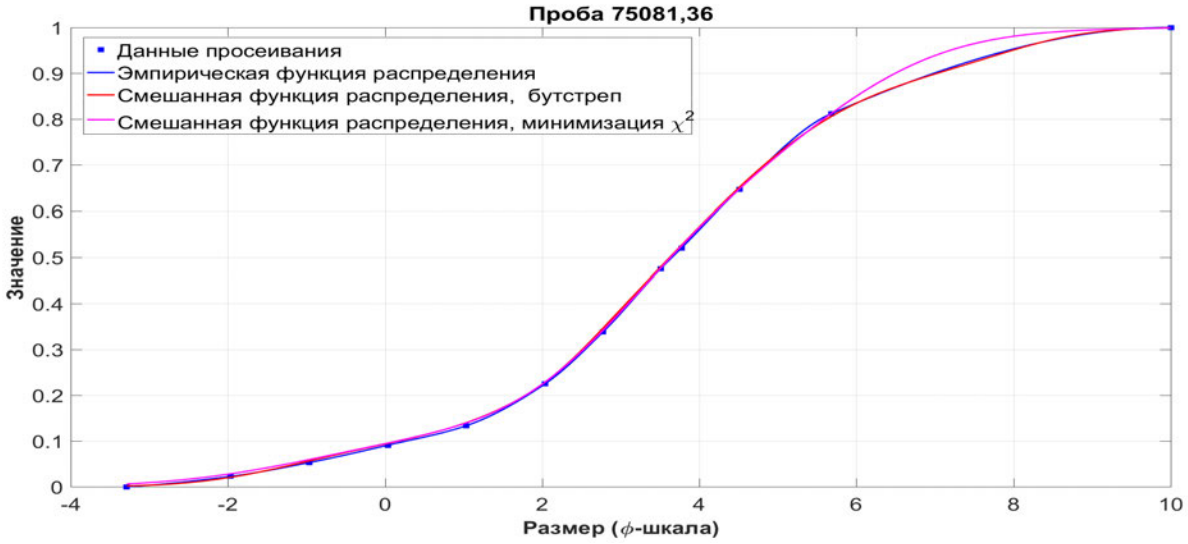

Рис. 4.19. Проба лунного грунта 75081,36 (миссия «Апполон-17»)

На рис. 4.20 приведены результаты проверки с помощью критерия однородности χ^2 для $k = 4$ (см. формулу (4.4)).

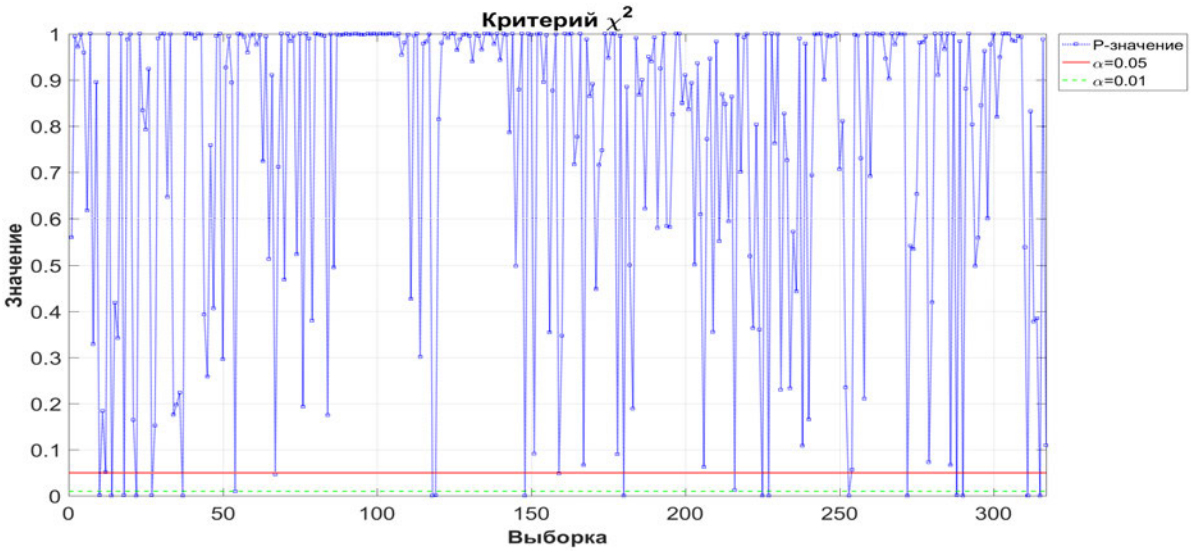

Рис. 4.20. Ошибки аппроксимации (критерий χ^2)

На графиках нанесены стандартные критические уровни 0,05 и 0,01. Критический уровень 0,05 для критерия χ^2 превышен P-значениями для 92,7% выборок (294 из 317) при аппроксимации трехкомпонентными смесями и для 93,1% выборок (295 из 317) при аппроксимации четырехкомпонентными смесями. Критический уровень 0,01 превышен P-значениями для 94,6% выборок (300 из 317) при аппроксимации трехкомпонентными смесями и для 94% выборок (298 из 317) в случае четырехкомпонентных распределений. Фактически, качество статистической аппроксимации одинаково как для $k = 3$, так и для $k = 4$. Это позволяет

использовать меньшее число параметров при решении задачи условной оптимизации.

Данный метод позволяет преодолеть указанные недостатки бутстреп-процедуры, при этом качество аппроксимации можно оценить как отличное. В качестве недостатка этого подхода можно указать поиск решения с помощью условной оптимизации, которая в зависимости от начального приближения может приводить к некорректным результатам. Поэтому требуется аккуратная настройка параметров, а также использование серии запусков для выбора наиболее удачных значений. Вычислительная сложность предложенной процедуры является умеренной, поэтому повторные расчеты в случае необходимости возможны.

4.4 Кластерный анализ параметров смесей

В данном разделе приводятся результаты кластерного анализа параметров сразу всех аппроксимирующих конечных нормальных смесей. При этом для разбиения по кластерам используется полный набор параметров (математические ожидания, среднеквадратические отклонения и веса), а также тривиальное обратное преобразование для перехода от ϕ-шкалы для размеров к метрической. Соответствующая процедура приведена в алгоритме 4.3.

Алгоритм 4.3. Кластеризация параметров вероятностной аппроксимации распределений размеров частиц лунного реголита

```
1:  function RegolithClustering(Params, NumClust)
2:      Params_μm ←Params2MKM(Params);        // Параметры в мкм
3:      // Кластеризация методом k-medoids
4:      [Clust_φ, Med_φ]←kMedoids(Params, NumClust);        // φ-шкала
5:      [Clust_μm, Med_μm]←kMedoids(Params_μm, NumClust);
6:      // Нечеткая кластеризация c-means
7:      [FClust_φ, Centers_φ]←FCM(Params, NumClust);
8:      [FClus_μm, Centers_μm]←FCM(Params_μm, NumClust);
9:      Clusters←[Clust_φ, Med_φ, Clust_μm, Med_μm, FClust_φ, Centers_φ,
        FClust_μm, Centers_μm];
10:     return Clusters;
```

На рис. 4.21–4.24 (верхние графики) проиллюстрирована взаимозависимость математического ожидания и среднего квадратического отклонения для ϕ-шкалы и стандартных единиц измерения для обоих рас-

смотренных выше методов. Размер и интенсивность цвета точек соответствуют их весам (см. формулу (4.4) и цветовую шкалу справа).

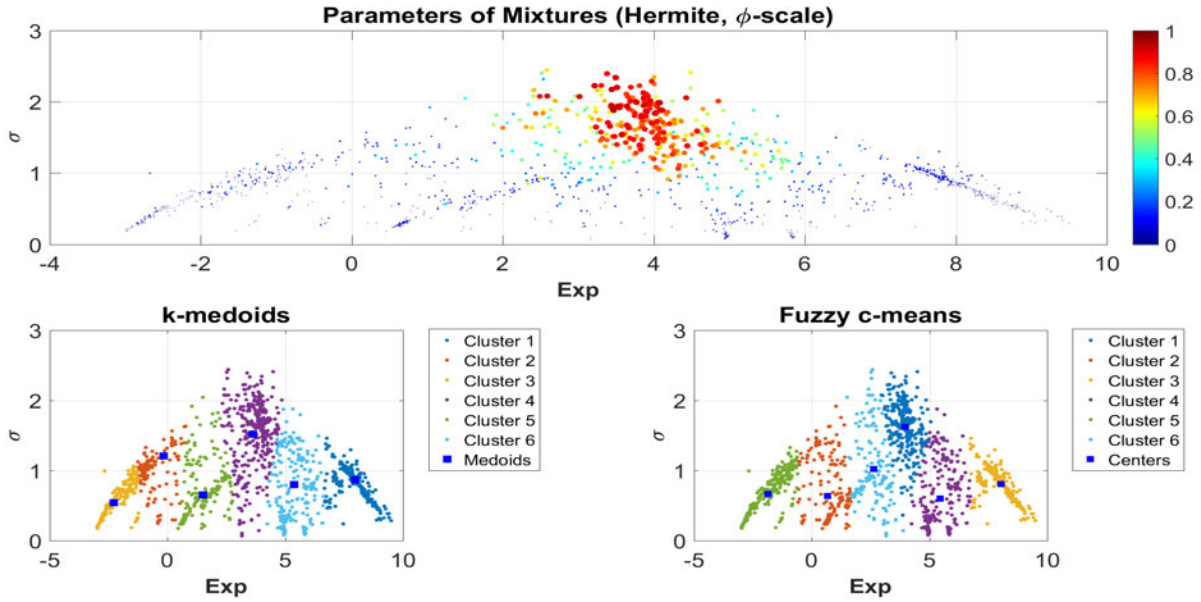

Рис. 4.21. Кластеризация параметров аппроксимирующих смесей (бутстреп, ϕ-шкала)

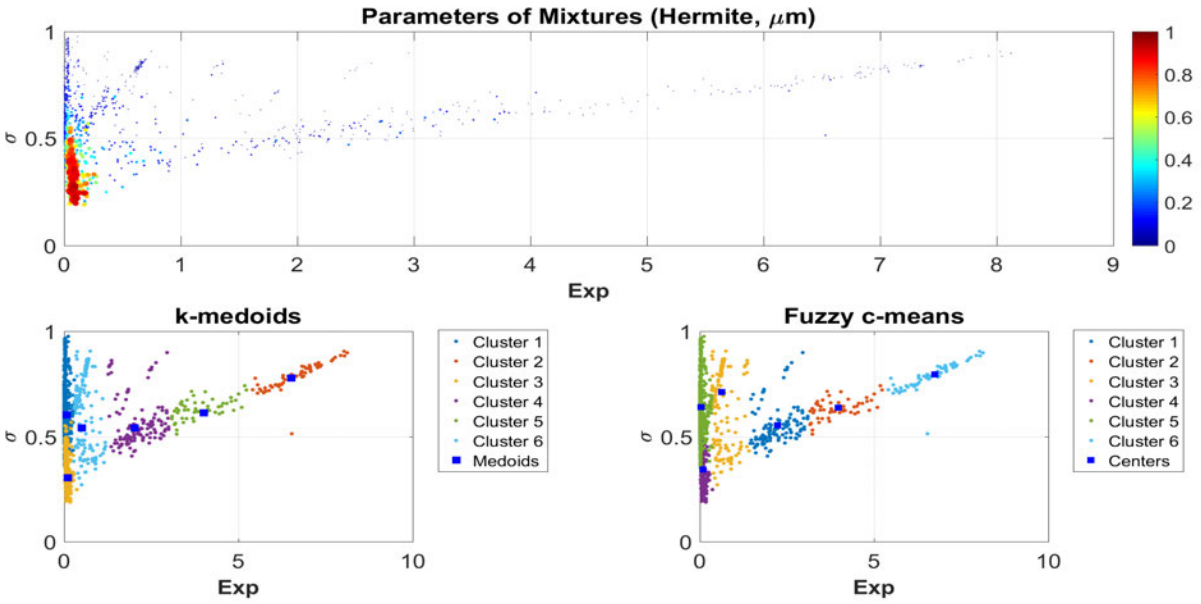

Рис. 4.22. Кластеризация параметров аппроксимирующих смесей (бутстреп, мкм)

Таким образом, представлено существенное уточнение базовой линейной аппроксимации для данной зависимости, предложенной в статье [112], которая, как видно из рис. 4.21–4.24, оказывается чрезмерным загрублением для анализируемых наблюдений.

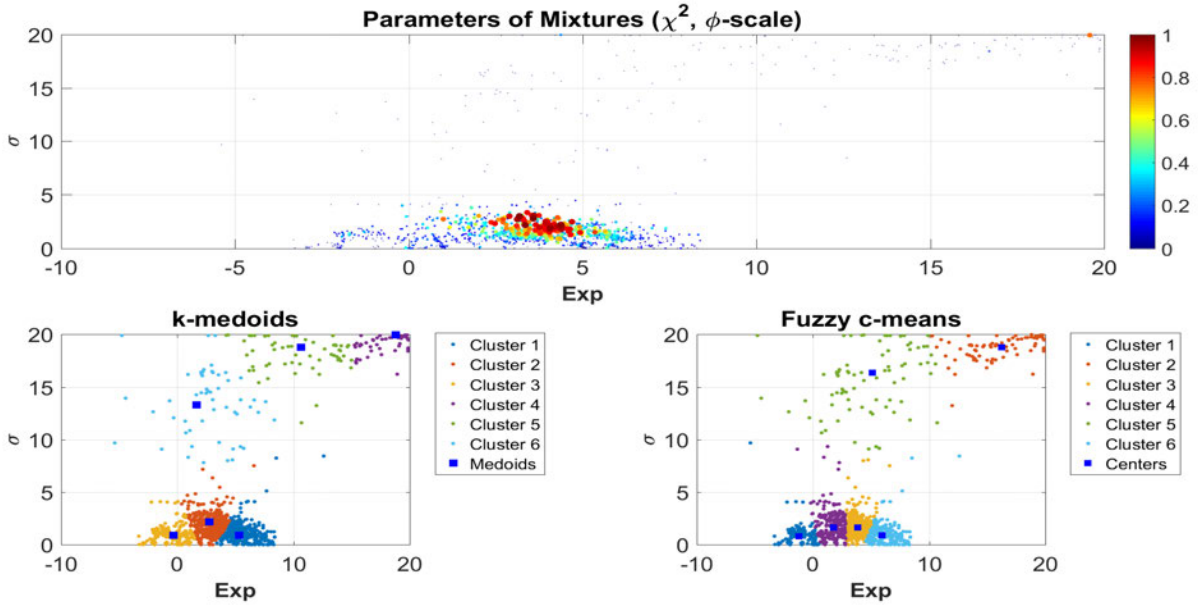

Рис. 4.23. Кластеризация параметров аппроксимирующих смесей (метод на основе статистики χ^2, ϕ-шкала)

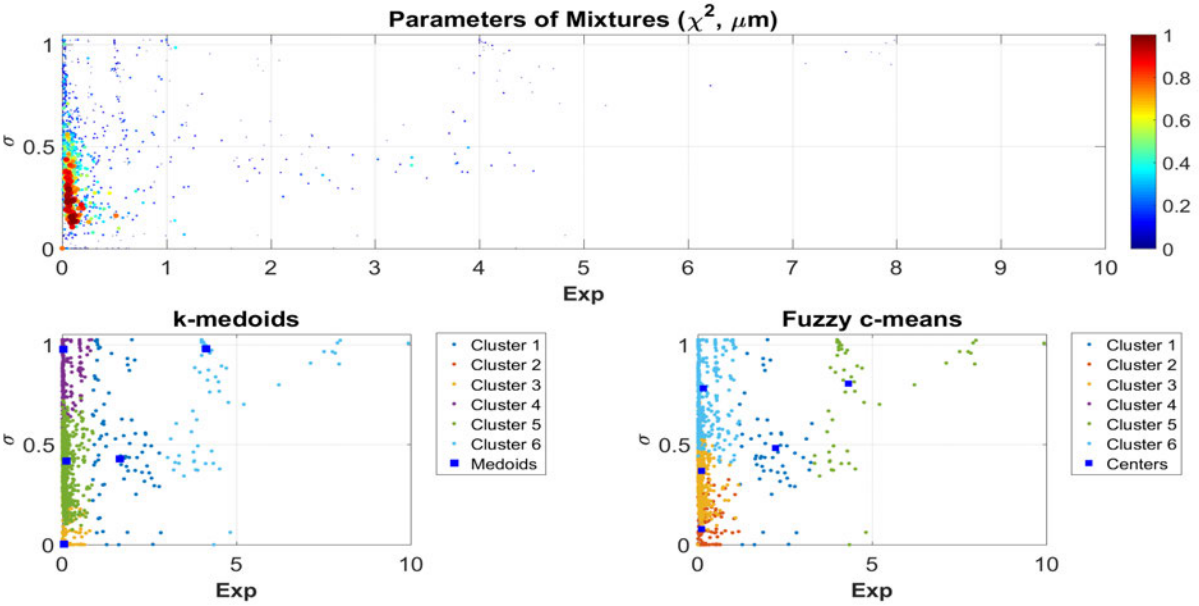

Рис. 4.24. Кластеризация параметров аппроксимирующих смесей (метод на основе статистики χ^2, мкм)

Два нижних графика на каждом из рис. 4.21–4.24 демонстрируют разбиение параметрического пространства на 6 групп методами k-медоид и нечеткой кластеризации c-средних (в этом случае в качестве окончательного решения выбирается кластер, для которого достигается максимальное значение величины вероятности членства для данного элемента среди всех возможных) для бутстреп-процедуры и метода на

основе статистики χ^2. Очевидно, что решения обоих методов в каждом из случаев оказываются достаточно близкими.

Полученные кластеры (при этом для отнесения параметров к тому или иному кластеру используется сразу вся тройка (a_i, σ_i, p_i)) могут быть использованы, например, для соотнесения с химическим составом проб или иными характеристиками реголита. Число кластеров выбрано для сопоставления с химическим составом смесей порошков, повторяющих состав лунного реголита, представленных в статье [111] (см. таблицу 4.1, в которой компоненты отсортированы по массовой доле).

Таблица 4.1. Химические составы смесей порошков реголита [111]

Смесь No. 1		**Смесь No. 2**	
Компонент	**Массовая доля**	**Компонент**	**Массовая доля**
$Si\,O_2$	49.45%	$Si\,O_2$	45.91%
$Ca\,O$	16.92%	$Al_2\,O_3$	23.68%
Al	13.5%	$Ca\,O$	15.71%
$Mg\,O$	10.8%	$Fe\,O$	8.07%
$Fe\,O$	8.7%	Mg	6.05%
$Ti\,O_2$	0.63%	$Ti\,O_2$	0.58%

Для кластеров, полученных в каждом из методов как для бутстреп-алгоритма, так и с помощью минимизации статистики χ^2, были определены веса (отношение числа попавших в данных кластер параметров к общему числу точек). Результаты представлены в таблице 4.2.

Таблица 4.2. Соотношение размеров кластеров

Бутстреп-алгоритм				**Метод на основе статистики χ^2**			
ϕ-шкала		**мкм**		**ϕ-шкала**		**мкм**	
k-meds	**c-means**	**k-meds**	**c-means**	**k-meds**	**c-means**	**k-meds**	**c-means**
25,08%	20,19%	42,43%	42,35%	35,09%	31,94%	51,50%	41,48%
20,35%	19,95%	24,61%	25,71%	33,44%	24,45%	18,53%	28,08%
17,27%	17,90%	13,80%	13,56%	18,61%	19,09%	17,67%	20,90%
15,22%	16,25%	9,70%	9,07%	4,73%	12,22%	6,62%	5,05%
12,54%	14,51%	5,28%	5,28%	4,34%	6,23%	5,52%	4,26%
9,54%	11,20%	4,18%	4,02%	3,79%	6,07%	0,16%	0,24%

Возможно и иное представление графиков 4.21–4.24 для метрической и ϕ-шкал, а именно – трехмерное изображение (см. рис. 4.25–4.28). При этом веса точек откладываются на отдельной шкале, но их размер при отрисовке по-прежнему определяется данным параметром.

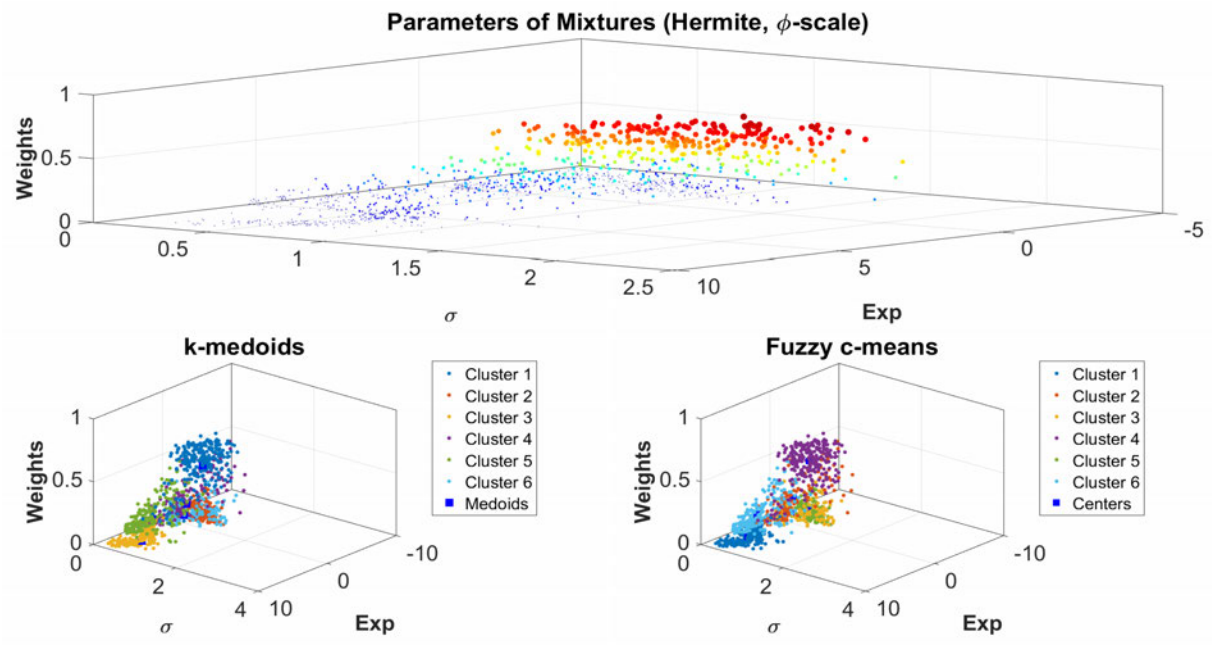

Рис. 4.25. Трехмерное изображение кластеров параметров (бутстреп, ϕ-шкала)

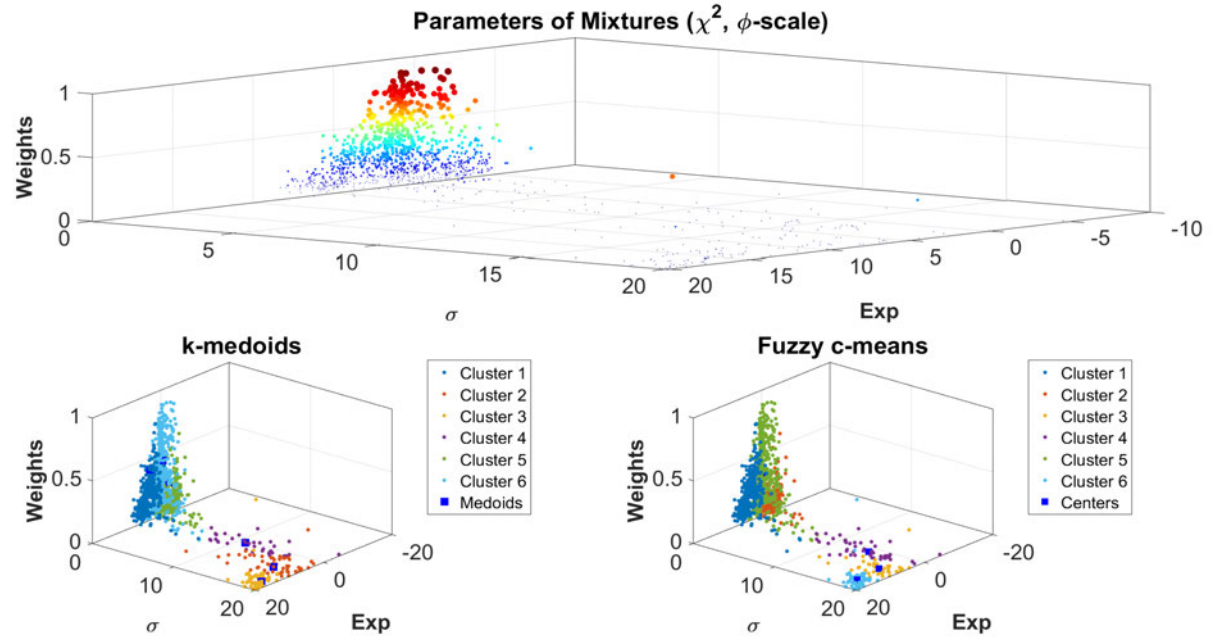

Рис. 4.26. Трехмерное изображение кластеров параметров (метод на основе статистики χ^2, ϕ-шкала)

Безусловно, нет точного совпадения результатов, полученных при анализе реальных лунных реголитов, с искусственно подготовленными в статье [111] порошками – в том числе из-за возможных вычислительных погрешностей, а также использования различных методов кластеризации. Однако определенное сходство явно прослеживается для достаточно большого числа столбцов, особенно стоит отметить кластеризацию

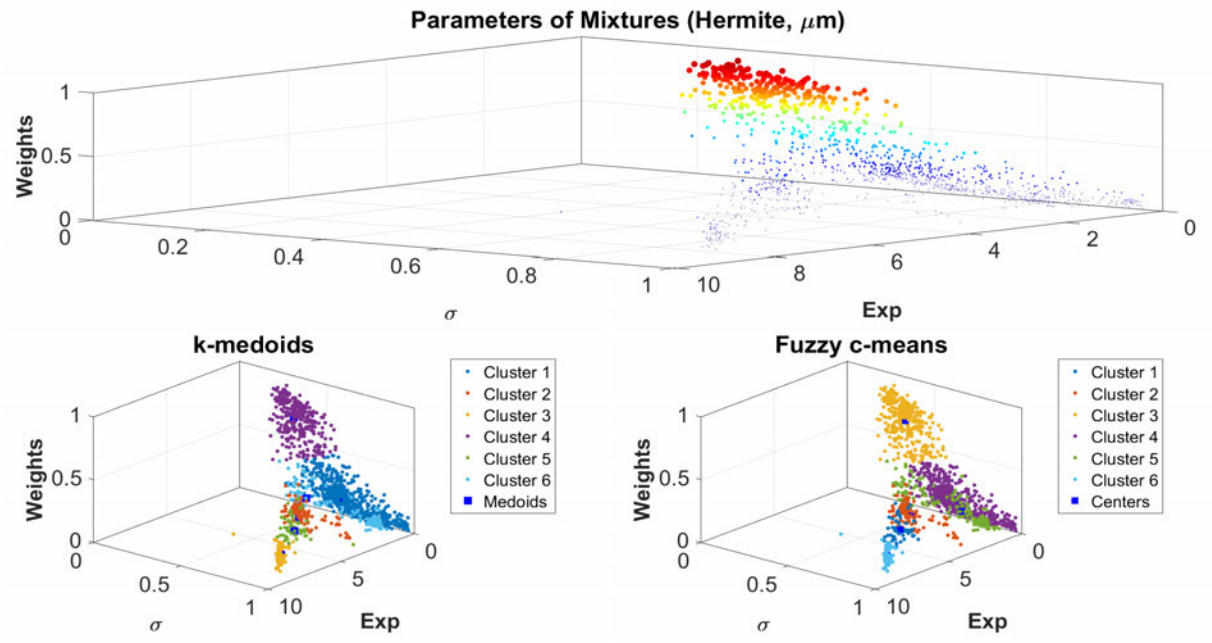

Рис. 4.27. Трехмерное изображение кластеров параметров (бутстреп, мкм)

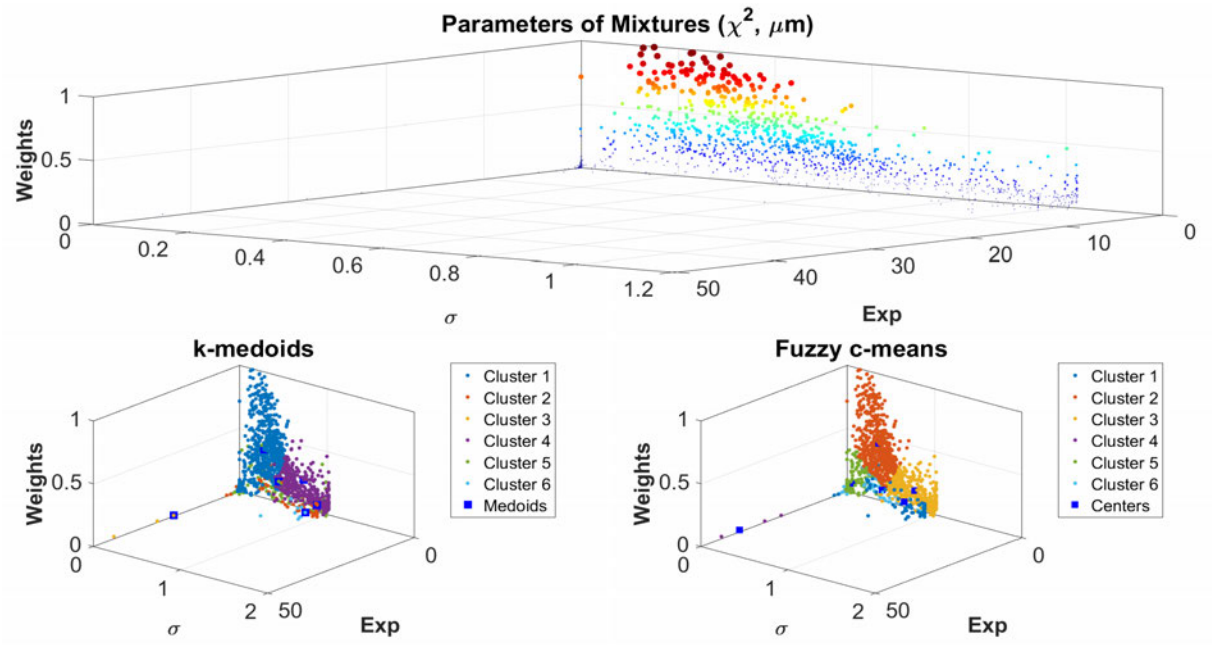

Рис. 4.28. Трехмерное изображение кластеров параметров (метод на основе статистики χ^2, мкм)

параметров в микрометрах при аппроксимации на основе статистики χ^2 (см., например, два самых правых столбца в таблице 4.2). Отметим, что результаты нечеткой кластеризации обладают несколько лучшей воспроизводимостью при разбиении по кластерам по сравнению с медоидами при повторных перезапусках методов кластеризации, однако принципиально выводы остаются неизменными.

4.5 Алгоритм аппроксимации распределений размеров частиц лунного реголита

Таким образом, статистическая обработка всех проб лунного реголита сводится к последовательному применению бутстреп-процедуры (см. алгоритм 4.1) и метода на основе минимизации статистики χ^2 (см. алгоритм 4.2) к каждому образцу из каталога `NASA` с целью формирования набора параметров аппроксимирующих конечных смесей, которые в дальнейшем подвергаются кластеризации (см. алгоритм 4.3). Полностью описанный метод представлен в алгоритме 4.4.

Алгоритм 4.4. Анализ данных лунного реголита

```
1:  function LunarRegolith(RegolithSamples, Size1, Size2)
2:      Init( );          // Загрузка данных и инициализация параметров
3:      Parpool( );  // Запуск инструментов параллельной обработки
4:      for all RegolithSamples do
5:          // PhiSize(i) – размер частиц для i-й выборки в ϕ-шкале
6:          // Values(i) – доля частиц соответствующего размера
7:          ECDF←Fit(PhiSize(i), Values(i));      // Интерполяция ECDF
8:          // Имитационное моделирование выборок (алгоритм 4.1)
9:          [Sample, TestSample]←GenSamples(ECDF, Size1, Size2);
10:         // EM-алгоритм для конечных нормальных смесей
11:         Params(i)←NormalApprox(Sample);
12:         // Ошибки аппроксимации (статистика Колмогорова)
13:         KSError(Params(i), ECDF, TestSample);
14:         // Минимизация статистики хи-квадрат (алгоритм 4.2)
15:         [ChiParam, ChiPval]=ChiApprox(PhiSize(i), Values(i));
16:     // Функция кластеризации параметров (алгоритм 4.3)
17:     Clusters←RegolithClustering(Params);
18:     Plot(RegolithSamples, Params, Clusters);          // Визуализация
19:     return ;
```

Алгоритм 4.4 реализован на языке программирования `MATLAB`. Для расчетов использовались ресурсы гибридного высокопроизводительного вычислительного кластера архитектуры `Intel x86_64`: сервер `Huawei XH 622 V3` (два процессора `Intel Xeon E5-2683V4` с тактовой частотой 2,1 ГГц (16 ядер), 512 Гб оперативной памяти и 2 видеокарты `NVIDIA Tesla K80`. Это позволило повысить скорость вычислений в

среднем с 39,3–53 с для одной пробы, обработанной с помощью стандартного настольного решения, до 13,9 с, полученных на вычислительном кластере. Таким образом, было получено почти четырехкратное ускорение вычислений, которое особенно важно при реализации бутстреп-подхода, для которого важную роль играет размер генерируемых выборок. Метод минимизации статистики χ^2 менее требователен к вычислительным ресурсам.

Полученные результаты могут оказаться весьма перспективными с точки зрения анализа и прогнозирования поведения плазмохимических процессов в пылевых структурах, возникающих при воздействии импульсного излучения гиротрона [110, 111].Указанные эксперименты ориентированы на моделирование процессов воздействия на частицы лунного реголита для разработки эффективных технологических решений, которые могут быть использованы, в частности, при подготовке новых космических миссий.

Глава 5

Вероятностные модели процессов в физике турбулентной плазмы

В данной главе описываются разработка и применение различных методов интеллектуального анализа данных на основе конечных смесей вероятностных распределений и их скользящего разделения в совокупности с нейросетевыми подходами для моделирования и изучения тонкой структуры процессов, наблюдаемых в экспериментах с турбулентной плазмой.

5.1 Методология вероятностного анализа тонкой структуры процессов с помощью спектров

Спектральный анализ является одним из самых мощных инструментов обработки сигналов экспериментов [100], в том числе и при исследовании плазменной турбулентности [145]. Как известно, плазма является состоянием вещества с большим числом степеней свободы, и расшифровка спектров плазменной турбулентности является некорректной задачей. Это усложняет задачу идентификации, так как при известном числе процессов форма гармоник в амплитудном спектре остается неизвестной.

Отметим, что комплексные Фурье-спектры, измеренные рефлектометром на краю плазмы, в интервале стационарных макропараметров плазмы изменяются существенным образом. Не сохраняется форма и по-

луширина спектра. Доплеровский сдвиг, который связан с полоидальным вращением флуктуаций (и плазмы), изменяется в течение эксперимента (разряда) в стационарных условиях более чем в четыре раза. Использование подобных данных для оценки скорости флуктуаций приводило бы к выводам о том, что в стационарных условиях не сохраняется либо фазовая скорость колебаний, либо полоидальная скорость вращения плазмы (или радиальное электрическое поле), либо сразу обе. В течение разряда плазмы в стеллараторе Л-2М изменяется и Фурье-спектр коротковолновой турбулентности, измеренной методом рассеяния излучения гиротрона в центре плазмы. Однако именно спектры дают возможность определить тип неустойчивости, механизм формирования турбулентности, доли ионно-звуковых солитонов и дрейфовых вихрей. Поэтому необходима разработка инструментов для прикладного решения описанной некорректной задачи.

Для изучения комплексных частотных спектров флуктуаций для оценки скорости их движения (суммарной фазовой и полоидальной) по частотному сдвигу, а также выделения стохастических процессов в статье [63] автором был предложен специальный статистический алгоритм на основе конечных нормальных смесей, который в данном разделе будет использован для анализа спектров с внесением ряда важных модификаций, в частности для оценивания односторонних спектров. Для этого будут применяться в том числе и конечные логнормальные и гамма-смеси, носители которых сосредоточены на положительной полуоси. Данный бутстреп-метод будет использован для анализа низкочастотной плазменной турбулентности на краю и в центре плазменного шнура в стеллараторе Л-2М.

5.1.1 Описание алгоритма

Идея метода основана на интерпретации спектра как плотности некоторого неизвестного вероятностного распределения, которое в дальнейшем используется для моделирования выборки из этого распределения, что позволяет оценить параметры с помощью одной из модификаций EM-алгоритмов с выбором соответствующего базового семейства. В дальнейшем будут приведены примеры для конечных нормальных, логнормальных и гамма-распределений. В алгоритме 5.1, в котором продемонстрирована реализация данного метода, соответствующие настройки задаются с помощью векторного параметра `options`.

Алгоритм 5.1. Бутстреп-метод анализа спектров

```
1: function PhysBootstrap(Spectrum, options)
2:     // Интерполяция дискретных точек спектра
3:     F_x ←Fit(Spectrum, options.x);// options.x – область определения
4:     // Случайный вектор для метода обратных функций
5:     r←Rand(options.X_αSize);
6:     for i=1:options.X_αSize do        // Имитационное моделирование
7:         X_α,i ←fsolve(F_x, r_i);          // Метод обратных функций
8:     // EM-алгоритм для смоделированной выборки
9:     Params←EMs(X_α, options.EM);
10:    F_mixt←FApprox(Params, options.distribution);
11:    plot(Spectrum, F_mixt, options.plot);              // Визуализация
12:    return [X_α, F_mixt];
```

Отметим, что в разделе 4.2 подобный подход был успешно использован для симуляции выборок для оценивания параметров неизвестного эмпирического распределения размеров частиц лунного реголита. Ключевое отличие от рассматриваемого в этом разделе алгоритма заключается в объекте аппроксимации: здесь спектр только *интерпретируется* как плотность, но на самом деле таковым не является.

5.1.2 Анализ экспериментальных данных

В данном разделе рассмотрим применение алгоритма 5.1 для обработки экспериментальных данных низкочастотной структурной турбулентности на стеллараторе Л-2М, полученных методами доплеровской рефлектометрии [358] и рассеянием излучения второй гармоники гиротрона на флуктуациях [382]. Для уменьшения шумовых спектральных компонент используется метод Велча [103]. Исходные данные представляют набор выборок одинакового объема – 4096 элементов (такое количество наблюдений обеспечивает хорошее разрешение по частоте). Каждый такой ряд отвечает за определенный момент времени в работе диагностики и представляет собой двухсторонний спектр с нулевой частотой, эквивалентной несущей частоте специализированного сверхвысокочастотного генератора.

Сначала рассмотрим аппроксимацию двусторонних спектров с использованием конечных нормальных смесей. На рисунке 5.1 изображена эволюция во времени характерного для диагностики доплеровской

рефлектометрии сглаженного полного комплексного спектра для одной серии (серия номер 58515) с 55 мс по 58 мс. По оси абсцисс отложены значения частот в МГц, по оси ординат – полученные значения спектра. График соответствует стационарному режиму для макропараметров плазмы, в котором сохраняются неизменными такие величины как плотность, электронная температура и некоторые другие.

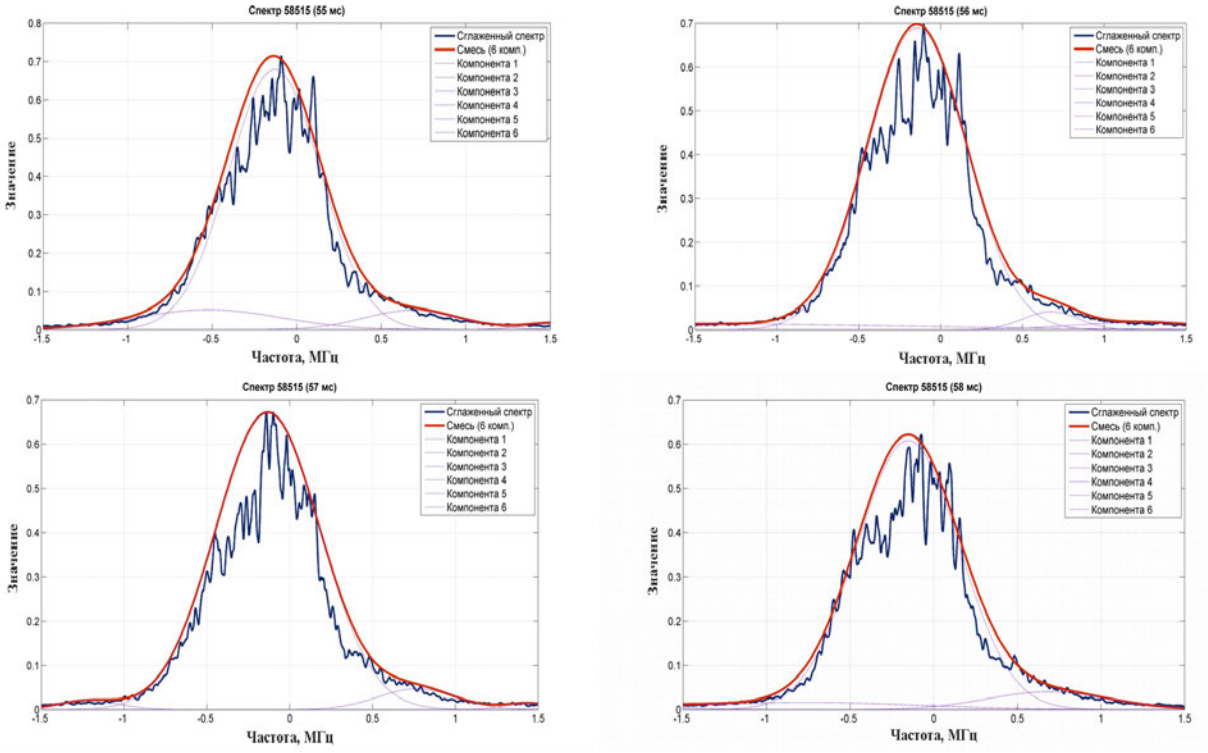

Рис. 5.1. Аппроксимация спектра 58515 с 55 мс по 58 мс

Оценивание параметров производилось с помощью метода EM-типа (см. строку 10 в алгоритме 5.1) с возможными шестью компонентами в смеси. Объем имитационной выборки составлял 30000 элементов. Использовался стандартный критерий останова (2.5) с $\varepsilon = 10^{-8}$.

Из шести компонент в каждом случае значимыми оказались лишь три (фиолетовые пунктирные линии на рисунке 5.1), при этом их форма и положение сохраняются для каждого из рассмотренных временных интервалов. Вклад остальных компонент в итоговую функцию (красная сплошная линия) незначителен (их уровень лежит ниже экспериментальной ошибки измерения). Приближающая спектр кривая сохраняет стационарное положение относительно оси частот. Максимальная компонента в данном разложении спектра может быть проинтерпретирована как доплеровский сдвиг, который в свою очередь прямо пропорционален скорости полоидального вращения плазмы в области работы диагности-

ки доплеровского рефлектометра в тороидальной установке стелларатора Л-2М. Такое поведение максимальной компоненты соответствует физическим данным и позволяет вычислять полоидальную скорость вращения плазмы более точно, чем это дают другие методы обнаружения максимума в спектре. Остальные две значимые компоненты, вероятно, связаны с фазовыми скоростями флуктуаций.

Также обрабатывались данные диагностики рассеяния излучения на второй гармонике гиротрона, ориентированной на измерение коротковолновых флуктуаций плазмы вблизи центра плазменного шнура.

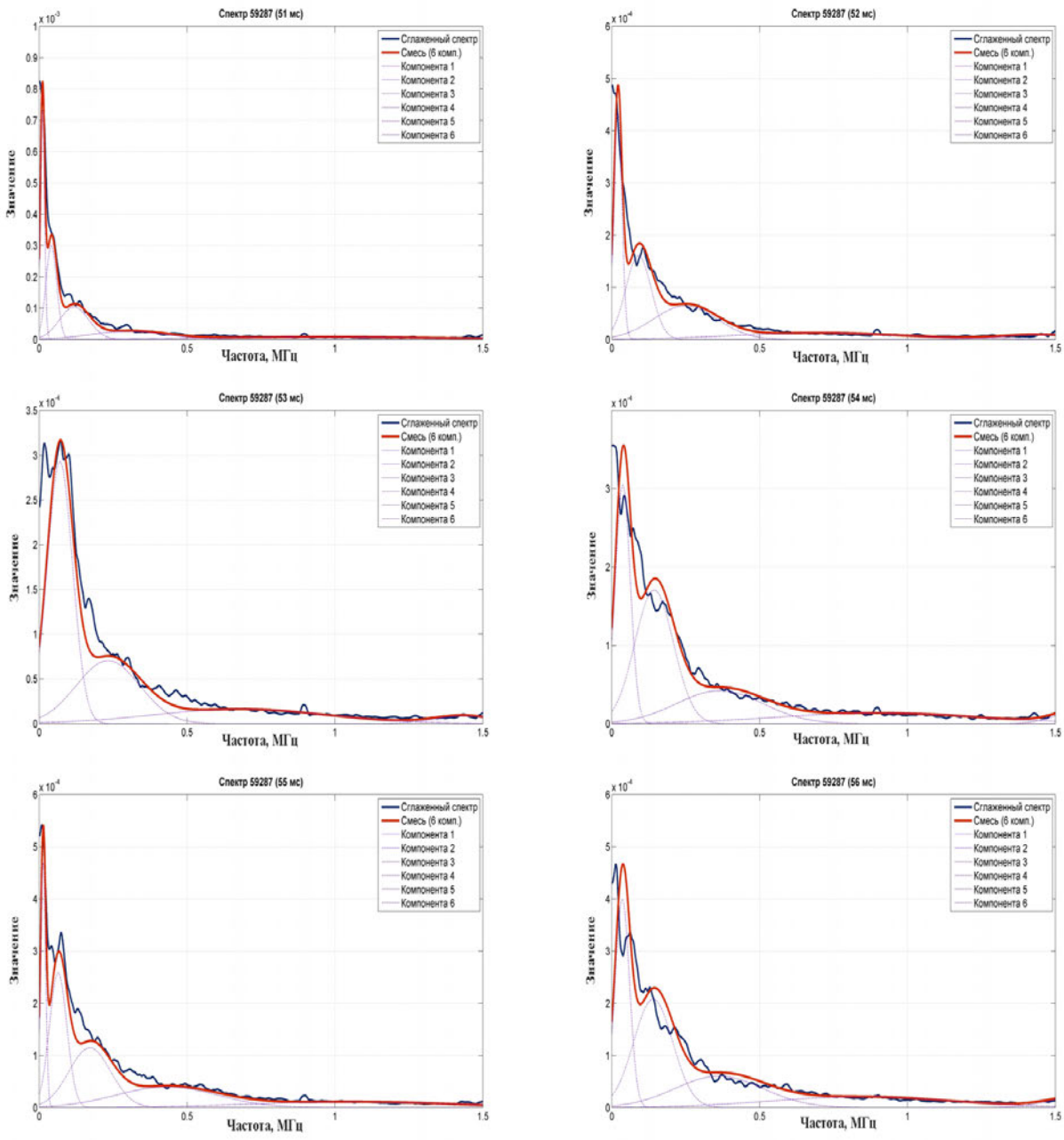

Рис. 5.2. Аппроксимация спектра 59287 с 51 мс по 56 мс

Исходные данные представляют собой односторонний несимметричный Фурье-спектр, убывающий с увеличением частоты. На рисунке 5.2 изображена эволюция спектров (серия номер 59287), измеренных в разряде стелларатора в 51 мс – 56 мс эксперимента. По оси абсцисс отложены значения частот в МГц, по оси ординат – полученные значения спектра. Процедура анализа принципиально не отличается от случая диагностики доплеровской рефлектометрии. Настройки вычислительных методов остаются без изменений. Количество точек анализа, как и ранее, 4096, размер моделируемых выборок в этом случае также составлял 30000 элементов. Снова использованы конечные смеси нормальных законов.

Разложение спектра на компоненты в данном случае дает представление о поведении различных типов колебаний, существующих и эволюционирующих во времени в плазме. Выделяются три доминирующих компоненты, у которых меняется пропорциональный состав. Это соответствует перекачке энергии между турбулентностями различного типа. Также после разложения получаются компоненты с малыми весами. С физической точки зрения их можно отнести к шумам, недостаточному отношению сигнал/шум в измерительном тракте и относительно низкому разрешению по амплитуде исходной временной выборки. Отметим, что подобные результаты получаются и для других серий диагностик, полученных в аналогичном режиме (см. рисунок 5.3, серия номер 59679).

Трехмерные изображения эволюции во времени аппроксимаций спектра позволяют наглядно изобразить приближения для всего эксперимента. Для рассматриваемой серии 59679 пример представлен на рисунке 5.4. Наиболее информативны сечения в моменты времени 51–56 мс, которые изображены на графике сплошными линиями. Аппроксимация поверхностью служит для улучшения визуального восприятия промежутков между ними. Отметим, что программные инструменты построения изображений этого раздела рассмотрены далее в параграфе 7.3.

Для аппроксимации односторонних спектров разумно использовать распределения с носителем, сосредоточенным на положительной полуоси. Применение конечных смесей логнормальных и гамма-распределений для одного из спектров с 59679 для 54 мс продемонстрировано на рисунках 5.5 и 5.6. В отличие от случая нормальных смесей, здесь результирующая кривая выполняет сглаживание спектра. При этом для логнормальных смесей можно отметить хорошее повторение основных особенностей, с выявлением нетривиальных компонент.

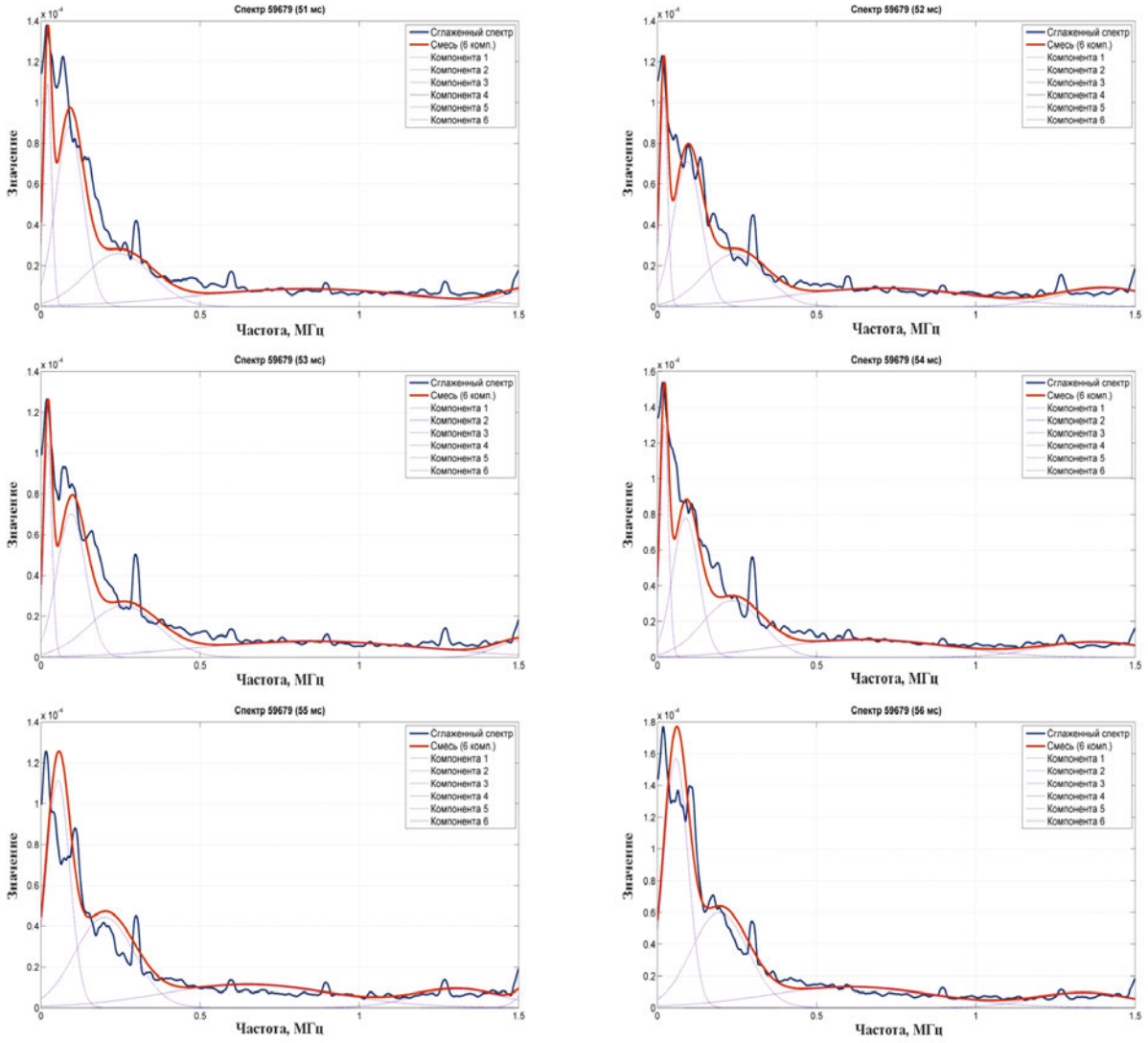

Рис. 5.3. Аппроксимация спектра 59679 с 51 мс по 56 мс

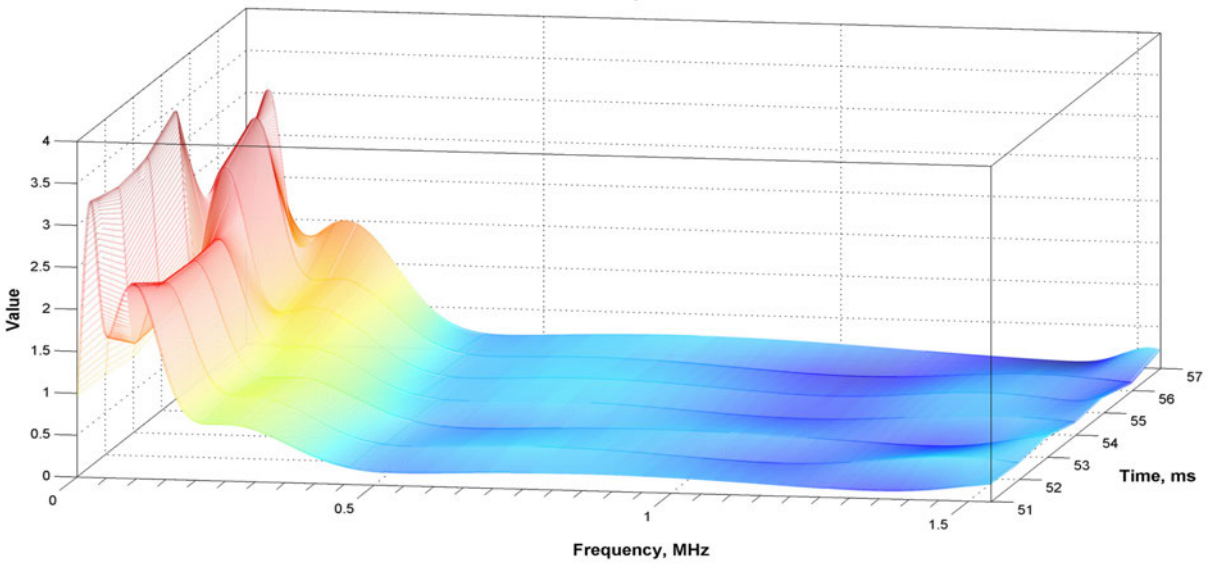

Рис. 5.4. Трехмерная визуализация эволюции во времени аппроксимаций спектра 59679

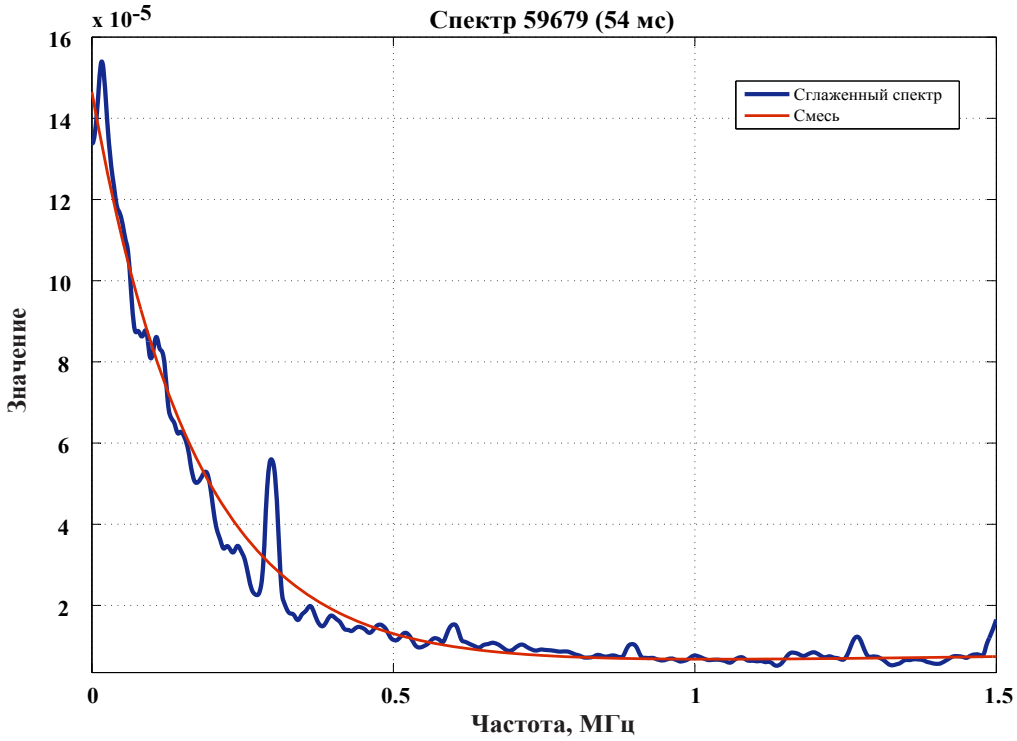

Рис. 5.5. Аппроксимация спектра 59679 (54 мс), гамма-распределения

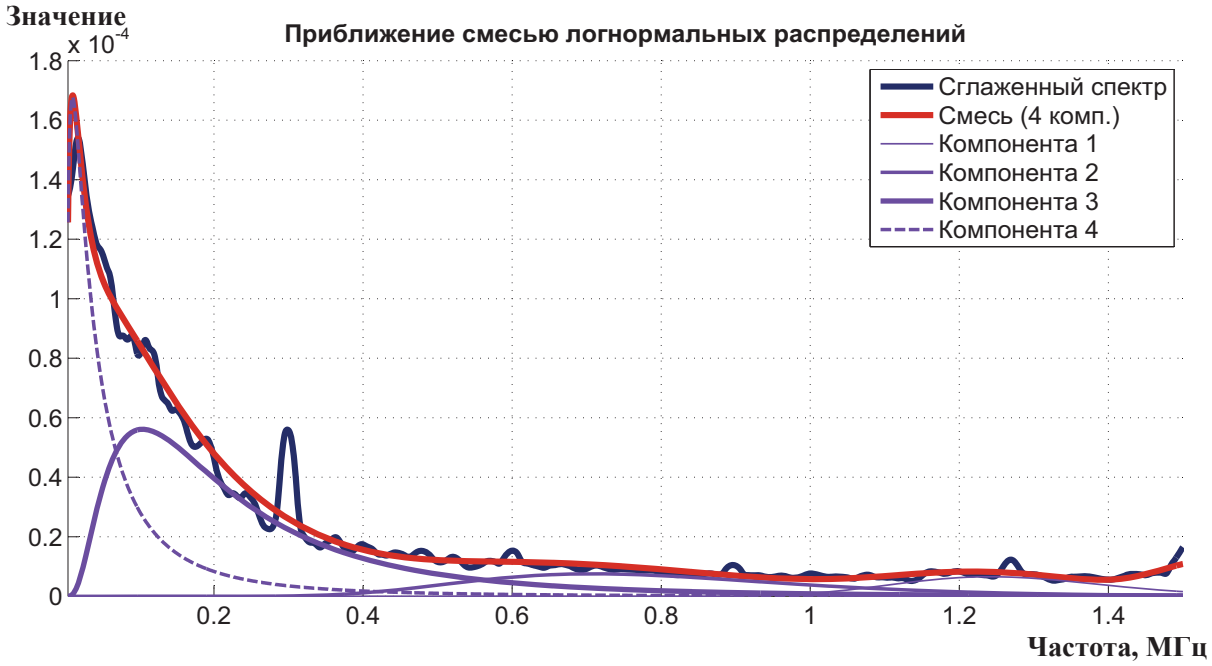

Рис. 5.6. Аппроксимация спектра 59679 (54 мс), логнормальные распределения

Отметим, что конечные смеси гамма-распределений успешно применялись и при непосредственном моделировании процессов в данных совершенно иной физической природы в рамках СРС-метода, например, для биржевой книги заявок [227] и интернет-трафика [229] (также см. раздел **??**).

5.2 Вероятностно-статистический подход к анализу эволюции характеристик микротурбулентности

Традиционно при анализе турбулентного состояния плазмы исследователи пытаются установить связь между скоростями роста неустойчивых режимов, условиями их возбуждения и спектрами флуктуаций, полученными с помощью гирокинетического моделирования или в реальных экспериментах. При этом основное внимание уделяется стационарным режимам, необходимым для работы в устойчивом состоянии будущего управляемого термоядерного реактора, однако нелинейной стадией развития турбулентности, ее насыщения, образования вихрей и их хаотизации обычно пренебрегают. В статье [1] автором была предложена методология анализа статистических характеристик турбулентных пульсаций с использованием конечных нормальных смесей, метода их скользящего разделения и различных модификаций EM-алгоритмов, для изучения физических характеристик при изменении условий возбуждения микроволнового поля. Продемонстрирована высокая согласованность модельных результатов с реальными турбулентными процессами. В данном разделе указанный вероятностно-статистический подход развивается для анализа эволюции характеристик микротурбулентности в переходном процессе при электронно-циклотронном резонансном (ЭЦР) нагреве плазмы стелларатора Л-2М.

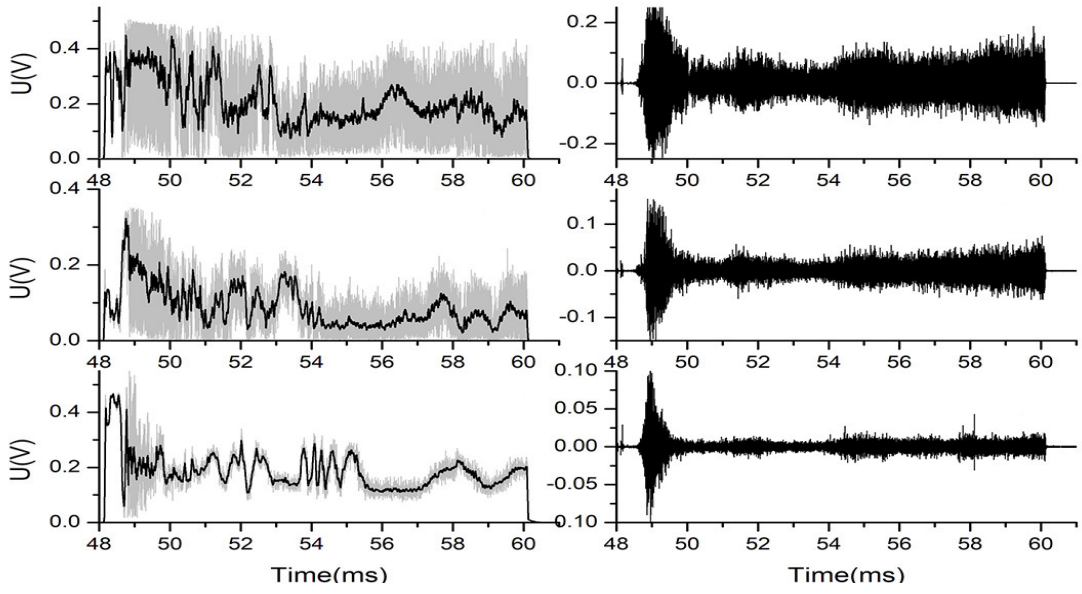

Рис. 5.7. Типичная форма диагностических сигналов

В качестве экспериментальных данных используются ансамбли диагностик, которые учитывают флуктуации плотности плазмы даже в центральных областях плазменного столба. На рисунке 5.7 приведен вид графиков экспериментальных диагностик коротковолновых флуктуаций плотности для различных значений волновых чисел (слева), а также их приращения (справа). Два верхних графика соответствуют значению 0,2 м$^{-1}$, в то время как нижние – величине 0,3 м$^{-1}$. Физическая составляющих указанных экспериментов подробно описана в статьях [146, 147].

5.2.1 Алгоритм анализа физических данных

Из рисунка 5.7 очевидно наличие в данных существенных трендов, поэтому для корректности применения моделей типа (1.7) необходимо осуществить предварительный переход к приращениям. Для них осуществляется запуск СРС-метода с одной из модификаций EM-алгоритма, по полученным оценкам параметров смеси строятся математическое ожидание, дисперсия, коэффициенты асимметрии и эксцесса. Кроме того, предусмотрен вывод вида плотностей (компоненты смеси (1.7)) и их весов для заданных временных меток. Данная процедура описана ниже в алгоритме 5.2.

Алгоритм 5.2. Алгоритм обработки экспериментальных данных

```
1:  function PhysDataProcessing(Data, options)
2:      dData← Diff(Data, options.Diff);      // Переход к приращениям
3:      Params← EMs(dData, options.EM);                  // CPC-метод
4:      // Моментные характеристики (3.9)–(3.12)
5:      [Exp, Var, Skew, Kurt]← Moments(Params);
6:      // Вывод моментных характеристик
7:      PlotMoments(Exp, Var, Skew, Kurt, options.Moments);
8:      // Вывод плотностей для заданного диапазона
9:      PlotPDFs(Params, options.PDF);
10:     return ;
```

Необходимо отметить, что функция `PhysDataProcessing` (алгоритм 5.2) должна быть запущена для каждого обрабатываемого ансамбля. Предполагается, что необходимые эксперименты были проведены до этапа вероятностного анализа, сформированы соответствующие наборы `Data`, проведена необходимая предобработка наблюдений, позволяющая перейти к исследованию соответствующих физических эффектов.

5.2.2 Статистическое определение количества формирующих процессов

В статье [1] с помощью классического и стохастического EM-алгоритмов было впервые определено число процессов, которые формируют исходную ионно-звуковую турбулентность. Оказалось, что несмотря на присутствие различных стохастических факторов, их около 3–5, а увеличение числа компонент в смеси указывает на появление дополнительных процессов в плазме.

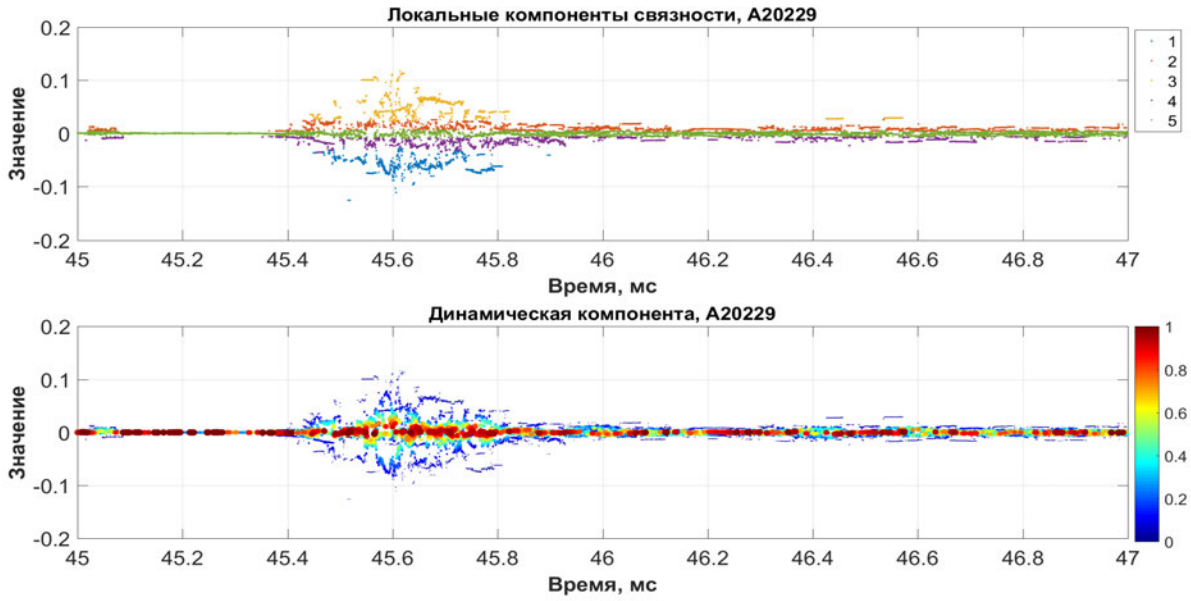

Рис. 5.8. Ансамбль A20229, динамическая компонента (45–47 мс)

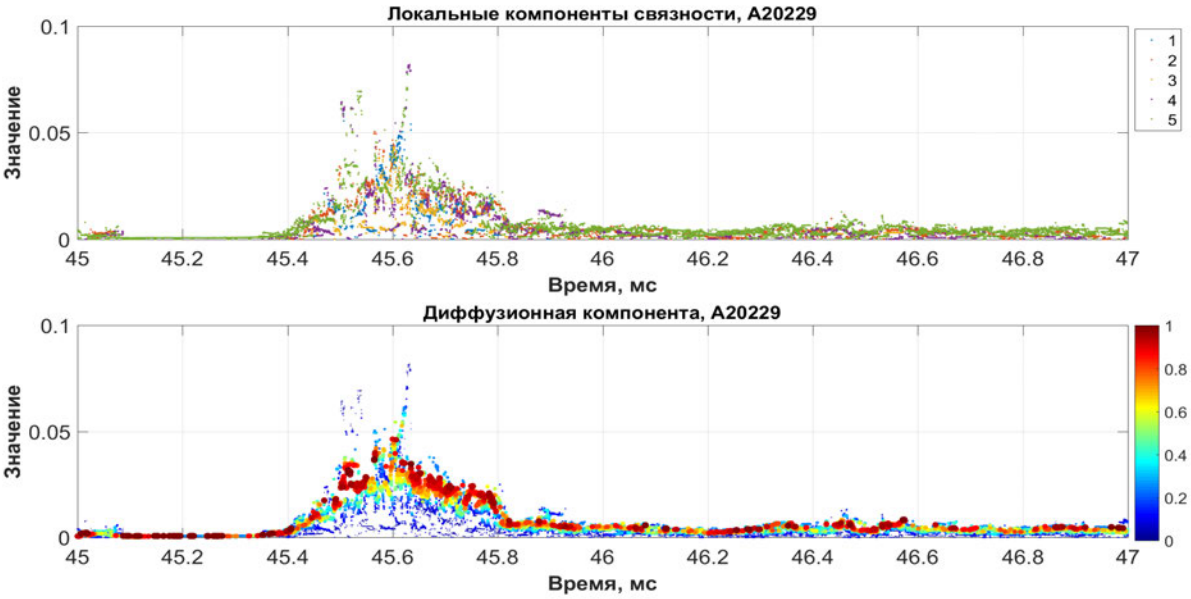

Рис. 5.9. Ансамбль A20229, диффузионная компонента (45–47 мс)

С помощью метода, предложенного в разделе 3.3, продемонстрируем автоматическое определения числа формирующих компонент (и их изменений во времени) для нескольких ансамблей экспериментальных данных. На рисунках 5.8–5.15 приведена визуализация решений комбинации жадного алгоритма и методов кластеризации (верхние графики), а также соответствующий вид оценок CPC-метода для динамической и диффузионной компонент (нижние графики). В процессе анализа используются значения `Params`, полученные применением функции `EMs` (см. строку 3 в алгоритме 5.2).

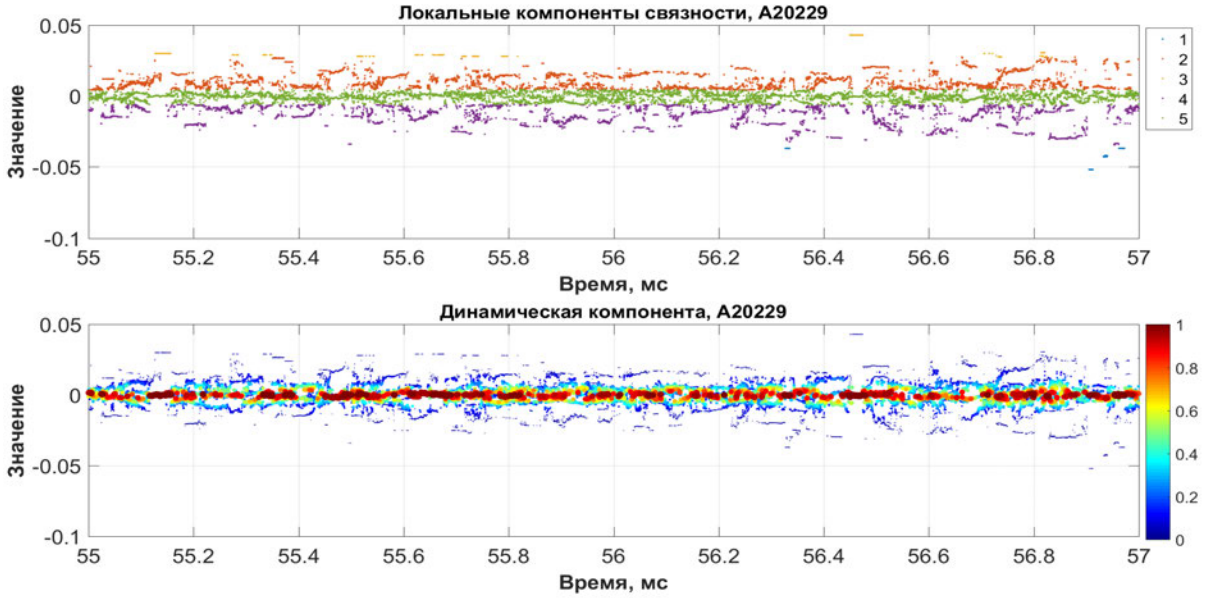

Рис. 5.10. Ансамбль A20229, динамическая компонента (55–57 мс)

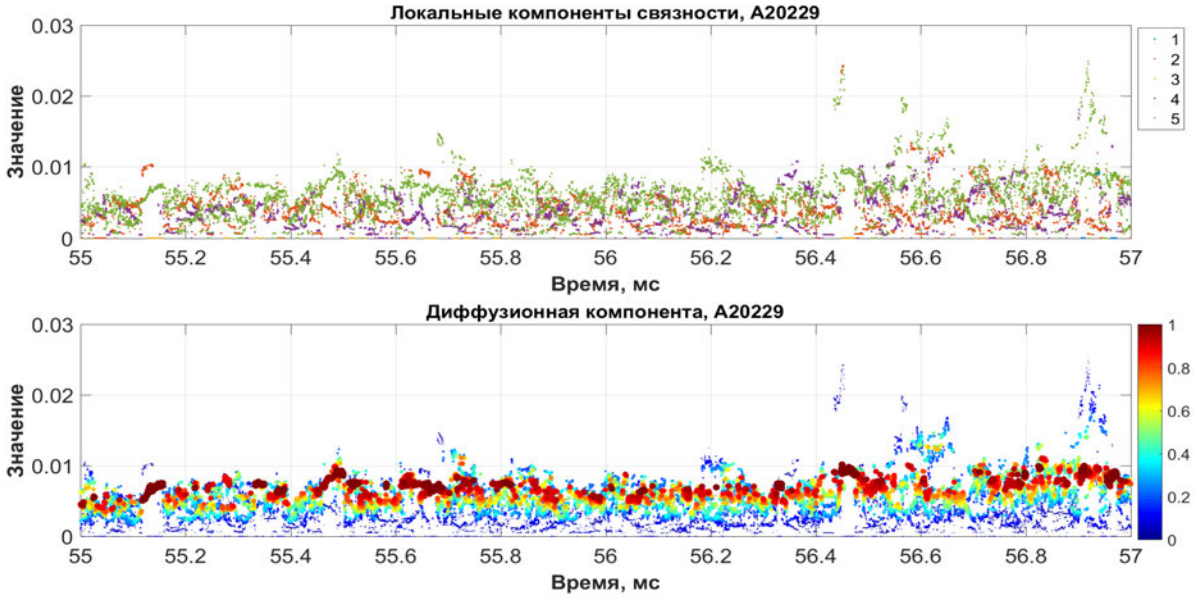

Рис. 5.11. Ансамбль A20229, диффузионная компонента (55–57 мс)

Видно, что в начале эксперимента наблюдается максимальное число процессов, которые достаточно быстро исчезают после прекращения воздействия и лишь кратковременно могут появиться в дальнейшем. Например, на рисунках 5.8 и 5.10 оранжевая компонента с номером 3 присутствует в период 45,4–45,8 мс, затем исчезает и краткосрочно наблюдается, например, в интервалах 46,4–46,6 мс, 55,6–55,8 мс.

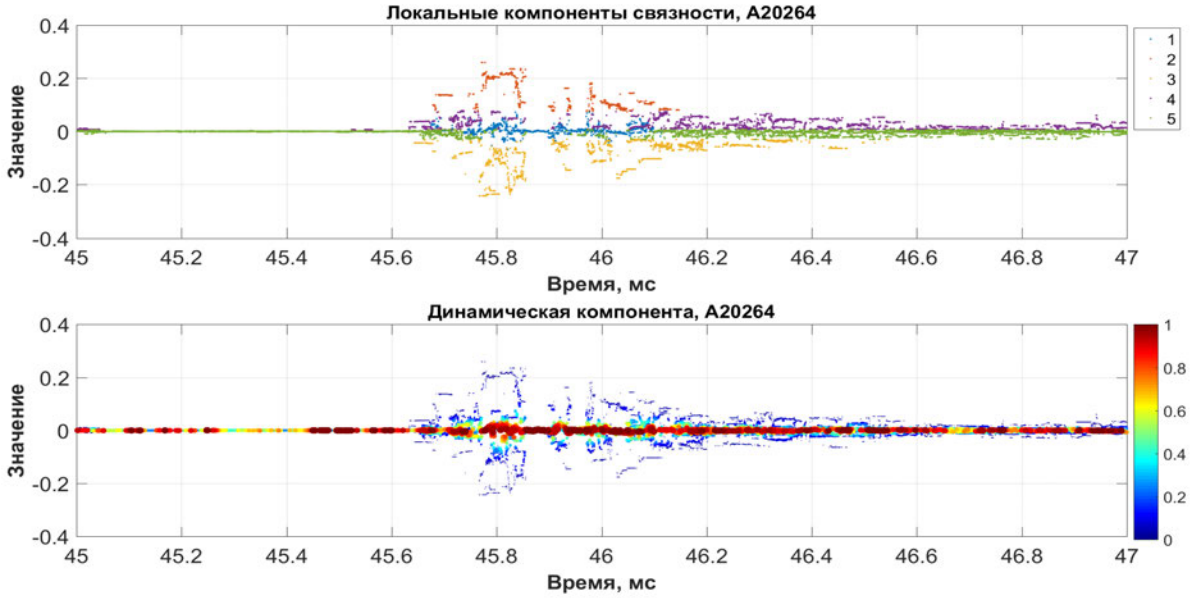

Рис. 5.12. Ансамбль A20264, динамическая компонента (45–47 мс)

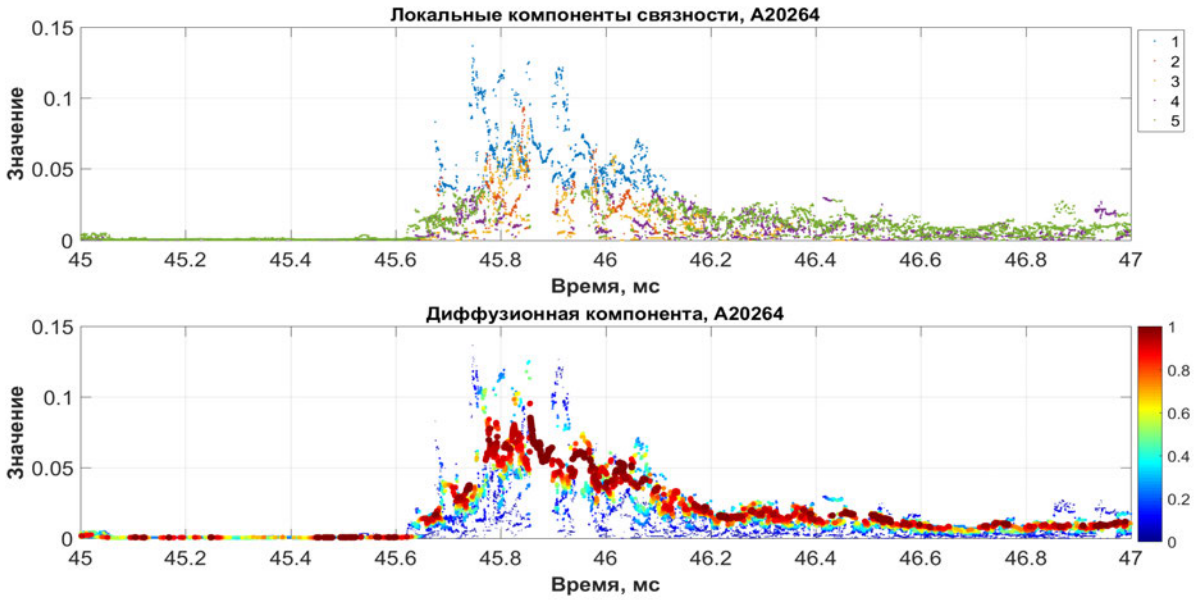

Рис. 5.13. Ансамбль A19692, диффузионная компонента (45–47 мс)

Остальные компоненты (например, зеленая и фиолетовая на тех же графиках) присутствуют более устойчиво, эволюционируя во времени с

точки зрения флуктуаций их величин (в смысле математических ожиданий и среднеквадратических отклонений), а также вклада в общую структуру процесса, описываемого весом (см. нижние графики).

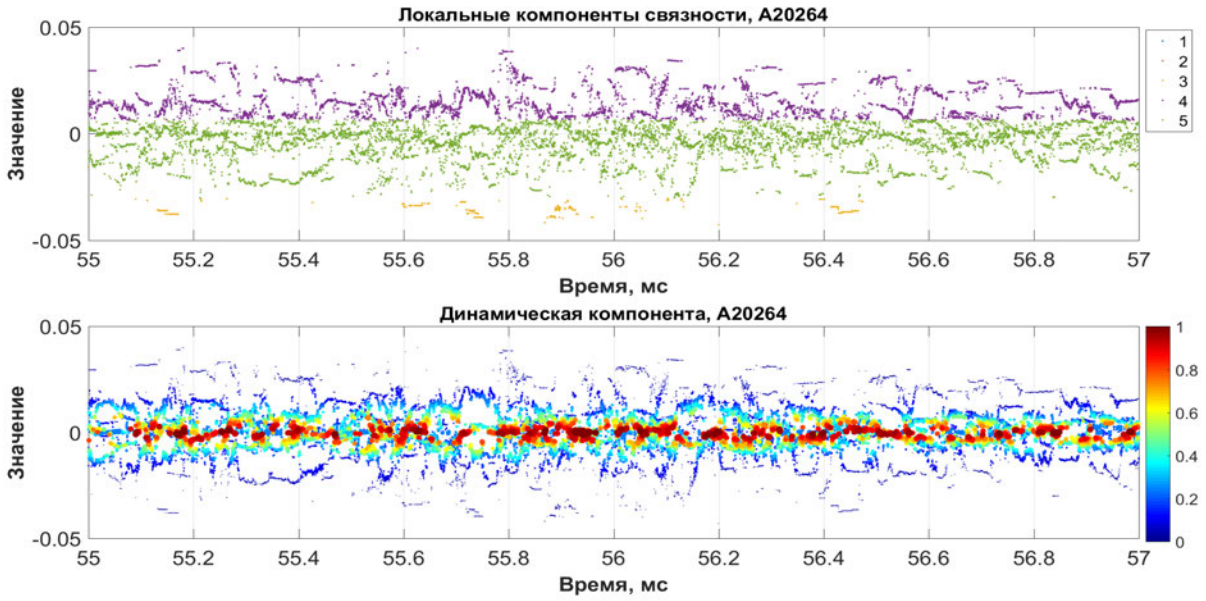

Рис. 5.14. Ансамбль A20264, динамическая компонента (55–57 мс)

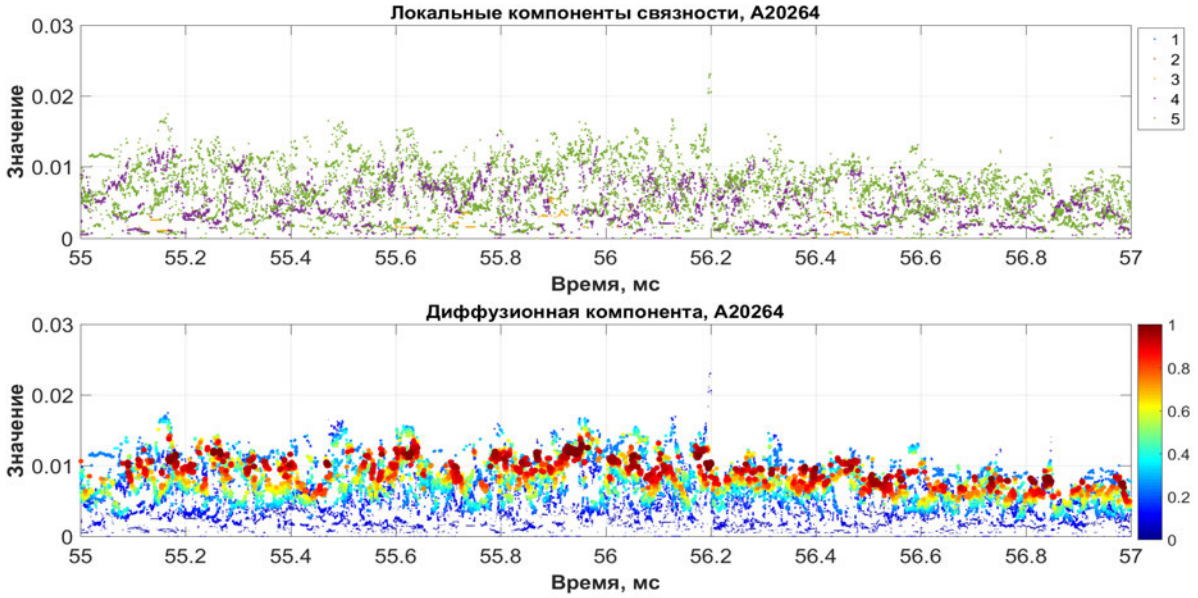

Рис. 5.15. Ансамбль A20264, диффузионная компонента (55–57 мс)

Общее число компонент в процессе автоматического анализа не превышало упомянутых выше 3–5, однако при необходимости оно может быть расширено за счет повышения чувствительности жадного алгоритма за счет выбора порогового значения в формуле (3.28). Аналогичные выводы справедливы и для других рассмотренных ансамблей.

5.2.3 Пример анализ экспериментальных данных

Опишем подробнее переходный процесс, наблюдаемый в эксперименте. Временная эволюция температуры электронов является основным показателем процесса. После включения вторичного гиротрона температура в области поглощения ЭЦР нагрева возрастает до момента времени 54 мс от начала эксперимента, когда она значительно падает, а вскоре после этого снова увеличивается, но медленнее, пока вторичный гиротрон не будет выключен (через 58 мс от начала эксперимента). Рост средней электронной плотности начинается одновременно с падением температуры (в районе 54 мс) и имеет сравнимую длительность, после которой наблюдается лишь незначительное снижение до выключения вторичного гиротрона. Импульсное введение примеси в плазму, вызванное распылением настенного покрытия, начинается через 53 мс от начала эксперимента.

На рисунке 5.16 представлены временная эволюция квадратов флуктуаций плотности и соответствующих приращений для локальных измерений в области гирорезонанса (два верхних графика), в области, сдвинутой на 2 см наружу от гирорезонанса (средний ряд графиков), и для измерений, усредненных по хорде (нижний ряд).

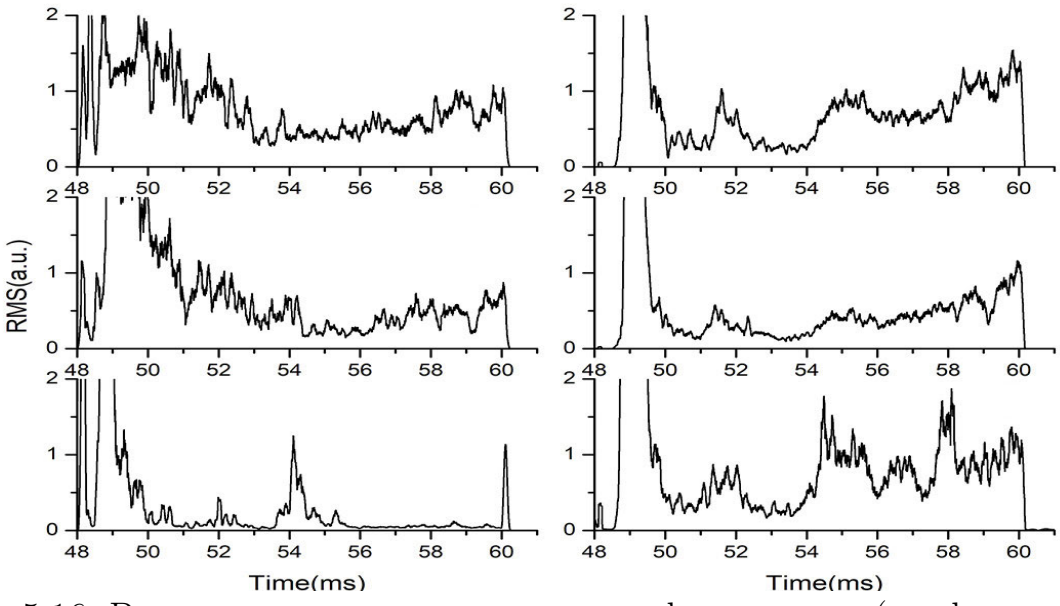

Рис. 5.16. Временная эволюция плотности флуктуаций (графики слева) и приращений (справа)

Все данные были усреднены за 0,1 мс. Реакция усредненных по хорде флуктуаций плотности на изменение параметров плазмы (левый нижний график) представляет собой всплеск высокой амплитуды, который

появляется с задержкой в 2 мс относительно включения вторичного гиротрона, что подразумевает задержку в 1 мс относительно импульсного введения примесей. Уровень флуктуаций плотности после этого становится выше. Поведение уровня флуктуаций плотности в области гироре-зонанса отличается: величина всплесков уменьшается до тех пор, пока вторичный гиротрон не будет выключен, после чего величина всплесков неуклонно возрастает до конца разряда.

Для полного анализируемого интервала эксперимента (50–60 мс) используется аппроксимация приращений четырехкомпонентной нормальной смесью. На рисунке 5.17 изображены плотности – ее компоненты – в некоторые специально выбранные моменты времени:

– до включения вторичного гиротрона (52,00–52,04 мс, первый график сверху);

– после импульсного введение примеси (53,8–53,84 мс, второй график сверху);

– до существенного увеличения величины приращений (54,42–54,46 мс, третий график сверху);

– непосредственно после этого (54,50–54,54 мс, третий график снизу);

– во время устойчивого роста (57,80–57,84 мс and 57,94–57,98 мс, второй и первый график снизу).

Каждая плотность изображается цветом, соотнесенным с ее весом (1.7): от темно-синего для величин, близких к нулю, до темно-красного, соответствующего единице. Точные значения приведены в легендах непосредственно на графиках.

Область эксперимента в 54–56 мс характеризуется значительным отклонением от нормального распределения – это наглядно видно и на графиках эволюции коэффициента эксцесса смеси (см. рисунок 5.18). Для этих величин характерно появление выбросов продолжительностью $0,1$–$0,3$ мс, причем до включения вторичного гиротрона (около 52 мс) наблюдаются более длительные периоды между единичными всплесками, а затем в интервале 54–56 мс наблюдаются более высокие амплитуды (примерно – двукратный рост) выбросов, сопровождающиеся сокращением интервалов между ними. Отклонения от нормального закона в области поглощения ЭЦР нагрева происходит в гораздо более длительные периоды времени, чем в области, охватывающей весь радиус плазмы: возможно, в области нагрева протекают более интенсивные аномальные транспортные процессы по сравнению с интенсивностью, усредненной по всему объему плазмы.

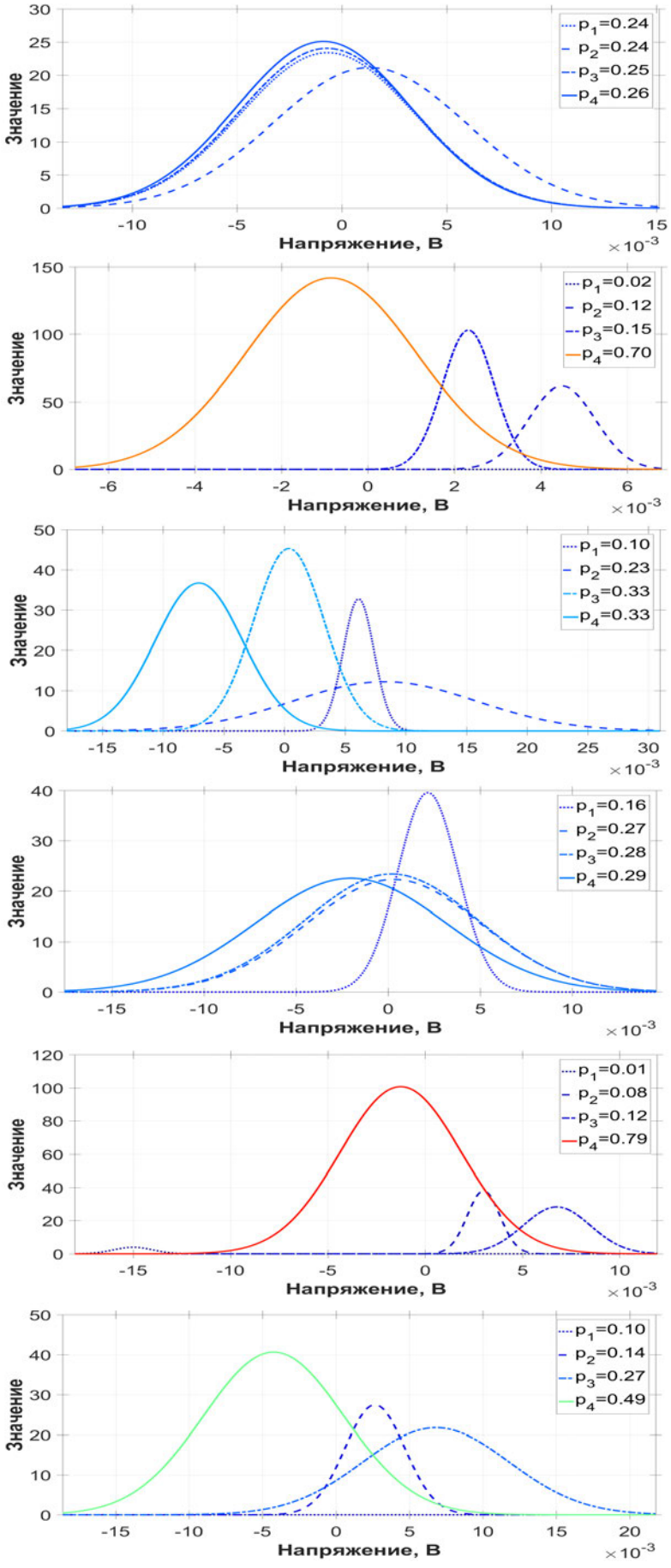

Рис. 5.17. Плотности для временных интервалов 52,00–52,04 мс, 53,8–53,84 мс, 54,42–54,46 мс, 54,5–54,54 мс, 57,8–57,84 мс и 57,94–58,98 мс

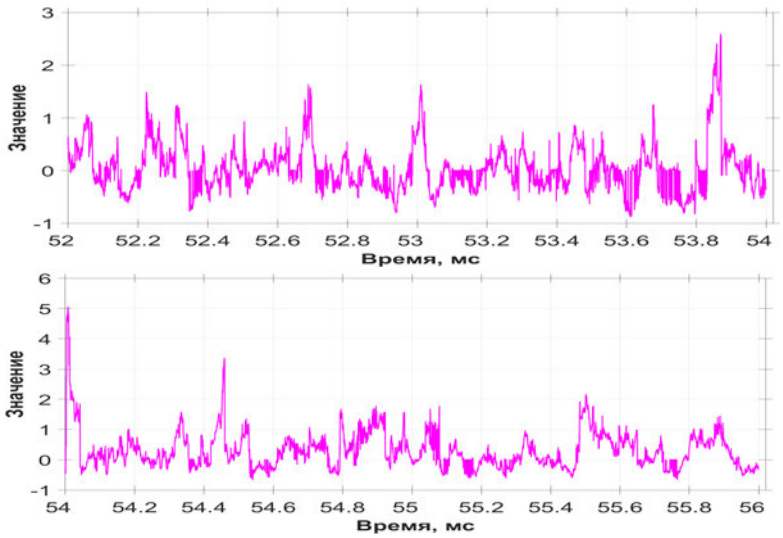

Рис. 5.18. Коэффициенты эксцесса для интервалов 52–54 мс и 54–56 мс

Было предложено оценивать вклад тяжелых хвостов по динамической (синяя кривая на рисунке 5.19) и диффузионной (зеленая кривая) составляющим (см. раздел 2.1).

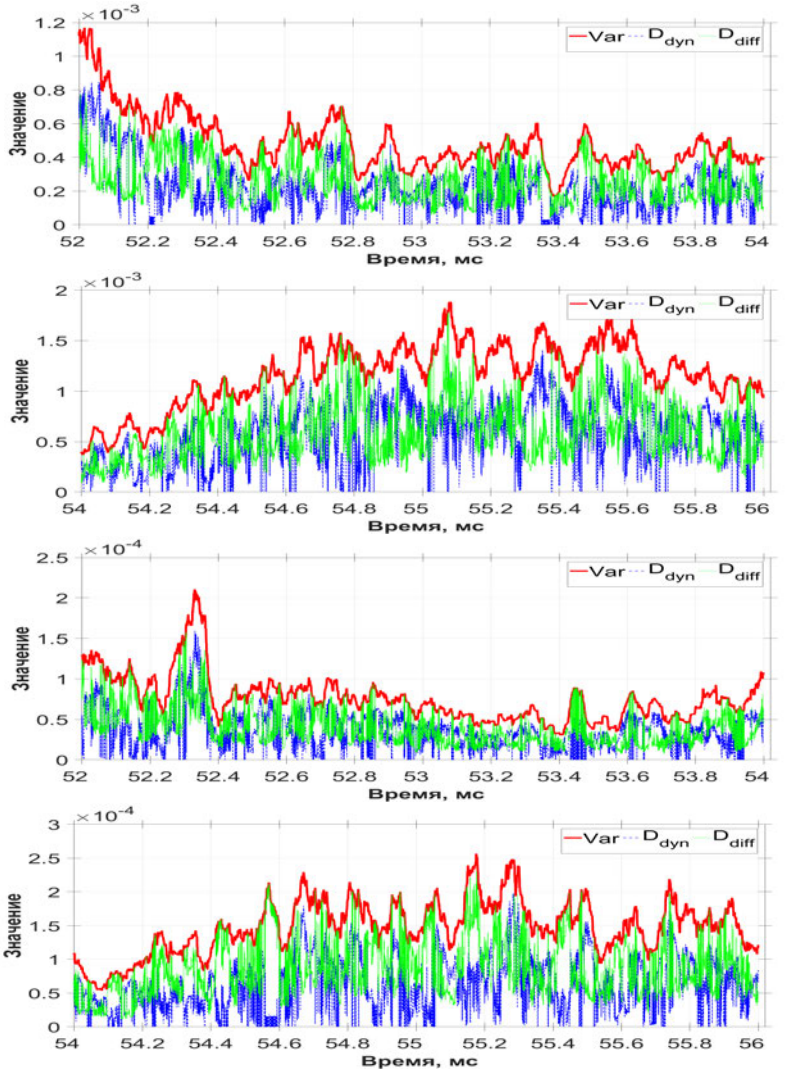

Рис. 5.19. Динамическая и диффузионная компоненты для двух областей

Установлено, что их соотношение остается приблизительно постоянным по всему плазменному разряду, если анализировать значения, усредненные по интервалам в 0,3–0,5 мс. Временная эволюция этих компонент и всей дисперсии близка к эволюции квадратов приращений флуктуаций. Таким образом, данная величина может использоваться наравне с коэффициентом эксцесса.

5.2.4 Прогнозирование экспериментальных данных с расширением признакового пространства

Методы машинного обучения и нейронные сети в исследованиях турбулентной плазмы позволяют добиваться заметных результатов как в вопросах моделирования наблюдаемых явлений [335, 336, 340, 363], так и в задачах анализа и прогнозирования нестабильностей и разрушительных для стеллараторов и токамаков эффектов [282, 350]. В данном разделе рассмотрим подход к решению задачи прогнозирования непосредственно экспериментальных наблюдений с помощью глубокой нейронной сети прямого распространения с двумя скрытыми слоями. Для повышения качества обучения будет предложено расширение признакового пространства за счет использования выборочных моментов, моментных характеристик модели, полученной с помощью алгоритма 5.2 для исходных данных (то есть для функции `Diff` в настройках `options.Diff` должна быть указана соответствующая опция), и описанных выше моментных характеристик для ряда приращений.

Итак, каждому вектору $\mathbf{X}$, который подается в качестве входного в нейронную сеть, ставится в соответствие набор величин $(\mathbb{E}_X, \mathbb{D}_X, \gamma_X, \kappa_X)$, определяемых формулами (3.1)–(3.4) и рассчитанных по наблюдениям, которые отстоят от первого элемента $\mathbf{X}$ (согласно его расположению в анализируемом ряде) на величину скользящего окна СРС-метода. Это позволяет увеличить объем данных для обучения без необходимости расширения экспериментальных выборок. Кроме того, указанные моменты не содержат информацию о том, как ведет себя ряд после данного наблюдения – и поэтому могут быть корректно использованы при построении прогнозов.

На рисунках 5.21–5.23 продемонстрированы полученные результаты прогнозирования на 1 (синяя линия) и 30 (зеленая пунктирная линия) шагов для нескольких экспериментальных рядов в некоторых случайно выбранных диапазонах. Отметим, что ансамбли `A19692` изуча-

лись в предыдущем разделе. Обучение нейронных сетей реализовано на языке программирования `Python` с задействованием гибридного высокопроизводительного вычислительного кластера с двумя процессорами `Power9` с тактовой частотой 2,0 ГГц (20 ядер) и 4 видеокартами `NVIDIA Volta V100` (общий объем памяти 16 Гб).

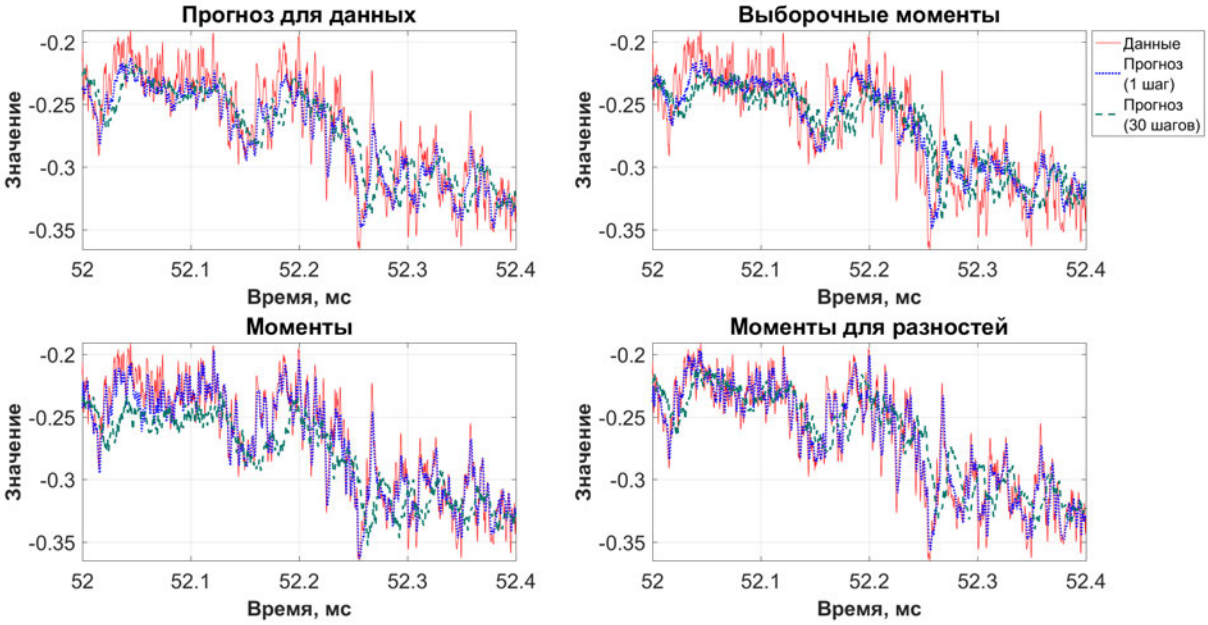

Рис. 5.20. Пример прогноза для ряда A19692 (52,00–52,40 мс)

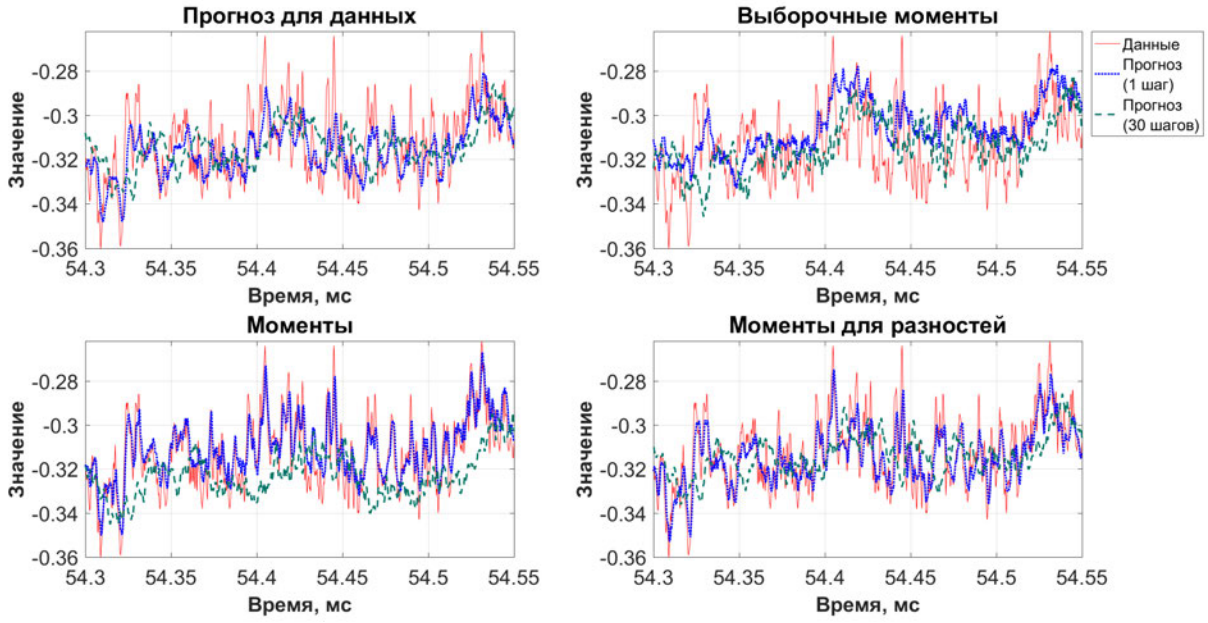

Рис. 5.21. Пример прогноза для ряда A19692 (54,30–54,55 мс)

На рисунках 5.21–5.23 изображены исходные данные и нейросетевые прогнозы на 1 и 30 шагов на основе только исходных наблюдений (слева вверху) и с расширением признакового пространства за счет:

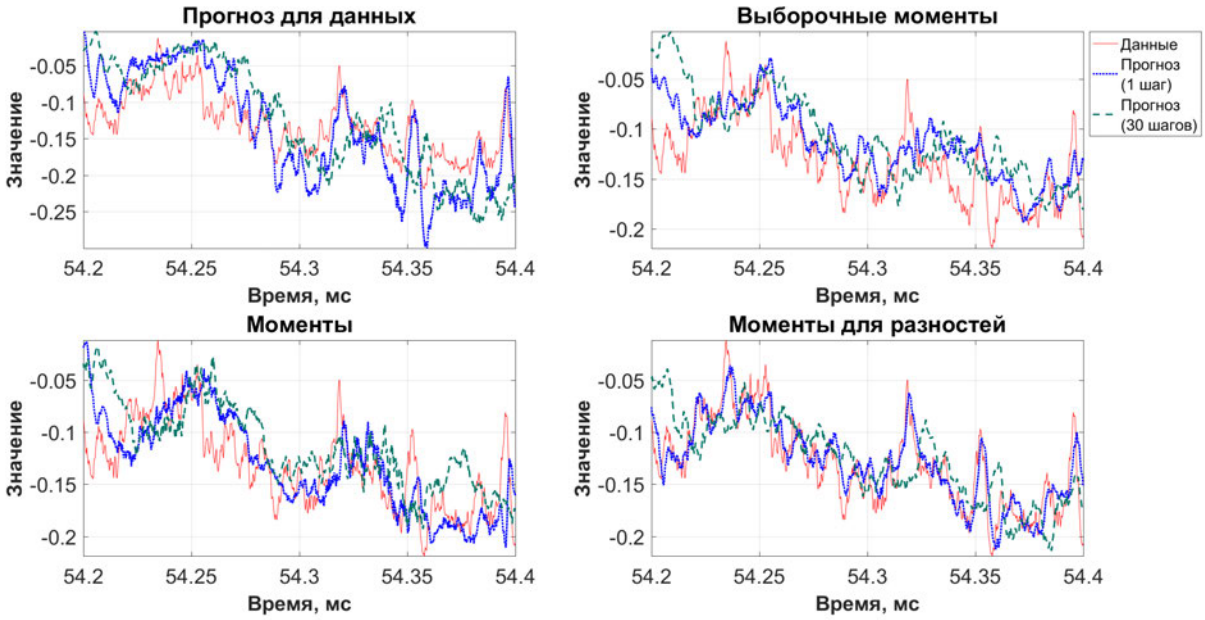

Рис. 5.22. Пример прогноза для ряда A20229 (54,20–54,40 мс)

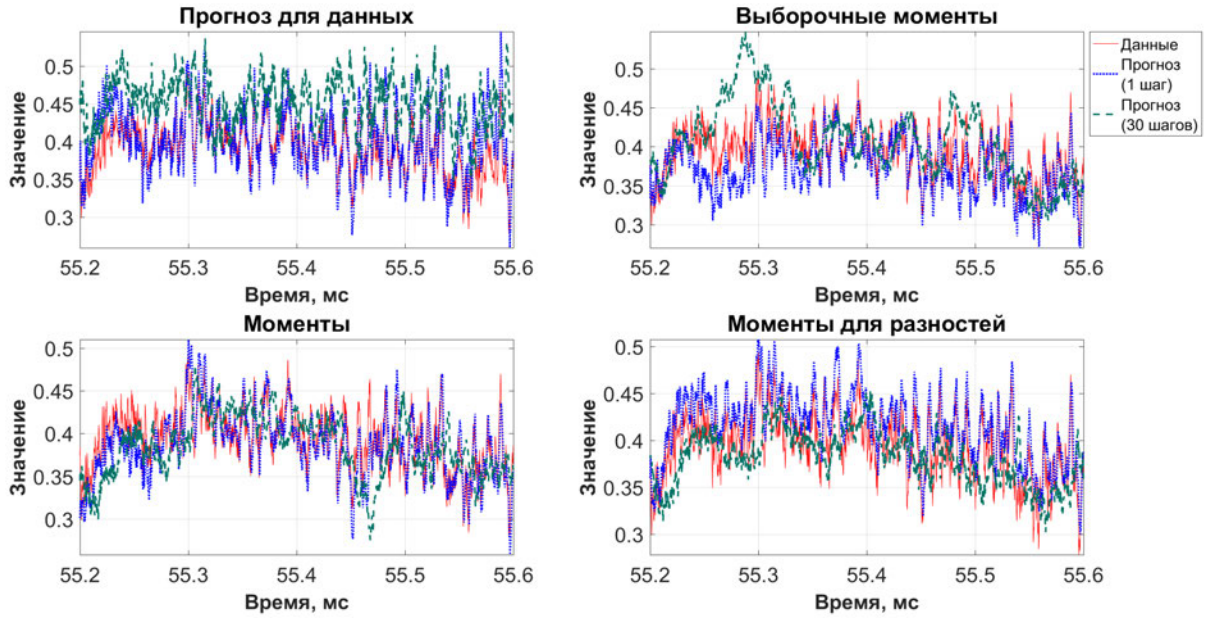

Рис. 5.23. Пример прогноза для ряда A20264 (55,20–55,60 мс)

– выборочных моментов (справа вверху);
– модельных моментов для исходных данных (слева внизу);
– смешанных нормальных моментов для приращений (справа внизу).

На рисунке 5.24 продемонстрирован эффект, получаемый за счет различных расширений признакового пространства. Для всех рядов получено повышение точности прогнозирования (относительно значений метрики `RMSE`) при использовании модельных моментов для приращений исходных наблюдений.

Наибольшая разница (см. рисунок 5.25) наблюдается со случаем, ко-

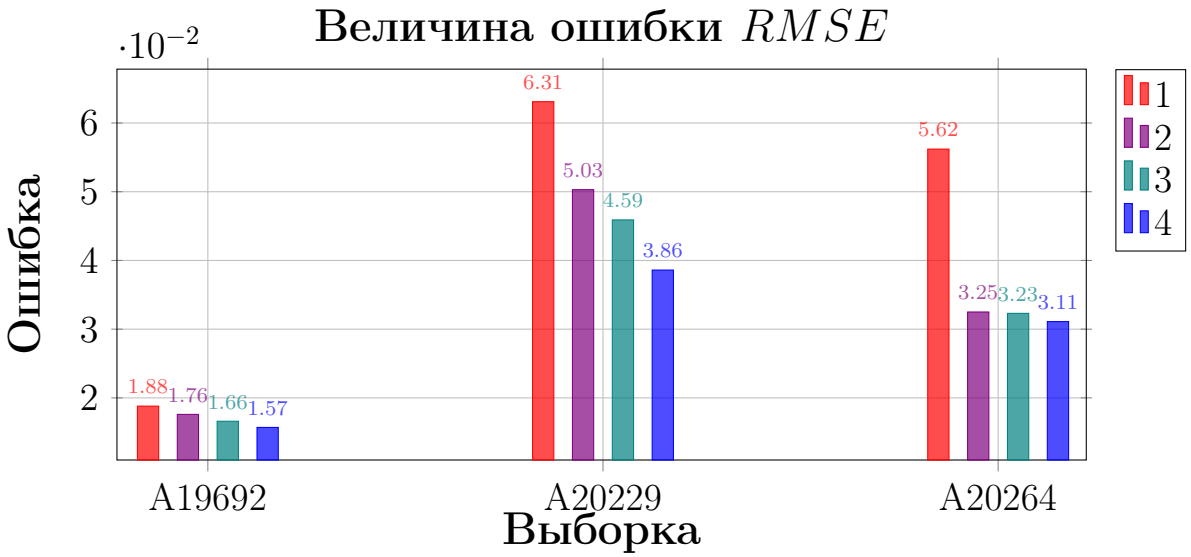

Рис. 5.24. Точность нейросетевых прогнозов исходных данных (1) и с расширенными обучающими наборами: выборочные (2), модельные моменты (3) и моменты смесей для приращений (4)

гда не произведено расширение признакового пространства – в такой ситуации преимущество может доходить до 80,71% (ряд `A20264`), наименьшее полученное увеличение точности составляет 19,75% (ряд `A19692`). По сравнению с выборочными моментами прирост эффективности составляет от 4,5% (ряд `A20264`) до 30,31% (ряд `A20229`).

Рис. 5.25. Прирост точности прогнозов относительно конфигураций для исходных данных (1), с выборочными (2) и модельными (3) моментами

Таким образом, в данном разделе продемонстрирован подход, который на основе построения прогнозов с помощью нейронных сетей для рассматриваемых экспериментальных данных показывает эффек-

тивность использования аппроксимационных моделей типа конечных нормальных смесей (1.7) именно для разностей процесса.

Прогнозирование результатов экспериментов представляет существенный прикладной интерес в следующих задачах обработки турбулентной плазмы:

– верификация рядов, полученных в рамках проведения эксперимента с едиными начальными условиями;

– восстановление сигналов при временном прекращении работы регистрационного оборудования;

– анализ профилей плотности токов увлечения и поглощения электронно-циклотронного нагрева (в частности, для термоядерного реактора ITER [137, 282]).

Предлагаемый подход может быть использован и для верификации смешанной модели, построенной для исходного ряда. Действительно, в случае значительного повышения точности прогноза данных при использовании именно характеристик аппроксимирующей смешанной модели, это может являться индикатором корректности ее выбора в качестве математического описания изучаемого явления.

5.3 Нейросетевое прогнозирование моментных характеристик

В предыдущем разделе было продемонстрировано, что использование смешанных вероятностных моделей позволяет существенным образом повысить качество анализа в плазменных процессах, при этом важную роль играют такие моментные характеристики, как математическое ожидание $\mathbb{E}$ (3.1), дисперсия $\mathbb{D}$ (3.2), коэффициенты асимметрии γ (3.3) и эксцесса κ (3.4). С учетом эволюции во времени предложенных моделей естественным является вопрос о возможности их прогнозирования.

Получение моментных характеристик с помощью СРС-метода требует определенных вычислительных ресурсов, поэтому в ряде ситуаций было бы полезно иметь нейросетевую модель, которая позволила бы строить прогнозы на несколько шагов вперед. В данном разделе обсуждаются возможные пути решения этой задачи. Анализ эффективности производится с задействованием ресурсов упомянутого выше гибридного высокопроизводительного вычислительного кластера архитектуры `Power9`. Это позволяет существенно повысить скорость обучения (в 12–27 раз), преж-

де всего рекуррентных сетей, по сравнению с настольными решениями. Обучение нейронных сетей реализовано с помощью библиотеки глубокого обучения `Keras`, фреймворка `TensorFlow` и языка программирования `Python`.

5.3.1 Задача классификации

Сначала рассмотрим подход, в рамках которого для каждого момента с помощью нейронной сети строится прогноз с точки зрения попадания в некоторый диапазон значений. Определяется отображение вида $f : \mathbf{X} \to \{0, \ldots, k-1\}$, $k \in \mathbb{N}$, которое каждому элементу тренировочной $\mathbf{X}_{train}$ и тестовой $\mathbf{X}_{test}$ выборок ставит в соответствие целое число из диапазона $[0, k-1]$, например, на основе разбиения выборочными квантилями:

– выбираются все уникальные значения из выборки;
– полученный ряд длины l сортируется по возрастанию;
– в диапазон с номером i попадают значения ряда с номерами от $\lfloor (i-1)/(k-1) \cdot l + 1 \rfloor$ до $\lfloor i/(k-1) \cdot l \rfloor$, где $\lfloor x \rfloor$ обозначает ближайшее целое число к x снизу.

Отметим, что данная функция f определяет единое правило для всей исходной выборки $\mathbf{X}$. Таким образом, от исходного ряда с непрерывными значениями осуществляется переход к дискретному. Тогда прогноз x_{pred} на один вперед сводится к присваиванию данной величине числа из диапазона $[0, k-1]$, то есть фактически решается классическая задача k-ичной классификации.

При необходимости построение прогноза на $m \in \mathbb{N}$ шагов вперед, действуя по описанному алгоритму, подобное решение необходимо принять сразу для набора $\{x_{pred,i}\}_{i=\overline{1,m}}$. Такие векторы будем называть паттернами. Отметим, при дискретизации ряда подобные комбинации длины m с некоторой частотой встречаются в преобразованной выборке. В дальнейшем в главе 6 на примере осадков будут рассмотрены примеры вероятностного подхода к анализу данных другого типа, а также проведено сравнение с результатами применения методов машинного обучения и нейронных сетей.

Для решения описанной задачи классификации была использована глубокая нейронная сеть прямого распространения с функцией активации «линейный выпрямитель» (*Rectified Linear Unit*) $\texttt{ReLU}(x) = \max(0, x)$ для входного и скрытых слоев, которая позволяет повысить скорость обучения [216]. Для выходного слоя используется `softmax` (обобщение

логистической функции на многомерного случая), которая представляет j-й выход в виде $y_j = e^{x_j} \cdot (\sum_i e^{x_i})^{-1}$. Выход нейронной сети состоит из k^m нейронов. Для отнесения полученного прогноза к одному из заранее определенных классов дополнительно используется преобразование массива выходных значений нейронной сети в унитарный код длины m – набор, состоящий из $M-1$ нулей и единственной единицы, расположенной в позиции с индексом, совпадающим с номером диапазона.

Количество скрытых слоев и нейронов в них, используемые методы адаптивного повышения скорости обучения [165], оптимизации параметров (`Adam` [285], `NAdam` [195], `AdaDelta` [434], `AdaMax` [165]) и борьбы с переобучением (включая L^2-регуляризацию [406] и дропаут [389]) представляют собой так называемые гиперпараметры нейронной сети, то есть величины, которые не изменяются в процессе обучения. Их число может существенно варьироваться в зависимости от сложности анализируемых данных и выбора соответствующей архитектуры. Для моментов конечных нормальных смесей, полученных при анализе экспериментальных данных турбулентной плазмы, было проведено сравнение результатов (подробнее см. статью [246]), получаемых для различных их комбинаций. В таблице 5.1 приведены результаты точности прогнозирования (k-ичной классификации) на 1, 3, 5 и 10 шагов по входному вектору в 150 наблюдений для различных значений величины k. Отметим, что ошибкой рассматривается несовпадение диапазона (класса), к которому было отнесено очередное наблюдение нейронной сетью с истинным, полученным описанным в начале раздела отображением.

Таблица 5.1. Точность прогнозирования моментных характеристик (k-ичная классификация)

Прогноз	**k**	$\mathbb{E}$	$\mathbb{D}$	$\boldsymbol{\gamma}$	$\boldsymbol{\kappa}$
1 шаг	3	99,7%	95,9%	95,5%	90,7%
	5	98,2%	95,4%	91,3%	86,5%
	10	97,3%	91,5%	87%	77,7%
	15	94,8%	87,3%	83%	72,8%
3 шага	3	99,6%	93,2%	90,9%	84,4%
	5	96,4%	89,9%	86,5%	71,5%
	10	93,8%	82,1%	76,5%	60,5%
5 шагов	3	99,6%	90,6%	87,3%	79,4%
	5	95%	85,1%	80,5%	65,1%
10 шагов	2	99,4%	91,4%	87,5%	75,1%

Таким образом, подобная постановка задачи прогнозирования позволяет определять диапазоны для большинства моментных характеристик с высокой точность ю(вплоть до 99,7%), однако исключается возможность точного восстановления их значений. Кроме того, как видно из таблицы 5.1, использование даже умеренных величин k ведет к росту ошибки, также повышается и время, необходимое для обучения. В следующем разделе будут предложены подходы к решению задачи регрессии, не использующие дискретизацию исходного ряда.

5.3.2 Задача регрессии

В данном разделе для решения задачи прогнозирования непрерывных значений моментных характеристик рассматриваются следующие архитектуры нейронных сетей:

(I) один скрытый слой, 60 нейронов;
(II) один скрытый слой, 100 нейронов;
(III) два скрытых слоя по 20 нейронов в каждом;
(IV) два скрытых слоя по 50 нейронов в каждом;
(V) два скрытых слоя по 100 нейронов в каждом;
(VI) три скрытых слоя по 20 нейронов в каждом;
(VII) три скрытых слоя по 50 нейронов в каждом.

Для каждой из них рассматривается реализация как в виде сети прямого распространения, так и с добавлением `LSTM`-слоев [255] (Long-Short Term Memory) – разновидности рекуррентных нейронных сетей, успешно зарекомендовавшей себя при решении задач обработки и прогнозирования различных временных рядов. Будет рассмотрена задача прогнозирования величины следующего наблюдения по 50 предшествующим наблюдениям. В качестве критерия качества построенного прогноза выбраны величина среднеквадратической ошибки `RMSE` для нормализованных (все наблюдения принадлежат сегменту $[0, 1]$) данных и функция потерь. В качестве метода оптимизации используется `Adam`, функции активации – `ReLU` для сетей прямого распространения и гиперболический тангенс и рациональная сигмоида $x/(1+|x|)$ для рекуррентных случаев. Подробнее выбор конфигурации и настройки методов описаны в статье [72].

На рисунке 5.26 представлены величины ошибок `RMSE` моделей, полученных в результате обучения 14 архитектур на основе базовых типов I–VII. Символ «r» рядом с римскими цифрами используется для обозначения рекуррентной модификации `LSTM`.

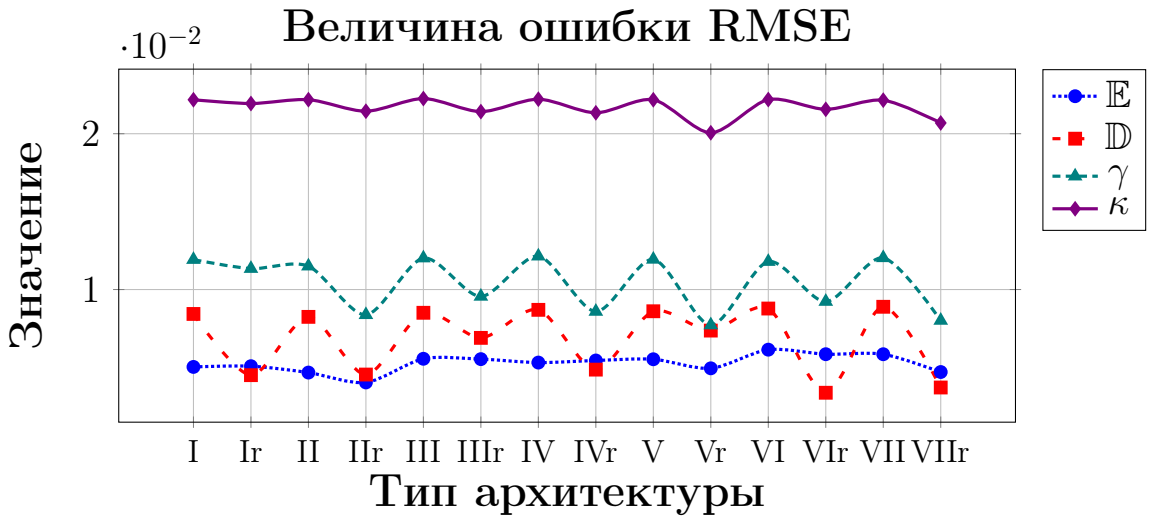

Рис. 5.26. Сравнение величины среднеквадратичных ошибок

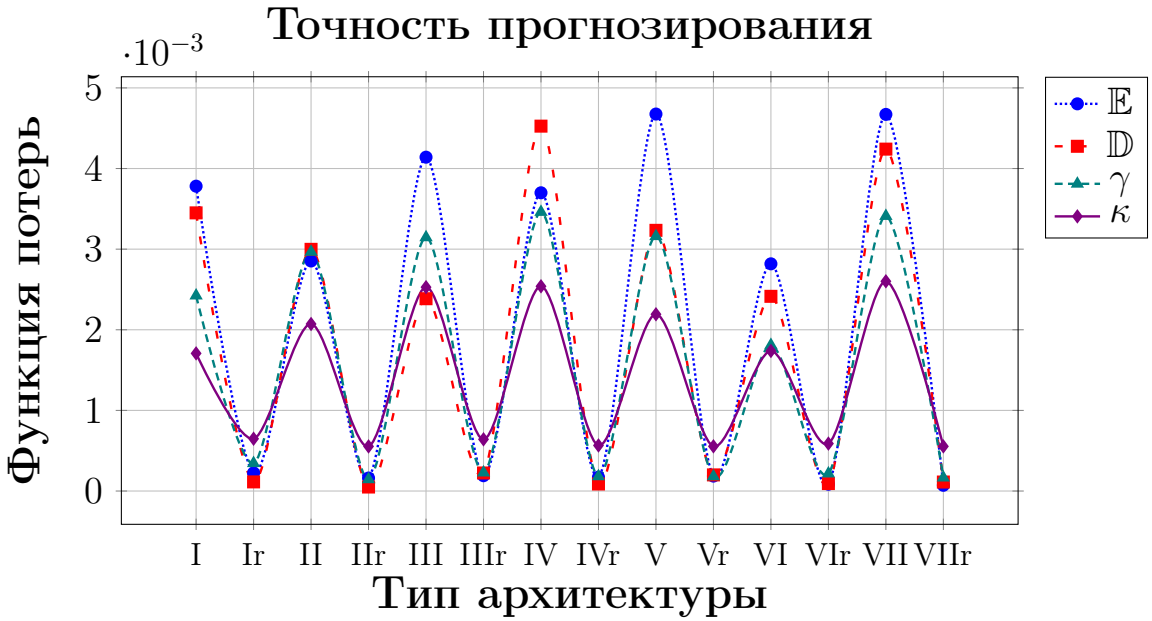

Рис. 5.27. Сравнение величины функции потерь

Рис. 5.28. Коэффициент эксцесса и прогнозы архитектуры VIIr

Использование рекуррентных архитектур во всех случаях несколько уменьшает значение ошибки, в среднем для всех рядов [72] – в 1,33 раза. Для математического ожидания и коэффициента эксцесса на лучших LSTM-архитектурах ошибка в среднем меньше на 10–20%, а для дисперсии и коэффициента асимметрии – на 30–60%. Этот эффект более наглядно проявляется для функции потерь (см. рисунок 5.27). В среднем для всех рядов в данной метрике разница составляет 20,3 раза, а в отдельных случаях (математические ожидания для конфигураций VII и VIIr) получается более чем 65-кратное уменьшение величины ошибки.

Можно выделить рекуррентные конфигурации IIr и VIIr, в которых для всех моментных характеристик сразу в обеих метриках получены либо наименьшие среди всех, либо близкие к этому значения. Таким образом, применение рекуррентных архитектур ведет к значительному повышению качества обучения в любой из рассматриваемых метрик. Для иллюстрации качества приближения данных обученными моделями на рисунке 5.28 продемонстрированы значения для коэффициента эксцесса (нормализованные данные) и аппроксимация ряда с помощью предсказаний, сделанных с применением рекуррентной архитектуры VIIr. Заметим (см. рисунок 5.26), что именно для четвертого момента величина ошибки для всех архитектур является наибольший, при этом можно утверждать (см. рисунок 5.28), что качество приближения моделью исходных данных является достаточно высоким.

Использование рекуррентных сетей позволило существенным образом повысить качество аппроксимации исходных данных. Однако усложнение конфигурации влечет за собой и дополнительную вычислительную нагрузку при обучении. В среднем сети прямого распространения обучались за 664 эпохи, в то время как рекуррентные модификации – за 687. Для рассматриваемых рядов было установлено, что для LSTM-конфигураций необходимое время обучения в среднем превышает результаты для классических в 47 раз (минимальное значение – 6 раз, максимальное – 90) в зависимости от архитектуры и анализируемого ряда. Для конфигураций IIr и VIIr лучшее время показывает именно первая – она обучается быстрее в среднем в 1,83 раза для одной моментной характеристики. Таким образом, архитектура IIr может быть использована в задачах, для которых наиболее критично быстродействие, а не точность аппроксимации.

5.3.3 Векторные прогнозы

В предыдущем разделе было представлено решение задачи регрессии для моментных характеристик, однако обсуждался прогноз только на один шаг вперед. Очевидно, для приложений наибольший интерес представляют более существенные периоды, например, 10–70 предсказаний по входным данным в 300–500 элементов. Ниже будет предложены соответствующие архитектуры. Кроме того, достаточно очевидно, что прогнозируемые ряды между собой связаны, так как описывают эволюцию одного и того же процесса. Поэтому достаточно естественной оказывается идея перехода к совместным (векторным) прогнозам для этих величин.

Эксперименты с глубокими сетями прямого распространения с 1–3 скрытыми слоями с различным количеством нейронов в каждом (50, 100, 150 и 200) показали, что время, которое необходимо для обучения одной «векторной» архитектуры для построения одношагового прогноза в среднем превосходит в 1,56 раза аналогичное значение для какого-либо одного момента, а точность получаемого значения (в терминах средней абсолютной ошибки `MAE`, Mean Absolute Error) возрастает на 8,2%. Таким образом, помимо ожидаемой взаимосвязи между рядами, в пользу «векторных» архитектур свидетельствуют ускорение вычислений по сравнению с последовательной обработкой отдельных рядов, при этом общая точность прогнозов для каждого ряда только увеличивается. Процедура тонкой настройки гиперпараметров в данном случае подробно описана в статье [247].

Для построения среднесрочных прогнозов воспользуемся следующими рекуррентными `LSTM`-архитектурами:

(Iv) один скрытый слой: 100 нейронов;

(IIv) два скрытых слоя: 150 и 100 нейронов;

(IIIv) три скрытых слоя: 200, 150 и 100 нейронов.

Теперь на вход каждой нейронной сети подается $4 \cdot N$ наблюдений, где N – ширина окна, на основе которого делается предсказание (например, $300, 400, 500$), на выходе получается вектор из – $4m$ наблюдений, где m – выбранная длина прогноза (например, $10, 30, 50, 70$). Используемая конфигурация гиперпараметров достаточно близка к описанной в разделе 5.3.2. Ее детали могут быть найдены в статье [75]. Для анализа результатов прогнозирования использованы классические метрики `RMSE` и `MAE`, при этом производится преварительная нормализация данных.

По результатам анализа 36 различных конфигураций были сделаны следующие выводы. Наибольший прирост точности получается в результате перехода от архитектур Iv с одним скрытым слоем к архитектурам IIv с двумя скрытым слоями: в среднем ошибка `RMSE` уменьшается на 23%, а `MAE` – на 28%. Однако прямым следствие подобного перехода становится повышение длительности обучения в среднем на 44%. Добавление еще одного скрытого слоя (то есть переход к архитектурам IIIv) увеличивает время обучения еще на 14%, при этом дополнительный выигрыш в точности для метрики `RMSE` составляет 3%, а для `MAE` – 2%. В ряде случаев ошибка может даже несколько возрастать.

На рис. 5.29 представлены примеры графиков моментов и сделанных прогнозов на 1 и 50 шагов для наблюдений с 8500 по 9000 (в терминах положения скользящего окна) для архитектуры IIv. Спрогнозированные ряды хорошо приближают исходные (даже с учетом их явной нестационарности), при этом существенные изменения локальных трендов (выбросы) ведут к естественному увеличению ошибки.

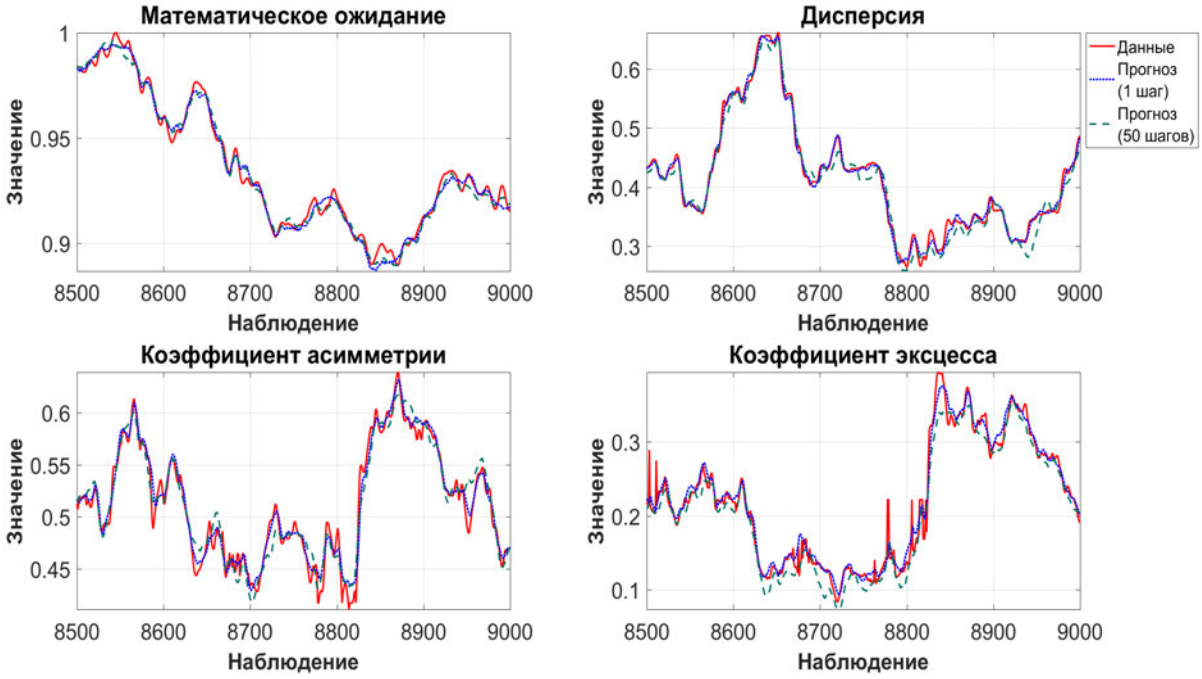

Рис. 5.29. Сравнение значений исходных данных (1), прогноза на 1 шаг (2) и на 50 шагов (3) для четырех моментов

Процедура нейросетевого прогнозирования моментных характеристик смесей представлена в алгоритме 5.3.

Параметры `options.TrainTest` позволяют определить произвольное разбиение на тестовую и тренировочную части для исходной выборки. Конфигурации архитектуры задаются набором опций

Алгоритм 5.3. Нейросетевое прогнозирование моментов

```
1: function ForecastMoments(Params, options)
2:     [𝔼, 𝔻, γ, κ]← Moments(Params);
3:     // Нормализация моментов
4:     NMoments← Normalization(𝔼, 𝔻, γ, κ);
5:     // Разделение на тренировочную и тестовые выборки
6:     [Train, Test] ← TrainTestData(NMoments, options.TrainTest);
7:     // Конфигурация архитектуры и гиперпараметров
8:     NN← Architecture(options.HyperParams);
9:     NNModel← NNTrain(NN, Train, Test);
10:    return NNModel;
```

`options.HyperParams` – и число слоев, и методы оптимизации, и `LSTM`-слои. Таким образом, данный алгоритм охватывает все рассмотренные в разделе 5.3 случаи.

Глава 6

Модели и методы анализа экстремальных явлений в метеорологии и океанологии

Изучение закономерностей и тенденций, связанных с выпадением осадков, прежде всего, экстремальных, является важной задачей в метеорологии. Во-первых, осадки являются значимыми параметрами климатологических и гидрологических моделей, а значит, необходимо анализировать усредненные объемы осадков, кумулятивные данные за каждый дождливый период или ежедневные наблюдения. В последнем случае результаты достаточно сильно зависят от точности измерений, более чувствительны к наличию пропусков [446]. Во-вторых, экстремальные по величине объемы осадков, особенно наблюдавшиеся в течение относительно короткого периода времени, ведут к появлению различных стихийных бедствий – наводнений, селевых потоков, оползней. Наконец, их изучение является крайне важным с точки зрения исследования процессов изменения климата [130, 142, 186, 193, 220, 256, 257, 284, 304, 338, 408]. Анализ и высокоточное прогнозирование подобных явлений необходимы для обеспечения безопасности и сохранения человеческих жизней.

При этом не существует однозначного определения того, что является действительно экстремальными осадками. Для различных климатических зон (и даже отдельных регионов в их пределах), одни и те же значения могут рассматриваться как вполне умеренные, так и приводить к катастрофическим последствиям. Наиболее распространенным

на практике является подход, основанный на классических результатах теории экстремальных значений. А именно, аномальными признаются величины, превышающие заданный для данной местности порог, который обычно определяется как квантиль некоторого уровня, например 0,95, для выбранного распределения [257]. Подобный алгоритм обычно называют методом превышения порогового значения [308].

Очевидно, что корректность выбора упомянутого распределения определяет и адекватность решений на основе соответствующего порогового значения. При этом разные модели будут давать отличающиеся друг от друга результаты даже для одних и тех же данных. В частности, изменения в максимальных значениях объемов осадков (и, соответственно, большее число превышений порога) не всегда означают необходимость признания их экстремальными – например, увеличение доли осадков подобного рода может объясняться увеличением интенсивности в сочетании с уменьшением числа дождливых дней [442, 444].

В данной главе результаты разделов 1.3, 3.1 и 3.3 используются для создания вероятностных моделей и методов исследования метеорологических (осадки и их интенсивности) и океанологических (турбулентные потоки тепла между океаном и атмосферой) данных. Особое внимание уделяется вопросам выявления экстремальных наблюдений в рассматриваемых пространственно-временных рядах. Используются как статистические подходы для оценивания неизвестных параметров, так и широкий набор алгоритмов машинного обучения и нейронных сетей для решения задач заполнения пропусков и прогнозирования.

6.1 Анализ осадков с использованием исторических паттернов

Построение адекватных вероятностно-статистических моделей для осадков является важной задачей, например, для анализа процессов в регионах, в которых сети датчиков не покрывают полностью зоны необходимого наблюдения [395]. Такие модели могут быть использованы для решения задач вероятностного прогнозирования [218, 302], формирования статистических сценариев [354].

Особенностью данных об осадках является наличие в них участков с подряд идущими нулевыми либо положительными значениями, которые называются «сухими» и «дождливыми» периодами. Обозначая через «D» или `Dry` отдельные значения в первых из них, через «W» или `Wet`

во вторых, получим цепочки следующего вида:

$$\ldots - D - \underbrace{W - W - W - \ldots - W}_{\text{«дождливый» период}} - D - \ldots$$

$$\ldots - W - \underbrace{D - D - D - \ldots - D}_{\text{«сухой» период}} - W - \ldots$$

Любые подмножества последовательностей из символов «D» и «W» образуют исторические паттерны, которые в данном разделе будут использованы для проведения анализа независимости наблюдений, а также прогнозирования наличия или отсутствия осадков в тот или иной день, а также их объемов. В качестве тестовых данных используются суточные наблюдения об объемах осадков за период 1950–2006 гг. в городах Потсдам и Элиста [443].

6.1.1 Дискретизация данных

Воспользуемся подходом к преобразованию исходных непрерывных данных к дискретным, описанным в разделе 5.3.1. Сначала будем считать соответствующую величину k равной двум. Рассмотрим преобразование исходных объемов суточных осадков V_{daily}, представляющих собой неотрицательные данные, по следующему правилу: если в данный i-й день наблюдалась какая-либо положительная величина, то она заменяется на единицу ($\widetilde{V}_{daily}^{(i)} = 1$«W»), иначе $\widetilde{V}_{daily}^{(i)}$ остается равной нулю (заменяется на «D»). Таким образом, исходный ряд, состоящий из непрерывных значений, становится дискретным, принимающим два возможных значения $\{0, 1\}$ (или «D» и «W», для большей наглядности). Данное упрощение позволяет анализировать непосредственно наличие или отсутствие осадков безотносительно к их объему, то есть решать задачу классификации наблюдений.

Для каждого набора, составленного из символов «D» и «W», в рамках исторических данных можно определить частоты его появления как отношение числа таких паттернов фиксированной длины N к общему числу возможных последовательностей 2^N, то есть фактически получить значения вероятностей (согласно классическому определению). Для Потсдама и Элисты проанализированы наблюдения за почти 60 лет для значений параметра N от 1 до 14, получены значения частот (вероятностей), определен паттерн с максимальным значением. Примеры для трех- и пятидневных паттернов для Потсдама и Элисты представлены в таблице 6.1 и на рисунках 6.1 и 6.2.

Таблица 6.1. Пример: значения частот для трехдневных паттернов

Паттерн	**Частоты**	
	Потсдам	**Элиста**
`Dry-Dry-Dry`	**0,3247**	**0,4864**
`Dry-Dry-Wet`	0,0930	0,1021
`Dry-Wet-Dry`	0,0582	0,0751
`Dry-Wet-Wet`	0,0956	0,0660
`Wet-Dry-Dry`	0,0930	0,1022
`Wet-Dry-Wet`	0,0609	0,0390
`Wet-Wet-Dry`	0,0957	0,0660
`Wet-Wet-Wet`	0,1789	0,0631

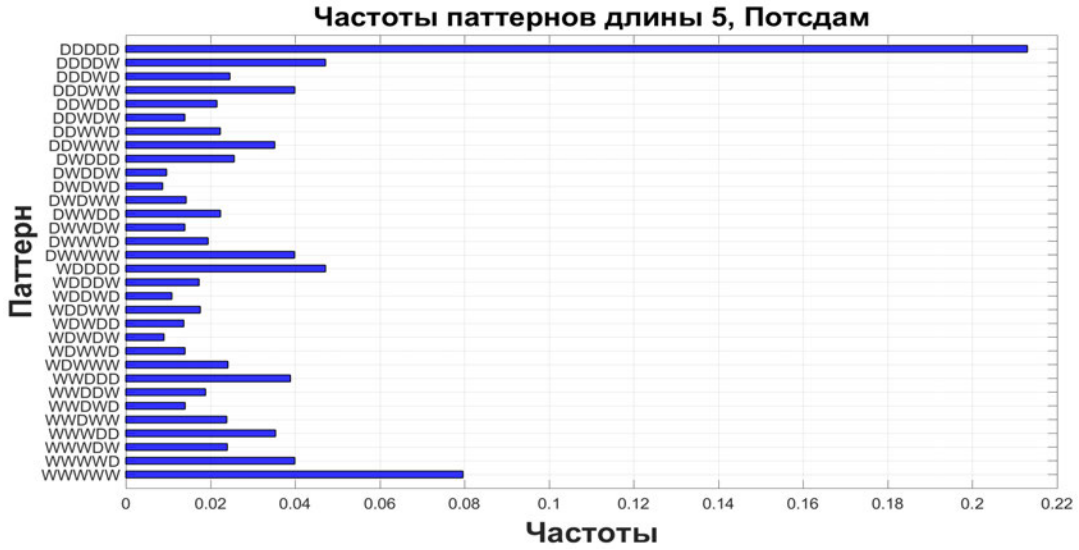

Рис. 6.1. Пример: частоты для пятидневных паттернов (Потсдам)

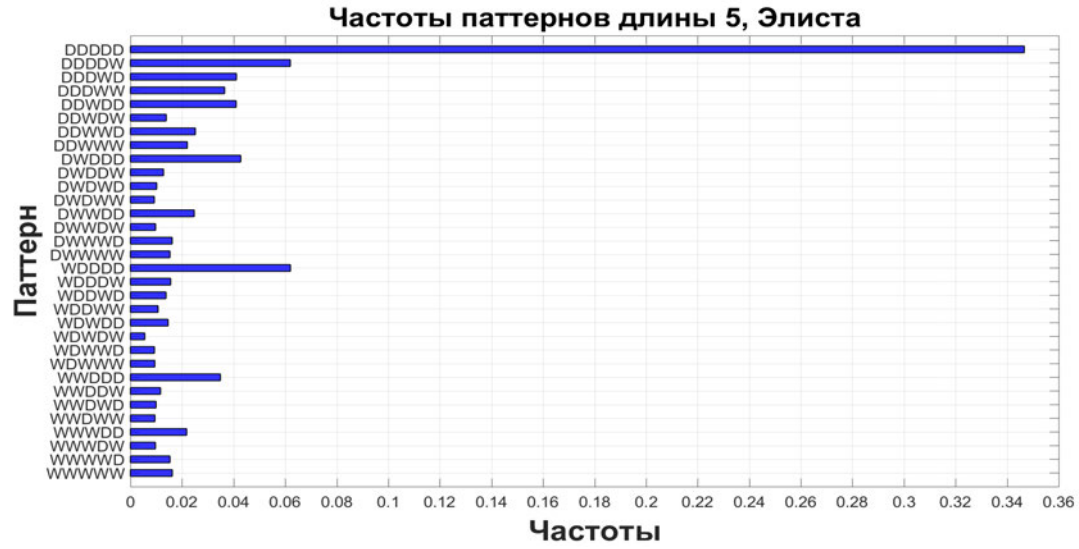

Рис. 6.2. Пример: частоты для пятидневных паттернов (Элиста)

Представленные таблицы и диаграммы позволяют делать определенные выводы о климатических зонах, в которых расположены соответствующие города. Так, в Потсдаме климат умеренный, продолжительные осадки не редкость (например, для трех подряд дней частота – 0,1789, см. таблицу 6.1), в то время как в Элисте климат резко континентальный с умеренным числом осадков (частота «трехдневного» дождя за период наблюдений составила всего 0,0631). Для четырнадцатидневных наборов максимальную частоту для обоих городов имеет последовательность из всех «сухих» дней, при этом для Элисты соответствующая величина равна 0,1138, а для Потсдама – 0,0671.

6.1.2 Проверка марковского свойства

В большинстве работ, посвященных статистическому анализу метеорологических данных, считается, что продолжительность периода выпадения осадков, измеренная в сутках (то есть число последовательных «дождливых» дней), подчиняется геометрическому распределению вероятностей [444]. Возможно, данные предположения базируются на классической интерпретации геометрического распределения в терминах испытаний Бернулли как распределения числа последовательных «дождливых» дней («успех») до первого дня без осадков («неудача»). Для изучаемых в работе городов проверим более слабое предположение, а именно наличие марковости.

Для этого потребуется вычисление условных вероятностей, но для различных значений N необходимые базовые величины для классического определения $\mathbb{P}(A|B) = \mathbb{P}(AB)/\mathbb{P}(B)$ для различных событий A и B являются известными. В таблице 6.2 представлены условные вероятности и модули их разностей для Потсдама и Элисты, демонстрирующие отсутствие свойства марковости у данных. Таким образом, последовательность «дождливых» и «сухих» дней не является марковской, поэтому использование схемы испытаний Бернулли некорректно. Развитие указанных вероятностных моделей будет предложено далее в разделе 6.3.

6.1.3 Вероятностное прогнозирование

Паттерны являются достаточно распространенным инструментом в рамках решения различных климатологических задач [124, 251, 356, 394]. С помощью введенной выше схемы дискретизации может быть реали-

Таблица 6.2. Таблица значений для условных вероятностей

Выражение	**Значение**	
	Потсдам	**Элиста**
$\mathbb{P}(\{DDD\} \mid \{DD\})$	0,7774	0,8264
$\mathbb{P}(\{DDW\} \mid \{DD\})$	0,2226	0,1736
$\mathbb{P}(\{DWD\} \mid \{DW\})$	0,3785	0,532
$\mathbb{P}(\{DWW\} \mid \{DW\})$	0,6215	0,468
$\mathbb{P}(\{WDD\} \mid \{WD\})$	0,6043	0,2761
$\mathbb{P}(\{WDW\} \mid \{WD\})$	0,3957	0,5114
$\mathbb{P}(\{WWD\} \mid \{WW\})$	0,3484	0,4887
$\mathbb{P}(\{WWW\} \mid \{WW\})$	0,6516	0,8264
$\lvert\mathbb{P}(\{DDD\} \mid \{DD\}) - \mathbb{P}(\{DD\} \mid \{D\})\rvert$	0,0466	0,0198
$\lvert\mathbb{P}(\{DDW\} \mid \{DD\}) - \mathbb{P}(\{DW\} \mid \{D\})\rvert$	0,0466	0,0198
$\lvert\mathbb{P}(\{DWD\} \mid \{DW\}) - \mathbb{P}(\{WD\} \mid \{W\})\rvert$	0,0192	0,0098
$\lvert\mathbb{P}(\{DWW\} \mid \{DW\}) - \mathbb{P}(\{WW\} \mid \{W\})\rvert$	0,0192	0,0098
$\lvert\mathbb{P}(\{WDD\} \mid \{DD\}) - \mathbb{P}(\{DD\} \mid \{D\})\rvert$	0,1265	0,827
$\lvert\mathbb{P}(\{WDW\} \mid \{WD\}) - \mathbb{P}(\{DW\} \mid \{D\})\rvert$	0,1265	0,827
$\lvert\mathbb{P}(\{WWD\} \mid WW) - \mathbb{P}(\{WD\} \mid \{W\})\rvert$	0,0107	0,0109
$\lvert\mathbb{P}(\{WWW\} \mid \{WW\}) - \mathbb{P}(\{WW\} \mid \{W\})\rvert$	0,0107	0,0109

зован базовый подход к прогнозированию наблюдений. А именно, по некоторой заданной части дискретизованного ряда с помощью вычисления соответствующих условных вероятностей можно определять вероятность появления после них определенных комбинаций. В отличие от стандартной для анализа данных практики, когда предсказываемое окно не должно превышать размер входных наблюдений, для исторических значений это правило может и нарушаться.

В качестве примера рассмотрим построение вероятностного прогноза на два следующих дня для Потсдама и Элисты при условии текущих наблюдений вида «`Wet-Wet-Dry-Dry`», то есть два дня подряд выпали осадки, в следующие двое суток они не регистрировались. В таблице 6.3 представлены вероятности соответствующих событий, при этом полужирным шрифтом выделено наиболее вероятное событие.

Таким образом, возможно формулировать утверждения вида: «Вероятность осадков *через* 2 дня в Потсдаме при текущих наблюдениях `Wet-Wet-Dry-Dry` составляет 0,3961, а вероятность отсутствия осадков *через* 2 дня – 0,6039»; «Вероятность осадков *через* 2 дня в Элисте при те-

кущих наблюдениях `Wet-Wet-Dry-Dry` составляет 0,2889, а вероятность отсутствия осадков *через* 2 дня – 0,7111». Здесь очевидным образом складываются значения второй и четвертой или первой и третьей строк из таблицы 6.3 соответственно.

Таблица 6.3. Пример: прогнозирование осадков на два следующих дня

Прогноз	**Вероятность**	
	Потсдам	**Элиста**
`Dry-Dry`	**0,4828**	**0,5852**
`Dry-Wet`	0,1909	0,1641
`Wet-Dry`	0,1211	0,1259
`Wet-Wet`	0,2053	0,1249

При этом получение новых данных может существенным образом изменить вероятности событий только при их значительном объеме, что позволяет говорить об устойчивости прогнозов. С вычислительной точки зрения обновление данных не является трудоемкой задачей. Можно отметить потенциальную возможность использования паттернов с точки зрения верификации ансамблей прогнозов – например, европейские климатические агентства достаточно точно предсказывают общий объем осадков, который выпадет за некоторый период, но остается актуальной задача определения структуры его распределения по дням.

Описанный анализ данных на основе паттернов может быть формализован в виде следующего алгоритма 6.1.

Алгоритм 6.1. Анализ метеорологических данных на основе паттернов

```
1:  function Patterns(Data, N)
2:      Data_Discr←Real2Discrete(Data, k); // Дискретизация данных
3:      ℙ←PatProb(Data_Discr, N);   // Частоты паттернов длины N
4:      Plot(ℙ, N);                 // Визуализация частот
5:      isMarkovChain(Data_Discr, N);   // Проверка на марковость
6:      repeat
7:          // Текущие наблюдения и длительность прогноза
8:          [Current, duration]←Input( );
9:          // Вероятностный прогноз
10:         Forecast←PrecipForecast(ℙ, Current, duration);
11:     until not(isEmpty(Current, duration))
12:     return ;
```

6.1.4 Бинарные нейросетевые прогнозы осадков

В данном разделе рассмотрим построение прогнозов для преобразованных наблюдений, однако вместо вероятностных подходов воспользуемся нейронными сетями. В качестве обучающих рядов задействуются исторические паттерны, однако в явном виде частота каждого из наборов не используется, а соответствующие процедуры реализуются в скрытых слоях нейронной сети. Таким образом, для каждого наблюдения, фактически, решается задача бинарной классификации. Для работы с нейросетями использована библиотека глубокого обучения `Keras`, фреймворк `TensorFlow` и язык программирования `Python`. Для повышения скорости обучения нейронных сетей расчеты производились на ресурсах гибридного высокопроизводительного вычислительного кластера архитектуры `Power9`. Графики для повышения наглядности подготовлены с помощью программного решения, разработанного на встроенном языке пакета `MATLAB`.

В качестве инструмента для решения задачи бинарной классификации входных наблюдений были выбраны нейронные сети прямого распространения с тремя скрытыми слоями. На вход такой нейросети подается ряд из N дискретных значений, описывающий чередование «сухих» и «дождливых» дней. С учетом наличия сезонных эффектов в данных, в качестве дополнительного входного атрибута используется параметр, содержащий номер месяца последнего дня в выборке. По результатам работы нейронная сеть относит полученные данные к одному из двух классов – временной промежуток, за которым следует день без осадков, или интервал, за которым следует дождливый день. В качестве функции активации нейронов скрытых слоев использована функция `ReLU`, для выходного слоя – традиционная для задач классификации функция `softmax` (см. раздел 5.3.1). В качестве метрики, описывающей качество обучения нейросети, использована среднеквадратическая ошибка `RMSE`. Прогноз строится на один или два дня вперед по 28 предшествующим значениям. В целом, используемая архитектура и соответствующие гиперпараметры достаточно близки к описанным в разделе 5.3.1.

На рисунках 6.3 и 6.4 на верхних графиках представлена зависимость изменения величины ошибки в процессе обучения для части выборки, используемой для настройки нейросети (80% от всех исходных данных), а на нижних – для тестового набора, в зависимости от эпохи обучения для одно- и двухдневного прогнозов для Потсдама.

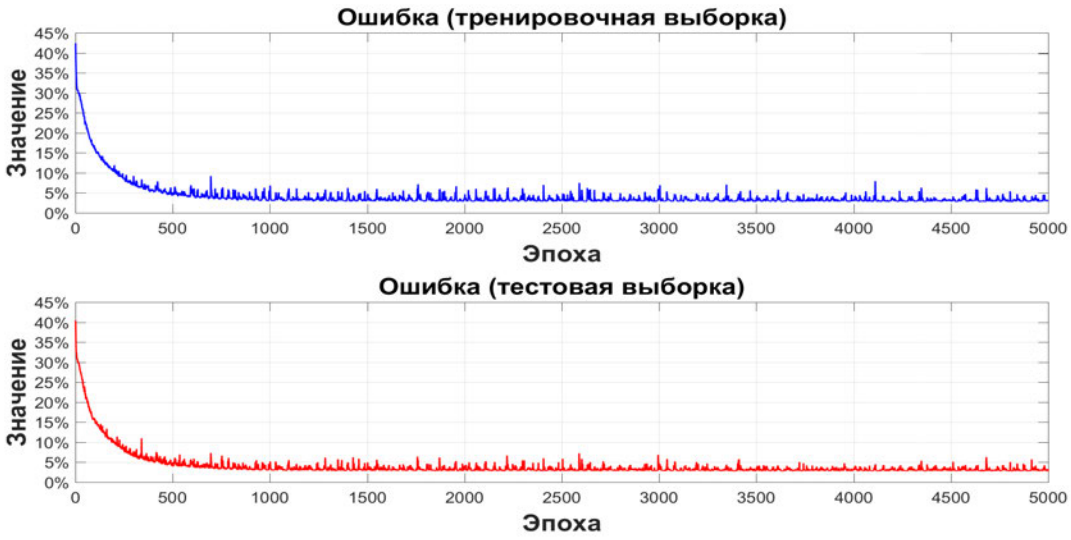

Рис. 6.3. Ошибки для однодневного прогноза, Потсдам

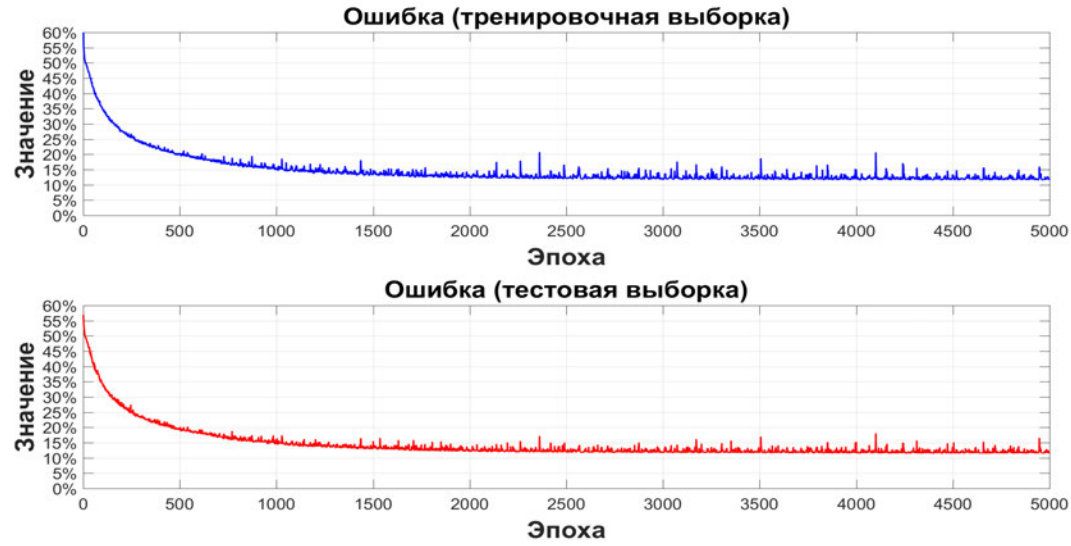

Рис. 6.4. Ошибки для двухдневного прогноза, Потсдам

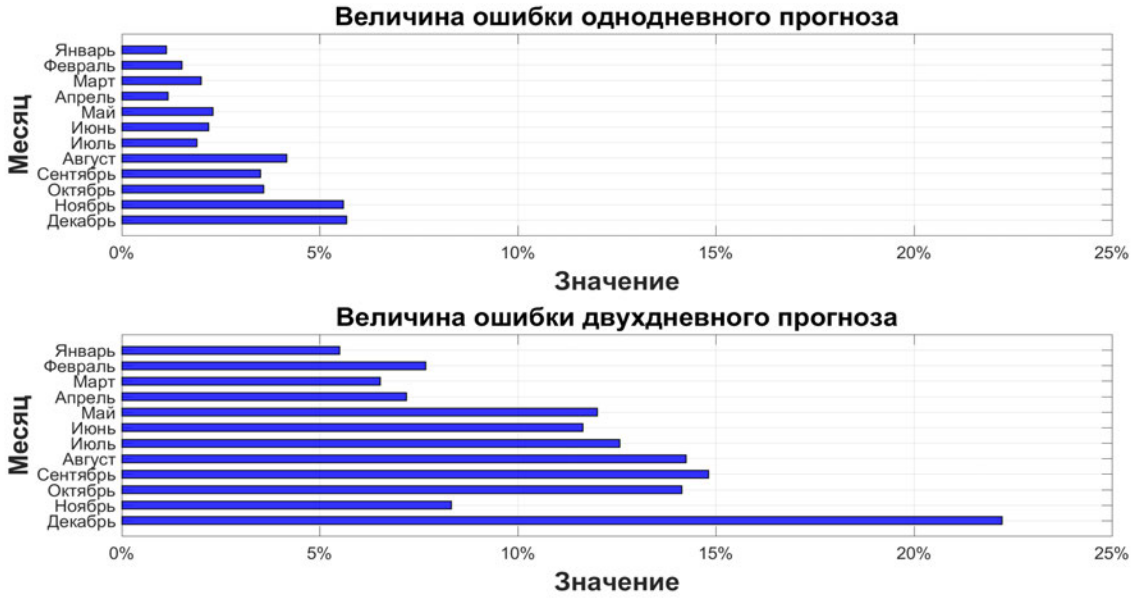

Рис. 6.5. Величины ошибок с учетом сезонного фактора, Потсдам

Кривые на графиках ведут себя в достаточной степени согласованно для каждой из длительностей прогнозов, что свидетельствует об отсутствии переобучения: величина ошибки получается примерно одинаковой как для обучающей, так и для тестовой части, которая непосредственно в процессе построения нейросети не участвует. Таким образом, можно ожидать, что модель построена корректно.

Для используемых тестовых наборов по итогам 5000 эпох обучения полученная точность для Потсдама составляет 97,1% для однодневного и 88,5% для двухдневного прогнозов. Аналогичные результаты получены и для Элисты – 96,9% и 90,1%. Необходимо отметить, что в наблюдениях для Элисты содержится большее количество нулей, что способствует некоторому повышению точности более длительных прогнозов.

Особый интерес представляет прогнозирование выпадения осадков с учетом присутствия в данных сезонного фактора. Было установлено, что расширение признакового пространства за счет включения дополнительного наблюдения, соответствующего номеру месяца заключительного элемента, в обучающую выборку позволяет повысить качество прогнозирования, особенно для случая двухдневного предсказания.

На рисунке 6.5 представлены столбчатые диаграммы, показывающие ошибки предсказания в процентах для каждого месяца для Потсдама. Точность прогнозирования для обоих городов приведена в таблице 6.4.

Таблица 6.4. Точность прогнозирования выпадения осадков

Месяц	**Точность**			
	Потсдам		**Элиста**	
	1 день	**2 дня**	**1 день**	**2 дня**
Январь	98,9%	94,5%	99,4%	93,6%
Февраль	98,5%	92,3%	100%	95%
Март	98%	93,5%	98,5%	93%
Апрель	98,8%	92,8%	97,7%	91,2%
Май	98%	88%	96,6%	89,7%
Июнь	97,8%	88,4%	95,7%	91,5%
Июль	98,1%	87,4%	95,8%	91,6%
Август	95,8%	85,8%	95,4%	90,9%
Сентябрь	96,5%	85,2%	96,6%	86,9%
Октябрь	96,4%	85,9%	97,6%	91,8%
Ноябрь	94,4%	91,7%	94%	91,8%
Декабрь	94,3%	77,8%	94,8%	82,7%

Очевидно снижение качества подобного прогнозирования при увеличении длительности соответствующего периода. Переход к двухдневным прогнозам влечет в среднем снижение точности на 8,52% и 6,03% для Потсдама и Элисты соответственно.

Для визуализации точности обучения нейронных сетей (см. рисунки 6.3–6.5) был создан специальный программный модуль на языке программирования пакета `MATLAB`, логика работы которого представлена в алгоритме 6.2.

Алгоритм 6.2. Пакетная обработка результатов обучения

```
1: function BatchNNVisualization(directory)
2:     // directory - имя каталога с результатами нейросетей
3:     while (true) do
4:         FName←dir(directory);          // Чтение очередного файла
5:         if (isEmpty(FName)) then
6:             break;
7:         PlotErr(FName);              // Ошибки обучения по эпохам
8:         PlotErrMonth(FName);  // Ошибки обучения по месяцам
9:     return ;
```

6.1.5 Решение задачи k-ичной классификации

Аналогично подходу, описанному в разделе 5.3.1, данная процедура была расширена путем перехода от двоичной модели дискретизации событий к k-ичной. Если в i-й день наблюдалась какая-либо положительная величина, то она заменяется на целое значение j из сегмента $[1, k-1]$ ($\widetilde{V}_{daily}^{(i)} = j$), соответствующих разбиению объема осадков на $k-2$ равных интервала; в противном случае величине $\widetilde{V}_{daily}^{(i)}$ приписывается нулевое значение. Одно- и двухдневные прогнозы строятся по 28 предыдущим наблюдениям, величина k равна 10. На рисунках 6.6 и 6.7 представлена зависимость изменения величины ошибки в процессе обучения для тренировочной и тестовой выборок для Потсдама. Для используемых тестовых данных по итогам 5000 эпох обучения полученная точность составляет 90,9% для однодневного и 73,1% для двухдневного прогнозов для Потсдама и 92,2% и 81,7% для Элисты.

Необходимо отметить, что по сравнению с задачей бинарной классификации в данном случае для стабилизации значения ошибки в процессе обучения предсказания требуется порядка 3000 эпох вместо 2000 ранее.

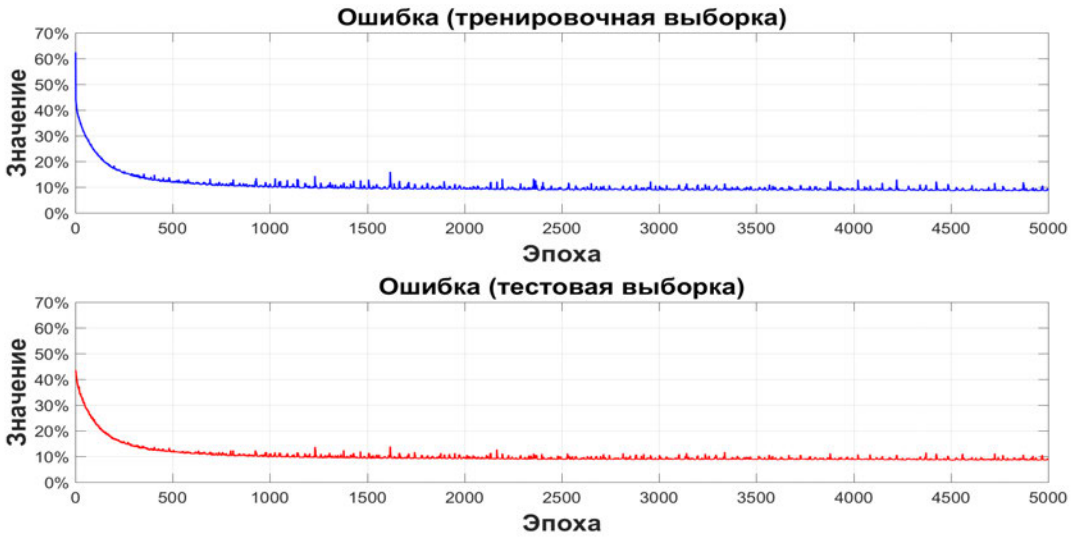

Рис. 6.6. Ошибки для однодневного прогноза, Потсдам

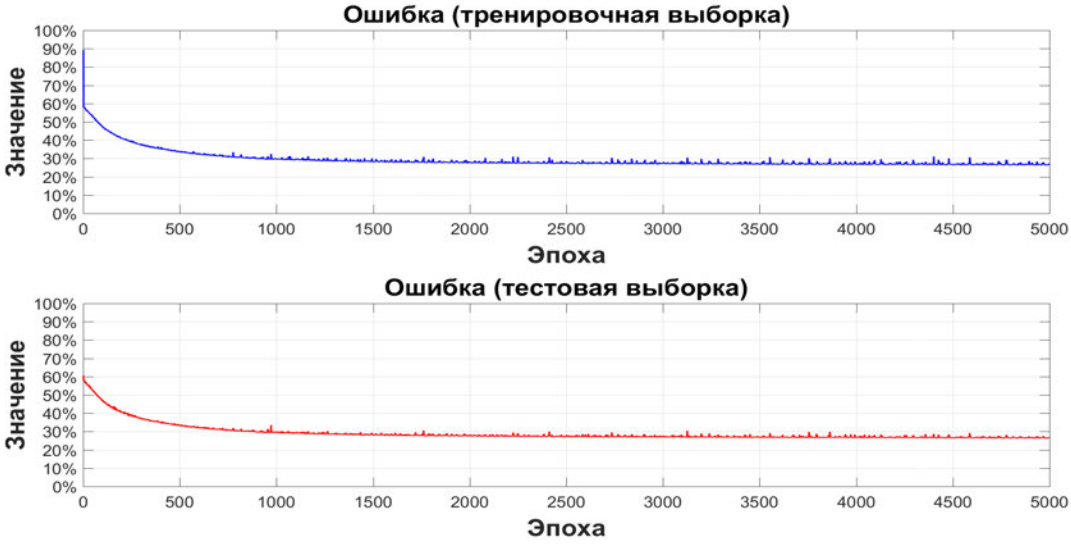

Рис. 6.7. Ошибки для двухдневного прогноза, Потсдам

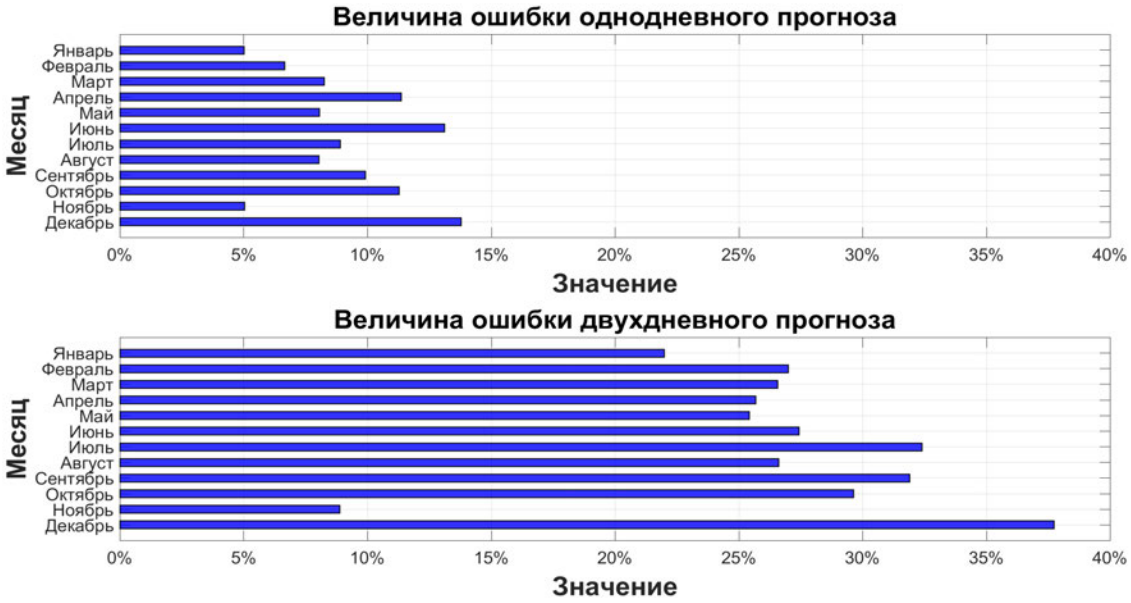

Рис. 6.8. Величины ошибок с учетом сезонного фактора, Потсдам

На рисунке 6.8 представлены соответствующие столбчатые диаграммы, показывающие величину ошибок предсказания в процентах для каждого месяца для Потсдама. Величины ошибок по месяцам для одно- и двухдневного прогнозов для обоих городов приведены в таблице 6.5.

Таблица 6.5. Точность k-ичного прогнозирования объемов осадков

Месяц	**Точность**			
	Потсдам		**Элиста**	
	1 день	**2 дня**	**1 день**	**2 дня**
Январь	95%	78%	92,2%	83,7%
Февраль	93,3%	73%	95,5%	86,5%
Март	91,7%	73,4%	94,7%	81,7%
Апрель	88,6%	74,3%	92,1%	81,2%
Май	92%	74,6%	91,1%	78,6%
Июнь	86,9%	72,6%	93,1%	83,3%
Июль	91,1%	67,6%	92,4%	82,6%
Август	92%	73,4%	92%	81,5%
Сентябрь	90,1%	68,1%	89,1%	80,1%
Октябрь	88,7%	70,4%	91,7%	79,4%
Ноябрь	95%	91,1%	94,6%	92%
Декабрь	86,2%	62,3%	89%	70,2%

По сравнению со случаем бинарных паттернов в случае прогноза на два дня величина ошибки возрастает весьма значительно, хотя средние значения точности для рядов (около 70–80%) являются все еще достаточно высокими.

6.1.6 Оптимизация конфигураций архитектур

Одним из наиболее популярных современных трендов обработки данных (в том числе и метеорологических) стало использование искусственных нейронных сетей. При этом на передний план, помимо выбора типа архитектуры, выходит задача корректного определения гиперпараметров [440]. Для каждой достаточно сложной задачи оптимальные настройки должны подбираться индивидуально, что представляет собой нетривиальную вычислительную проблему. Очевидно, что для реализации научных сервисов, например, в рамках цифровых платформ, желательно максимально автоматизировать исследовательский процесс, чтобы снизить зависимость результатов от квалификации пользователя в

обучении и конфигурации нейронных сетей. В данном разделе будет продемонстрирована возможность использования ряда известных методов для выбора оптимальных настроек гиперпараметров на примере анализа данных 22 европейских метеорологических станций (см. рисунок 6.9) за период 1904–1999 гг., расположенных в таких странах, как, например, Австрия, Германия, Голландия, Дания, Норвегия, Франция. В наблюдениях отсутствуют пропуски, поэтому нет необходимости их заполнения для корректного решения задач анализа и прогнозирования. Отметим, что вне зависимости от используемого метода предполагается, что при одинаковых значениях гиперпараметров должна получаться та же самая величина точности обучения, то есть имеет место воспроизводимость.

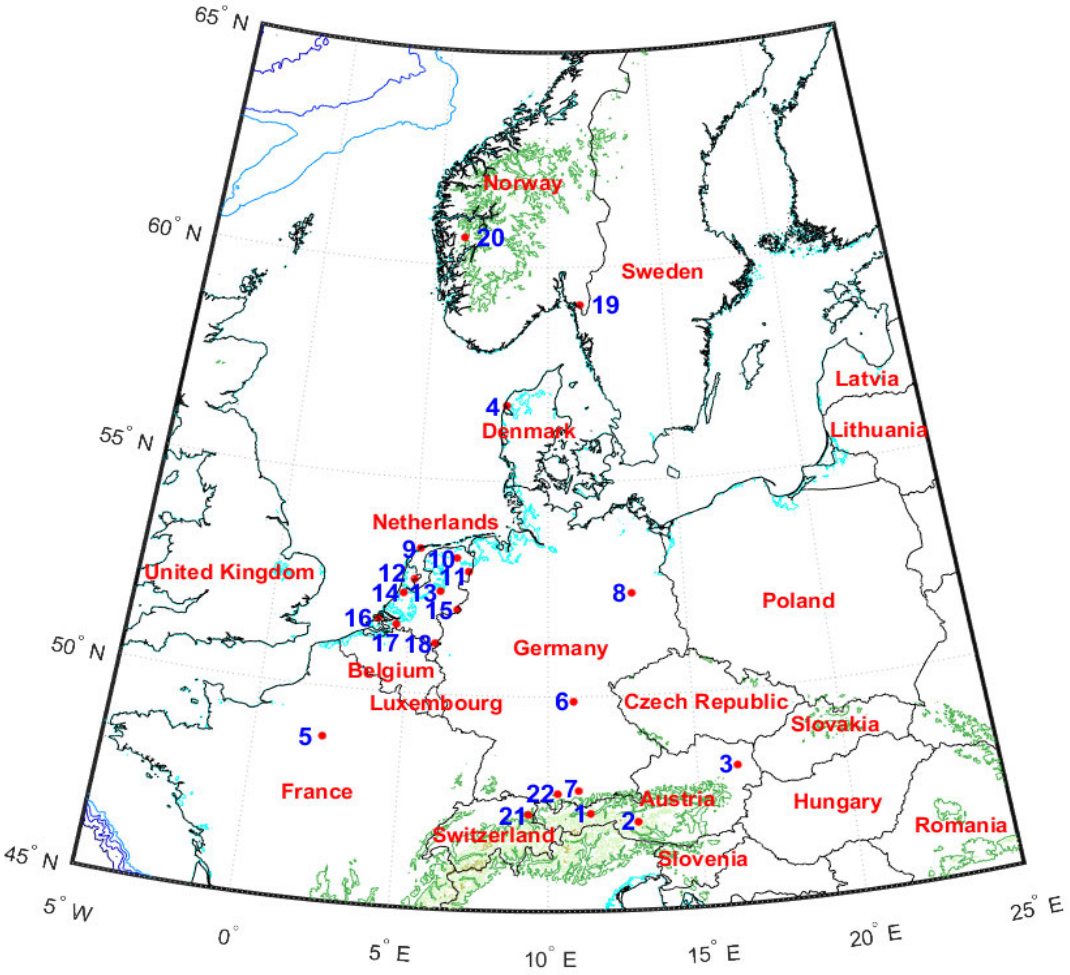

Рис. 6.9. Географическое расположение анализируемых станций

Наиболее простой и, очевидно, наименее эффективный, метод оптимизации гиперпараметров – ручной подбор соответствующих значений вектора. Другой популярный способ – поиск по решетке, в котором осуществляется полный перебор всех возможных комбинаций (подпространству) гиперпараметров. Однако в реальных задачах для корректного ре-

шения требуется значительное число гиперпараметров, а значит, размерность решетки быстро возрастает, что приводит к очень высокой вычислительной сложности данного метода. Эффективной альтернативой выступает метод случайного поиска: выбираются N случайных векторов в пространстве гиперпараметров, для каждой комбинации проводится обучение нейронной сети и определяется наилучшая с точки зрения полученной точности. Известно [152], что данный подход позволяет получить результаты как минимум не хуже, чем в случае поиска по решетке, при этом временные затраты снижаются значительным образом. Данный метод достаточно популярен при решении различных прикладных задач, в частности, в медицине [367, 410], кредитном скоринге [425], распознавании речи и почерка [255].

В таблицах 6.6 и 6.7 приведены примеры для метеостанций (см. карту на рисунке 6.9), расположенных на уровне моря, на равнине и в горах. Исследуются прогнозы на один и три дня вперед.

Таблица 6.6. Точность k-ичного прогнозирования на 1 шаг, $k = 10$

Страна (номер станции)	**Количество случайных выборов**			**Поиск по решетке**
	1 выбор	**5 выборов**	**10 выборов**	
Австрия (**3**)	83,49%	87,55%	87,62%	87,66%
Дания (**4**)	76,26%	80,99%	81,01%	81,05%
Франция (**5**)	79,96%	84,18%	84,22%	84,23%
Германия (**6**)	80,31%	83,52%	83,58%	83,60%
Германия (**8**)	82,42%	85,73%	85,75%	85,75%
Голландия (**9**)	76,49%	80,16%	80,21%	80,23%
Норвегия (**19**)	78,48%	82,51%	82,58%	82,60%

Таблица 6.7. Точность k-ичного прогнозирования на 3 шага, $k = 10$

Страна (номер станции)	**Количество случайных выборов**			**Поиск по решетке**
	1 выбор	**5 выборов**	**10 выборов**	
Австрия (**3**)	70,34%	71,10%	71,12%	71,14%
Дания (**4**)	57,13%	59,31%	59,36%	59,37%
Франция (**5**)	63,45%	63,80%	63,80%	63,80%
Германия (**6**)	63,04%	63,38%	63,39%	63,39%
Германия (**8**)	63,90%	66,14%	66,14%	66,14%
Голландия (**9**)	56,11%	58,74%	58,84%	58,85%
Норвегия (**19**)	59,74%	61,74%	61,74%	61,75%

В столбцах продемонстрирована средняя точность соответствующего числа случайных выборов, полученная для серии из пятидесяти независимых последовательных запусков. Из приведенных таблиц следует, что средняя точность случайного поиска уже для 5–10 выборов вполне сопоставима с полным перебором, при этом скорость может быть выше в 10–300 раз в зависимости от количества узлов решетки. Таким образом, для метеорологических данных и выбранной архитектуры подтверждается эффективность метода случайного поиска.

Задача обучения архитектур с различными комбинациями гиперпараметров (число скрытых слоев и нейронов в них, коэффициенты дропаута во входном и скрытых слоях, коэффициент изменения скорости обучения, методы оптимизации) является весьма трудоемкой, поэтому были использованы ресурсы гибридного высокопроизводительного вычислительного кластера архитектуры `Power9`. Его использование для сравнения конфигураций архитектур для решения задачи k-ичной классификации ($k = 10$) с входным вектором из 180 наблюдений позволило ускорить обучение не менее, чем в 5–8 раз. При этом при поиске по решетке используется более 3900 различных комбинаций элементов.

6.2 Анализ методов восстановления пропущенных значений в пространственно-временных метеорологических данных

Повышение эффективности алгоритмов машинного обучения привело к росту их востребованности как в задачах анализа результатов физических моделей предсказания погоды с целью получения более точного прогноза, так и в качестве самостоятельных инструментов исследования пространственно-временных метеорологических рядов, полученных со спутников и метеостанций. Такие наблюдения в больших объемах поступают с огромного числа датчиков и зачастую содержат пропуски, которые могут существенным образом повлиять на качество обучения методов или изменить решения статистических моделей анализа различных метеорологических явлений, например, экстремальных осадков, которые будут изучаться в следующих разделах данной главы. Поэтому весьма важной оказывается задача корректного заполнения пропусков в подобных данных [372].

Наиболее простой способ заключается в исключении из обработки ча-

сти выборки с пропусками до момента, когда данные станут полными, однако в метеорологических рядах пропущенное значение может появиться в любой случайный момент времени (в частности, из-за особенностей их регистрации). Другой подход основан на использовании данных реанализа, однако возможны существенные расхождения с реальными наблюдениями. Поэтому развиваются различные статистические и нейросетевые методы для обеспечения возможности корректного заполнения пропусков для последующего использования, например, в гидрологических моделях [144, 280, 402]. В том числе, в качестве дополнительных данных используются значения соседних станций [378], однако это возможно далеко не всегда. Таким образом, наибольший интерес представляют методы, использующие временные ряды только конкретной метеостанции, а дополнительные признаки для обучения извлекаются либо непосредственно из этих наблюдений, либо из иных метеорологических параметров, полученных в той же географической точке.

В данном разделе на примере пространственно-временных рядов, собранных более чем на 100 станциях России, Европы, Америки, Азии и Африки, то есть без привязки к каким-то конкретным регионам (в отличии от общепринятого подхода в метеорологии, когда подробно анализируются данные какой-либо страны [200, 258, 427]), проведен анализ эффективности методов машинного обучения для заполнения пропущенных значений в метеорологических данных. В качестве источника тестовых наблюдений использовалась открытая база `NNDC Climate Data` Национального управления океанических и атмосферных исследований (`NOAA`). Основной акцент был сделан на выборе наиболее универсальных методов, которые оставались бы эффективными при анализе данных со станций из регионов, отличающихся своим географическим местоположением от тестовых. Также это позволяет рассчитывать на сохранение эффективности развитых методов при анализе иных типов данных. Методы машинного обучения, а не нейронные сети, были выбраны по причине более высокой скорости их обучения.

6.2.1 Подготовка данных и используемые метрики точности

Для выбора универсальных методов заполнения пропущенных наблюдений использовались полные пространственно-временные ряды (или их части, не содержащие пропусков). Поэтому сначала было реали-

зовано случайное внедрение от одного до трех подряд идущих пропусков на одно окно [248], при этом их общее количество варьировалось вплоть до максимально возможного для выбранных настроек. Затем с помощью различных методов (исторические паттерны, алгоритмы машинного обучения) все эти пропуски заполнялись и проводилось сравнение с истинными значениями. При этом решались задачи и классификации, и регрессии. Предсказываемыми переменными являются факт выпадения и объемы осадков, а точка росы, средние температура и скорость ветра используются в ряде случаев в качестве дополнительных признаков.

Для оценивания корректности работы использовались несколько метрик. В частности, для задач классификации используется величина `ACC`, определяемая выражением

$$ACC = TPR + TNR, \tag{6.1}$$

где `TPR` (`True Positive Rate`) – доля верно угаданных случаев (классов) наличия осадков, а `TNR` (`True Negative Rate`) – доля верно угаданных случаев отсутствия осадков. Также для сравнения бинарных классификаторов используется площадь под `ROC`-кривой (`ROC AUC`) [270], которая описывает точность решения модели как вероятность принадлежности к определенному классу и задается выражением

$$ROC\,AUC = \int_0^1 TPR\left(FPR^{-1}(x)\right)\,dx, \tag{6.2}$$

где `FPR` (`False Positive Rate`) – доля неверно угаданных положительных классов. Она позволяет корректнее оценивать точность в случае, когда один из классов является доминирующим, что вполне соответствует исходным данным об осадках, в которых достаточно много нулевых значений. Кроме того, используется и стандартная для непрерывных данных метрика `RMSE`.

6.2.2 Заполнение пропусков на основе бинарной классификации

Сначала воспользуемся вероятностным подходом на основе паттернов. В данном разделе будут использованы паттерны длины $N = 5$, с помощью которых и задается минимальное расстояние между пропусками (в рамках паттерна их не может быть произвольное число). Рас-

сматриваются от одного до трех подряд пропущенных значений на окно, тогда их общее число в данных варьируется следующим образом:

– до 20% для единственного пропуска;
– до 33% для двух пропусков;
– до 43% для трех.

Опишем частотно-вероятностный подход к решению данной задачи.

1. Рассмотрим часть данных (после проведенной бинарной дискретизации) с пропуском, который обозначим символом $\mathcal{X}$:

$$\ldots - D - W - D - D - D - \boxed{\mathcal{X}} - W - D - D - W - \ldots$$

2. Рассмотрим все подвыборки длины $N = 5$, содержащие это пропущенное значение:

(a) $\ldots - D - \boxed{W - D - D - D - \mathcal{X}} - W - D - D - W - \ldots$

(b) $\ldots - D - W - \boxed{D - D - D - \mathcal{X} - W} - D - D - W - \ldots$

(c) $\ldots - D - W - D - \boxed{D - D - \mathcal{X} - W - D} - D - W - \ldots$

(d) $\ldots - D - W - D - D - \boxed{D - \mathcal{X} - W - D - D} - W - \ldots$

(e) $\ldots - D - W - D - D - D - \boxed{\mathcal{X} - W - D - D - W} - \ldots$

3. Пропущенное значение может быть только «D» или «W» (то есть решается задача бинарной классификации в смысле определения, были осадки в данный день или нет. Например, подвыборки из пункта 2a могут быть только такие:

$$W - D - D - D - \boxed{W} \quad \text{или} \quad W - D - D - D - \boxed{D}$$

4. Теперь необходимо выбрать подходящий паттерн с максимальной частотой (вероятностью) – и в качестве решения нужно использовать соответствующее значение из него:

$$\ldots - D - \boxed{W - D - D - D - \mathcal{X}} - W - D - D - W - \ldots \Rightarrow$$
$$\Rightarrow W - D - D - D - \boxed{D}$$
$$\ldots - D - W - \boxed{D - D - D - \mathcal{X} - W} - D - D - W - \ldots \Rightarrow$$
$$\Rightarrow D - D - D - \boxed{W} - W$$
$$\ldots - D - W - D - \boxed{D - D - \mathcal{X} - W - D} - D - W - \ldots \Rightarrow$$
$$\Rightarrow D - D - \boxed{D} - W - D$$

$$\ldots - D - W - D - D - \boxed{D - \mathcal{X} - W - D - D} - W - \ldots \Rightarrow$$
$$\Rightarrow D - \boxed{D} - W - D - D$$
$$\ldots - D - W - D - D - D - \boxed{\mathcal{X} - W - D - D - W} - \ldots \Rightarrow$$
$$\Rightarrow \boxed{W} - W - D - D - W$$

5. Тогда в качестве окончательного решения метода выбирается элемент, который чаще всего встречался в подвыборках, полученных на предыдущем шаге:

$$\ldots - D - W - D - D - D - \boxed{D} - W - D - D - W - \ldots$$

На рисунке 6.10 продемонстрирован пример применения данного подхода для различных долей пропущенных значений для Потсдама. Описанный алгоритм очень прост для реализации, однако полученная точность решения задачи бинарной классификации, особенно для случая нескольких подряд идущих пропусков, является достаточно умеренной или даже низкой.

Применим с той же целью метод опорных векторов (SVM) [184] с расширением признакового пространства за счет добавления данных об иных метеорологических характеристиках (точка росы, средние температура и скорость ветра и т.п.). Результаты для Потсдама представлены на рисунке 6.11. С увеличением общего количества пропущенных значений, точность уменьшается и для SVM, однако даже при доле исключеных значений в 40% от объема выборки при данная величина остается более 70% против 44,7% для паттернов. Кроме того, разница в точности при увеличении доли пропущенных значений (то есть между крайнелевой и крайне правой точками для каждого графика) не превышает 5,5%, что также является весьма хорошим результатом с точки зрения обработки реальных пространственно-временых рядов. Также, точность SVM-классификации достаточно близка друг к другу и для различного числа последовательных пропусков (не более 5,3%). Очевидно, что в данном случае усложнение используемого метода решения задачи является полностью оправданным.

Сравнение средней точности методов на основе паттернов и SVM представлена в таблице 6.8. В случае одного последовательного пропущенного значения разница составляет 7,3%, для двух и трех – 8,33% и 25,35%, соответственно. Таким образом, при увеличении количества пропусков точность вероятностного подхода существенно уменьшается, в то время как SVM менее чувствителен к этой ситуации.

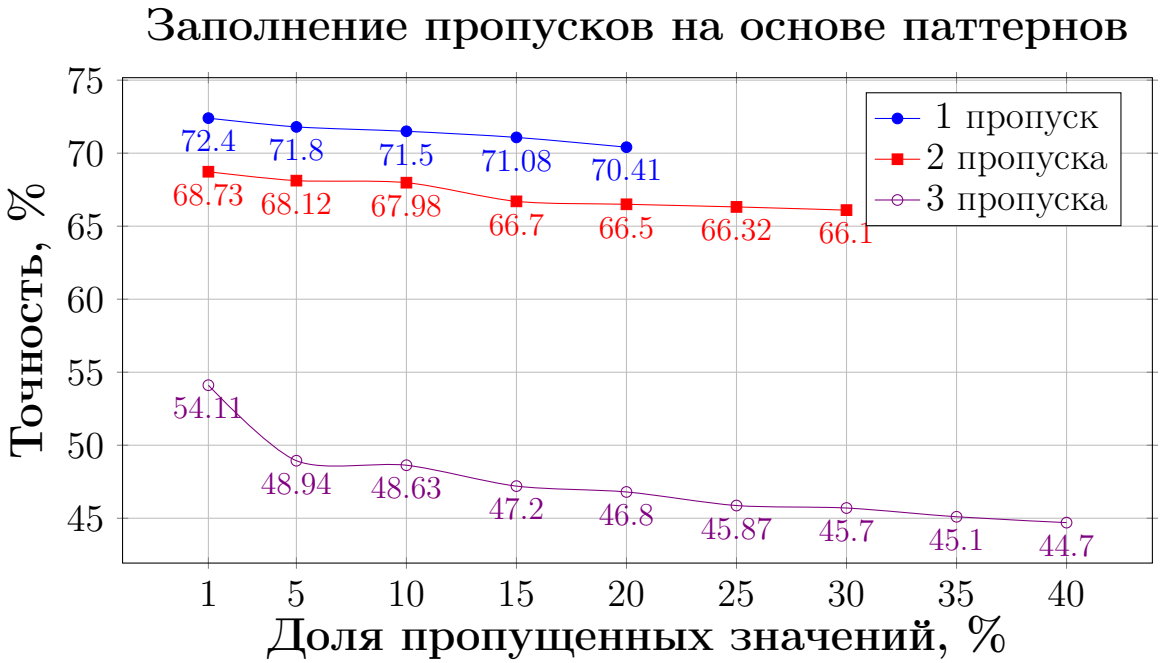

Рис. 6.10. Заполнение пропусков на основе паттернов, Потсдам

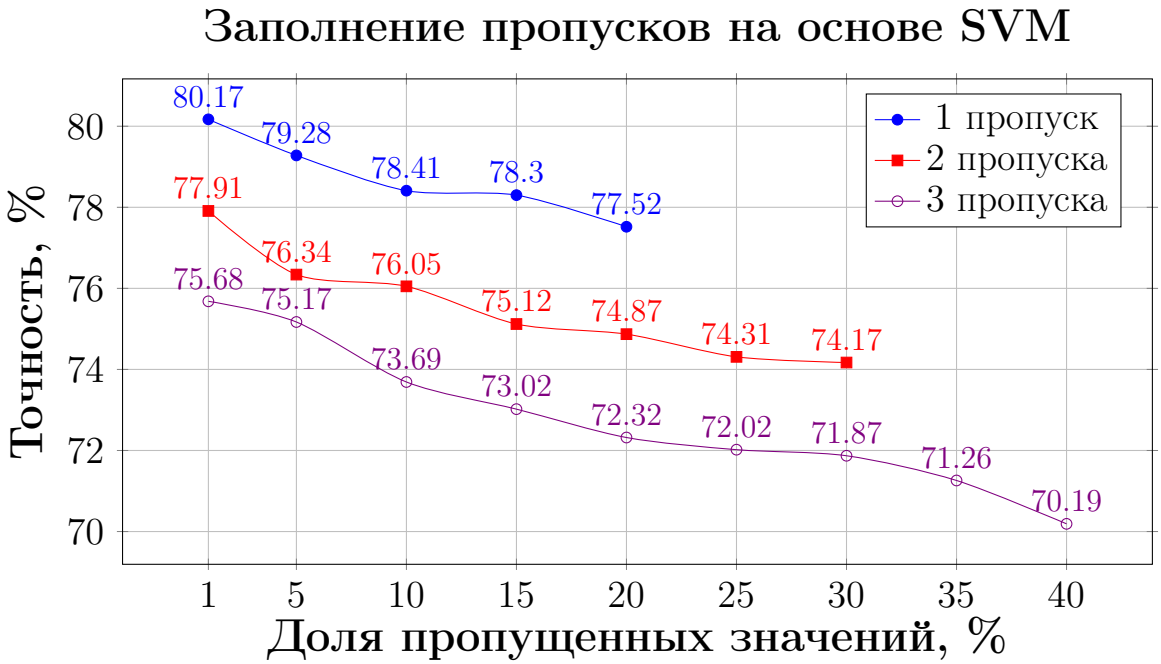

Рис. 6.11. Заполнение пропусков с использованием SVM, Потсдам

Таблица 6.8. Средняя точность заполнения пропусков с помощью паттернов и метода опорных векторов (классификация)

Количество последовательных пропусков	**Паттерны (Потсдам)**	**SVM (Потсдам)**
Один	71.44%	78.74%
Два	67.21%	75.54%
Три	47.45%	72.8%

Рассмотрим решение задачи классификации для случая одного подряд идущего пропуска с помощью экстремального градиентного `XGBoost_Logistic` [180] и категориального `CatBoost_Logistic` [357] бустинга с целевой функцией логистической регрессии для 14 тестовых станций в Германии (данные получены из базы `NOAA`). Категориальный бустинг – это предложенная компанией `Яндекс` модификация классического градиентного бустинга (`GB`) над деревьями решений [206], активное применяемая в коммерческих продуктах, но достаточно редко – в научных исследованиях (можно упомянуть лишь отдельные статьи [274, 359]). Метод `XGBoost` является наиболее часто применяемой для решения задач классификации и регрессии в значительном спектре прикладных областей [129, 175, 339, 407, 425] модификацией градиентного бустинга. Выбор функции логистической регрессии основан на возможности использования меньшего объема данных для обучения в этом случае.

В таблице 6.9 приведены результаты для лучшей из рассмотренных конфигураций по каждой станции в метриках `ACC` (6.1) и `ROC AUC` (6.2). Отметим, что всего для рассмотренных станций были протестированы более 8500 различных конфигураций, включая и задачу регрессии, которая будет рассмотрена в следующем разделе. Для извлечения признаков была использована библиотека `tsfresh` [183] для языка программирования `Python` с пакетами `NumPy`, `pandas`, `SciPy` и `scikit-learn`.

Таблица 6.9. Сравнение методов классификации на основе бустинга

Город	**Станция**	**Лучший метод**	**ACC**	**ROC AUC**
Берлин	93850	`XGBoost_Logistic`	86,11%	95,64%
Берлин	103810	`CatBoost_Logistic`	80,56%	82,19%
Доберлуг	94900	`CatBoost_Logistic`	83,33%	88,1%
Доберлуг	104900	`CatBoost_Logistic`	86,11%	84,23%
Хольцдорф	104760	`XGBoost_Logistic`	61,11%	72,1%
Линденберг	93930	`CatBoost_Logistic`	75%	76,92%
Линденберг	103930	`CatBoost_Logistic`	86,11%	86,55%
Нойруппин	92700	`XGBoost_Logistic`	72,22%	75,6%
Нойруппин	102700	`CatBoost_Logistic`	69,44%	66,9%
Потсдам	93790	`XGBoost_Logistic`	63,89%	68,75%
Потсдам	103790	`XGBoost_Logistic`	66,7%	78,41%
Визенбург	103680	`XGBoost_Logistic`	66,7%	70,63%
Виттенберг	94740	`XGBoost_Logistic`	77,78%	74,6%
Виттенберг	104740	`XGBoost_Logistic`	69,44%	73,96%

В таблице 6.10 приведены усредненные сразу по всем вариантам параметров значения точности для каждой модели.

Таблица 6.10. Усредненные результаты предсказания методов классификации

Метод	Метрика	
	ACC	ROC AUC
XGBoost_Logistic	72,22%	78,26%
CatBoost_Logistic	71,03%	75,96%

В данном случае метод XGBoost_Logistic продемонстрировал в среднем несколько более высокие (на 1,19 и 2,3% в метриках ACC и ROC AUC, соответственно) показатели точности по сравнению с CatBoost_Logistic. Однако, как видно из таблицы 6.9, для 6 из 14 станций в качестве лучшего алгоритма был выбран категориальный бустинг.

6.2.3 Заполнение пропущенных значений на основе классификации и регрессии

В этом разделе рассмотрим вопрос восстановления значений у пропущенных данных. При этом существенным образом будут использоваться сведения о бинарной классификации, изученные выше, с целью установления «сухих» и «дождливых» периодов. Действительно, в случае высокой точности работы методов классификации пропуски в «сухих» периодах должны заполняться строго нулевыми значениями, чтобы не увеличивать ошибку RMSE. Для определения величины пропусков в «дождливые» периоды будут использоваться следующие алгоритмы:

– заполнение средними значениями (Means): все отсутствующие элементы заменяются средним арифметическим выбранной подвыборки или полного временного ряда;
– метод k ближайших соседей (k-NN) [132];
– EM-алгоритм (EM);
– метод опорных векторов SVM;
– случайные леса (RFs) [135, 160, 201, 400];
– методы градиентного бустинга GB, XGBoost и CatBoost.

Сначала для оценивания точности заполнения пропусков в «дождливые» периоды воспользуемся нормализованной метрикой RMSE. А именно, рассмотрим величины $\varepsilon_m = {V_m}^{-1} RMSE_m$, где $RMSE_m$ среднеквадратическая ошибка для m-ого «дождливого» периода, а V_m общий объем

осадков за тот же период. При этом значения V_m определяются с использованием полных данных. Фактически проводится нормализация наблюдений за каждый «дождливый» период. Это позволяет сравнивать ошибки ε_m для разных «дождливых» периодов, а также вычислять их среднее значение для всех интервалов с пропусками и выражать точность в относительных единицах. Вектор-строка $\boldsymbol{\varepsilon} = \{\varepsilon_m\}_{m=\overline{1,L}}$ соответствует последовательности ошибок для всех «дождливых» периодов, содержащих пропущенные значения. Тогда общая ошибка равна $Err = L^{-1}\boldsymbol{\varepsilon}\mathbf{1}_{L\times 1}$, где $\mathbf{1}_{L\times 1}$ – вектор-столбец из L единиц.

Таблицы 6.11 и 6.12 демонстрируют сравнение точностей различных алгоритмов машинного обучения для заполнения искусственных пропусков в данных на примере Потсдама и Элисты для различных долей пропущенных значений для одного (см. также рисунок 6.12) и двух подряд идущих пропусков. В качестве классификатора в таблицах используется `SVM`. На рисунке 6.12 также приведено сравнение с историческими паттернами, для которых лучшие результаты были получены в комбинации с EM-алгоритмом.

Таблица 6.11. Точность гибридных методов, 1 пропуск (Потсдам)

Доля пропусков	**Метод**			
	SVM+XGBoost	**SVM+EM**	**SVM+RF**	**SVM+k-NN**
1%	89.74%	84.27%	85.79%	86.20%
5%	88.94%	81.94%	84.49%	84.76%
10%	87.52%	81.01%	81.70%	79.33%
15%	84.51%	80.87%	79.83%	77.39%
20%	82.89%	78.81%	76.79%	75.06%

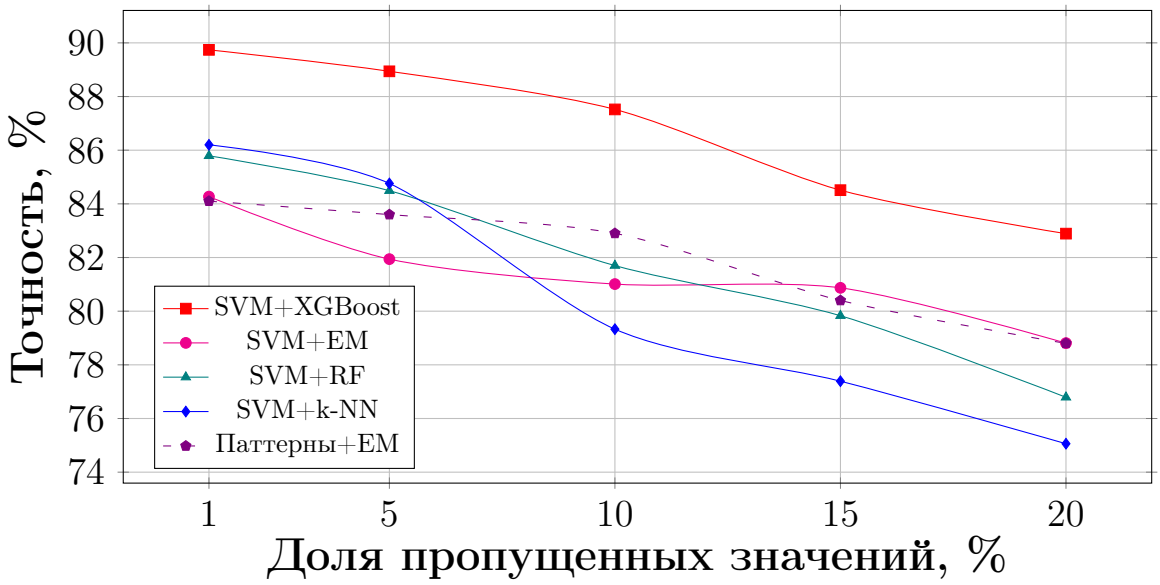

Рис. 6.12. Точность гибридных методов, 1 пропуск (Потсдам)

Таблица 6.12. Точность гибридных методов, 2 пропуска (Потсдам)

Доля пропусков	**Метод**			
	SVM+XGBoost	**SVM+EM**	**SVM+RF**	**SVM+k-NN**
1%	82.63%	79.51%	71.96%	83.75%
5%	76.37%	75.15%	71.56%	74.87%
10%	75.92%	73.32%	72.04%	70.11%
15%	74.27%	72.67%	68.22%	67.14%
20%	73.74%	71.08%	66.22%	67.95%
25%	73.19%	69.88%	68.33%	70.59%
30%	71.39%	68.79%	65.99%	66.53%

Таким образом, лучшие результаты получаются в случае использования метода опорных векторов для классификации и экстремального градиентного бустинга для регрессии. В терминах нормализованной ошибки `RMSE` для комбинации `SVM+XGBoost` точность даже для высоких уровней не опускается ниже 70–80%, что является очень хорошим результатом. Остальные методы показывают несколько более низкие значения, однако достаточно успешно себя проявляют случайные леса. Отметим, что соединительные линии на рисунке 6.12 (как и на остальных графиках этого раздела) носят иллюстративный характер – кривые строятся по точкам, отмеченным маркерами, в то время как промежуточные уровни пропусков не рассматриваются.

В таблице 6.13 приведено время, затраченное на обучение каждого метода (в секундах) на обычном настольном решении. Очевидно, что использование высокопроизводительных вычислительных кластеров в данном случае не требуется.

Таблица 6.13. Среднее время обучения алгоритмов машинного обучения

Доля пропусков	**Метод**			
	SVM+XGBoost	**SVM+EM**	**SVM+RF**	**SVM+k-NN**
1%	2.03	1.71	1.98	1.82
5%	1.99	1.93	1.88	1.74
10%	1.89	2.01	1.76	1.60
15%	1.64	2.55	1.54	1.39
20%	1.53	2.89	1.29	1.13
25%	1.18	3.05	0.98	0.84
30%	1.00	3.13	1.01	0.77

Для рассмотренных ранее 14 метеостанций в Германии воспользуемся методами `XGBoost` и `CatBoost` с различными целевыми функциями для восстановления пропусков с помощью «чистой» регресии (то есть

без предварительного анализа с помощью бинарного классификатора). В `XGBoost_RMSE` и `CatBoost_RMSE` используется `RMSE`, `CatBoost_MAE` – средняя абсолютная ошибка (`MAE`), а в `XGBoost_Gamma` – гамма-регрессия с логарифмической связью. В таблице 6.14 приведены результаты лучшей модели для каждой станции в метриках `ACC` и `RMSE`.

Таблица 6.14. Сравнение методов регрессии на основе бустинга

Город	**Станция**	**Лучший метод**	**ACC**	**RMSE**
Берлин	93850	`XGBoost_Gamma`	83,33%	0,0631
Берлин	103810	`XGBoost_Gamma`	69,44%	0,1111
Доберлуг	94900	`XGBoost_Gamma`	86,11%	0,075
Доберлуг	104900	`CatBoost_MAE`	83,33%	0,0876
Хольцдорф	104760	`XGBoost_Gamma`	75%	0,1027
Линденберг	93930	`XGBoost_Gamma`	75%	0,0721
Линденберг	103930	`XGBoost_Gamma`	80,56%	0,0912
Нойруппин	92700	`XGBoost_Gamma`	66,7%	0,061
Нойруппин	102700	`CatBoost_MAE`	66,7%	0,06429
Потсдам	93790	`XGBoost_Gamma`	66,7%	0,09095
Потсдам	103790	`XGBoost_Gamma`	66,7%	0,0886
Визенбург	103680	`XGBoost_Gamma`	61,11%	0,1767
Виттенберг	94740	`CatBoost_MAE`	72,22%	0,0591
Виттенберг	104740	`XGBoost_RMSE`	77,78%	0,0734

Сравнение результатов точности классификации для модели регрессии по сравнению с полученными ранее (см. таблицы 6.9 и 6.14, столбцы `ACC`) заметно хуже в большинстве справляются с задачей классификации. Рассмотрим ряд гибридных моделей на основе будстинга, в которых сначала будет решаться задача классификации, и только затем регрессии:

– `XGBoost_Logistic+Mean`: Сочетание выхода классификатора $\{c_t^{cls}\}$ и среднего значения объемов осадков $m = |T|^{-1} \sum_{t \in T} y_t$. Тогда выход гибридной модели определяется как $\{y_t^{pred} = c_t^{cls} \cdot m\}$.

– `XGBoost_Logistic+RMSE`: Вместо среднего значения m используется выход простого регрессора $\{y_t^{reg}\}$. Тогда выход новой модели определим как $\{y_t^{pred} = c_t^{cls} \cdot y_t^{reg}\}$.

– `XGBoost_Logistic+RMSE+Sigmoid`: Выход классификатора $\{p_t^{cls}\}$ определяет вероятность принадлежности к классу, соответствующему выпадению осадков. Зададим функцию связи $s(x) = \left(1 + e^{-\alpha(x-\beta)}\right)^{-1}$, где α и β – некоторые заданные действительные коэффициенты, кото-

рая позволяет использовать вместо бинарного решения классификатора $\{c_t^{cls}\}$ набор непрерывных значений, а кроме того, предоставляет возможность гибкого подбора коэффициентов, наиболее подходящих для конкретных данных. Определим выход как $\{y_t^{pred} = s\left(p_t^{cls}\right) \cdot y_t^{reg}\}$. Для рассматриваемых тестовых данных наилучшие результаты были получены для конфигураций с $\alpha = 10$ и $\beta = 0{,}45$.

В таблице 6.15 приведены 8 из 14 метеостанций, для которых были улучшены результаты чисто регрессионных моделей (см. таблицу 6.14). При этом увеличение точности классификации составило от 2% до 8% в зависимости от местоположения. Таким образом, использование гибридных моделей оказывается вполне оправданным с точки зрения более точной обработки данных и корректного заполнения пропусков в них.

Таблица 6.15. Точность для гибридных моделей на основе бустинга

Город	**Станция**	**Лучший метод**	**ACC**	**RMSE**
Берлин	93850	`XGBoost_Logistic+RMSE+Sigmoid`	91,67%	0,0588
Берлин	103810	`XGBoost_Logistic+RMSE`	75%	0,1012
Нойруппин	92700	`XGBoost_Logistic+RMSE`	72,22%	0,0556
Нойруппин	102700	`XGBoost_Logistic+RMSE+Sigmoid`	69,44%	0,05933
Потсдам	93790	`XGBoost_Logistic+RMSE+Sigmoid`	69,44%	0,0697
Потсдам	103790	`XGBoost_Logistic+RMSE+Sigmoid`	69,44%	0,0777
Визенбург	103680	`XGBoost_Logistic+RMSE`	66,7%	0,1418
Виттенберг	94740	`XGBoost_Logistic+RMSE+Sigmoid`	80,56%	0,052

В таблице 6.16 приведены усредненные сразу по всем вариантам параметров значения точности как для чисто регрессионных, так и для гибридных моделей в порядке возрастания величины `ACC` (6.1). Также для сравнения представлена простая модель заполнения средним.

Таблица 6.16. Усредненная точность для различных моделей

	Метрика	
Метод	**ACC**	**RMSE**
`Mean`	45,24%	0,0801
`CatBoost_RMSE`	47,02%	0,0756
`XGBoost_RMSE`	58,73%	0,0759
`CatBoost_MAE`	60,91%	0,0804
`XGBoost_Gamma`	70,83%	0,0877
`XGBoost_Logistic+Mean`	72,22%	0,0808
`XGBoost_Logistic+RMSE`	72,42%	0,0787
`XGBoost_Logistic+RMSE+Sigmoid`	0,034%	0,0763

Отметим, что и величины ошибок RMSE для непрерывных значений также остаются весьма умеренными. Итак, в большинстве случаев для работы с осадками более перспективным представляется использование экстремального градиентного бустинга, однако для отдельных рядов результаты могут быть улучшены за счет категориальных моделей.

6.2.4 Сравнение методов восстановления пропусков для всех тестовых метеостанций

При сравнении различных классификаторов для более 100 тестовых станций установлено, что наиболее высокие значения в среднем получаются для метода экстремального градиентного бустинга XGBClass (в среднем 83,41%), в то время как случайные леса и метод опорных векторов демонстрируют несколько более низкие, но достаточно близкие значения – 82,74% и 82,08%, соответственно. Величина RMSE для нормированных данных для различных долей пропусков в данных варьировалась от 0,01 до 0,06 (средние значения приведены в таблице 6.17). Таким образом, в качестве наиболее универсального метода может быть рекомендован экстремальный градиентный бустинг для обеих задач, однако в ряде ситуаций возможно получение некоторого дополнительного преимущества, например, за счет использования случайных лесов.

Таблица 6.17. Средние значения ошибок восстановления пропусков для всех случаев по всем анализируемым метеостанциям

Методы	RMSE
RFs+GB, RFs+XGBoost, XGBClass+GB, XGBClass+XGBoost	**0,031**
RFs+RFs, SVM+GB, SVM+RF, SVM+XGBoost, XGBClass+RF	0,032
RFs+EM, SVM+EM, XGBClass+EM	0,034

На рисунке 6.13 приведен пример географической карты с нанесенными на нее метеостанциями в разных странах и значениями RMSE для уровней в 1%, 10% и 20% пропущенных значений (случай одного подряд идущего пропуска). Также указано, использование каких именно комбинаций методов классификации и регресии привело к получению указанных значений. Легко видеть, что для станций в разных точках наиболее эффективными оказываются различные методы, однако в большинстве случаев указаны именно алгоритмы на основе экстремального градиентного бустинга.

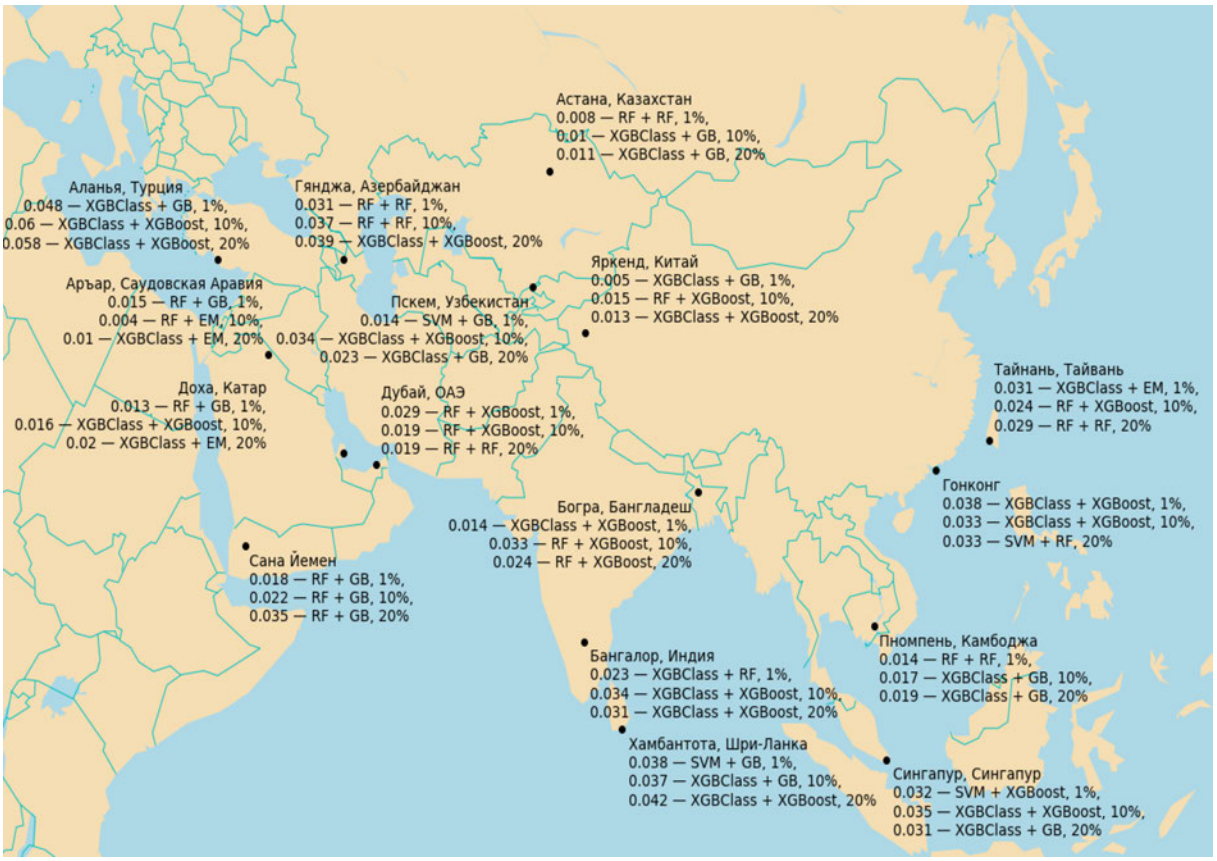

Рис. 6.13. Метеостанции и точность заполнения пропусков

В алгоритме 6.3 приведена схема реализации метода сравнения результатов различных методов машинного обучения для заполнения пропущенных значений в пространственно-временных данных.

Алгоритм 6.3. Тестирование методов заполнения пропусков в данных

```
1:  function MLTestImputation(Data, MLs)
2:      i ← 1;
3:      for all (Data) do                    // Набор тестовых метеостанций
4:          MVs ← MissingValues(Data_i);     // Внедрение пропусков
5:          F ← FeaturesSelection(Data_i);   // Извлечение признаков
6:          j ← 1;
7:          for all (MLs) do                 // Для набора алгоритмов
8:              // Заполнение искусственно внесенных пропусков
9:              [Class_{i,j}, Regr_{i,j}, Hybrid_{i,j}] ← Imput(Data_i, MVs, F, MLs_j);
10:             j++;                         // Выбор следующего алгоритма
11:         Plot(Class_i, Regr_i, Hybrid_i); // Визуализация результатов
12:         i++;                             // Выбор следующей метеостанции
13:     return [Class, Regr, Hybrid];        // Содержат данные о точности
```

6.3 Аппроксимации продолжительностей и объемов осадков за «дождливые» периоды

Модели экстремальных событий, описанные в разделе 1.3, существенным образом использовали тот факт, что интервалы времени между элементами имеют классическое или обобщенное отрицательное биномиальное распределение. В этом параграфе покажем, что для осадков в качестве таких объектов рассматриваются длины «дождливых» периодов. Также обсуждаются вопросы распределения объемов осадков, выпавших за эти периоды. Кроме того, предложена процедура функционального оценивания параметров обобщенных отрицательных биномиальных и гамма-распределений.

6.3.1 Отрицательное биномиальное распределение как модель длительностей «дождливых» периодов

Продолжительность «дождливого» периода не может составлять менее одного дня (иначе соответствующие наблюдения относятся к «сухим» периодам), при этом классическое определение отрицательного биномиального распределения с параметрами $r > 0$ и $p \in (0, 1)$ подразумевает возможность и нулевых значений. Легко видеть, что для случайной величины $X = Y + 1$ справедливы соотношения (с учетом условия $X \geqslant 1$):

$$\mathbb{P}(X = k) = \mathbb{P}(X - 1 = k - 1) = \frac{\Gamma(r + k - 1)}{\Gamma(r)\Gamma(k - 1 + 1)} p^r (1 - p)^{k-1} =$$
$$= \frac{\Gamma(r + k - 1)}{\Gamma(r)\Gamma(k)} p^r (1 - p)^{k-1}, \quad k = 1, 2, 3, \ldots$$

Поэтому для проверки гипотезы об отрицательной биномиальности распределения длительностей достаточно вычесть из исходных данных единицу и использовать стандартные процедуры. На рисунках 6.14 и 6.15 приведены примеры статистической подгонки соответствующих распределений для Потсдама и Элисты. P-значение по χ^2-критерию составило 0,2444 в первом случае и 0,1238 во втором, то есть гипотеза не отвергается. Кроме того, наглядно видно и визуальное соответствие гистограммы и подогнанных законов. Значения параметров отрицательного биноми-

ального распределения составили $r = 0.847$, $p = 0.322$ для Потсдама и $r = 0.876$, $p = 0.489$ для Элисты. Оказалось, что в обоих случаях параметр формы r меньше единицы.

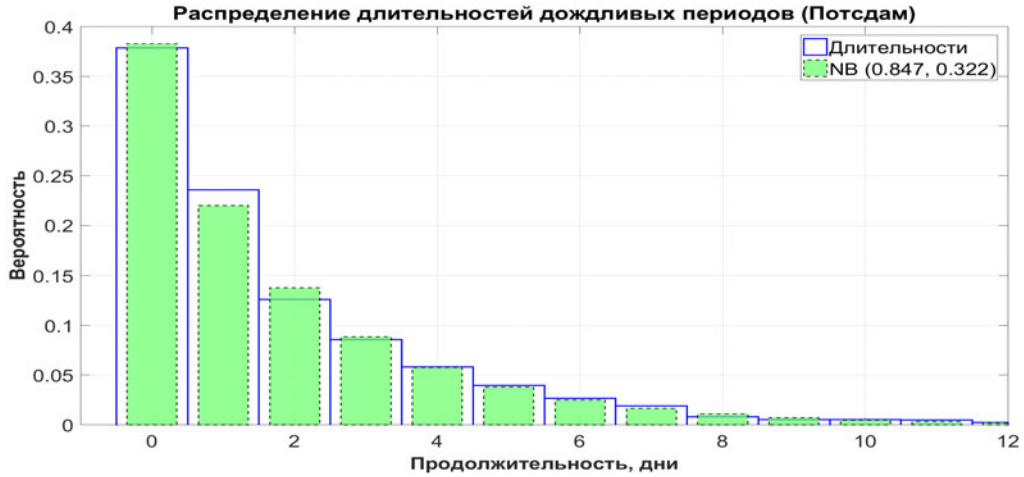

Рис. 6.14. Распределение длительностей периодов, Потсдам

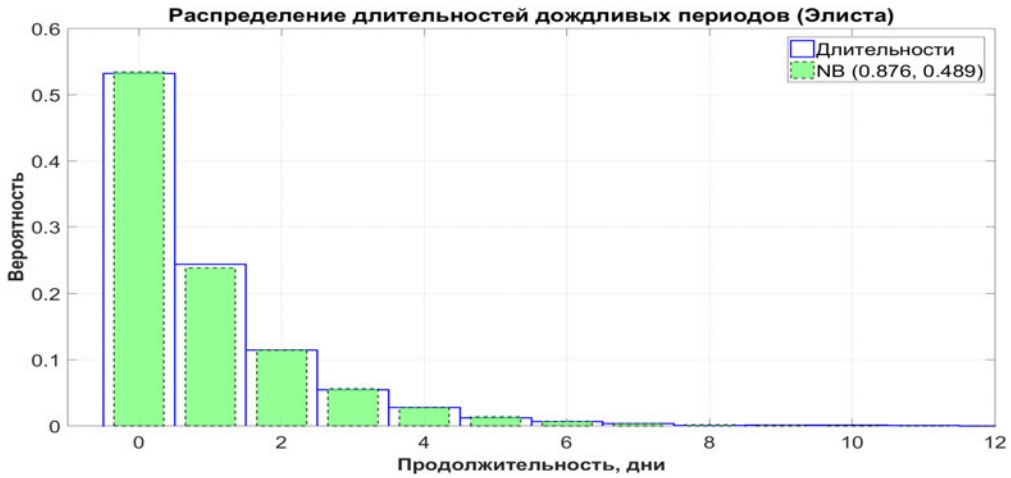

Рис. 6.15. Распределение длительностей периодов, Элиста

Можно предложить следующее объяснение данного вида распределения у длительностей «дождливых» периодов. Осадки представляют собой временной процесс. Пусть его проекция на ось абсцисс является непустым связанным (случайным) множеством. Основываясь на ряде естественных предположений для пуассоновской случайной меры [286], можно считать, что если плотность распределения «дождливых» дней равномерна по времени, то размер каждой проекции должен иметь пуассоновское распределение. Однако данное предположение из-за различных эффектов (например, наличия сезонных трендов) нарушается случайным образом, что приводит к необходимости использования смешанных пуассоновских распределений. В частности, отрицательное биномиальное является таковым (со смешивающим гамма-распределением) [254].

Действительно, для всех целых $k \geqslant 0$ имеем:

$$\frac{1}{k!}\int\limits_0^\infty e^{-\lambda}\lambda^k \frac{1}{\Gamma(r)}\Big(\frac{p}{1-p}\Big)^r \lambda^{r-1}\exp\Big\{-\frac{\lambda p}{1-p}\Big\}d\lambda =$$

$$= \Big(\frac{p}{1-p}\Big)^r \frac{1}{k!\Gamma(r)}\int\limits_0^\infty \exp\Big\{-\frac{\lambda}{1-p}\Big\}\lambda^{k+r-1}d\lambda = \frac{\Gamma(k+r)}{k!\Gamma(r)}p^r(1-p)^k.$$

Таким образом, если параметр λ пуассоновского распределения случаен и сам распределен как гамма с параметрами r и $p/(1-p)$, то соответствующее смешанное пуассоновское распределение будет отрицательным биномиальным с параметрами r и p.

Как известно [151, 286], пуассоновское распределение является наилучшей вероятностной моделью для дискретных хаотических стохастических процессов, адекватность которой обусловлена универсальным принципом неубывания энтропии в замкнутых системах. Тогда можно считать, что смешивающее гамма-распределение «случайного» параметра пуассоновского распределения в отрицательной биномиальной модели, являющееся и смешанным экспоненциальным при $0 < r \leqslant 1$ [215], описывает статистические закономерности случайных изменений внешних факторов.

В случае, если параметр $0 < r \leqslant 1$, можно показать [91], что любое отрицательное биномиальное распределение является смешанным геометрическим [90]. Указанное представление можно проинтерпретировать в терминах испытаний Бернулли со случайной вероятностью успеха. Сначала в результате «предварительного» эксперимента определяется значение вероятности успеха, а потом рассматриваемая случайная величина определяется как число успехов до первой неудачи в последовательности испытаний Бернулли с так определенной случайной вероятностью успеха. Такая интерпретация позволяет привести дополнительные аргументы, объясняющие адекватность отрицательной биномиальной модели для распределения продолжительности дождливых периодов. А именно, можно предположить, что последовательность «дождливых» и «сухих» дней не является независимой, но является условно независимой при фиксированном значении случайной величины, определяющей значение вероятности успеха, которое меняется от одного «дождливого» периода к другому (например, в зависимости от времени года) и определяется факторами, внешними по отношению к исследуемой локальной системе. Отрицательная биномиальная модель длительности «дождливых»

периодов позволяет получить асимптотические аппроксимации для таких важных характеристик выпадения осадков как распределение суммарного количества осадков, выпавших за один дождливый период, и распределение экстремального суточного количества осадков, которые будут рассмотрены в следующих разделах.

Аппроксимация распределения длительностей «дождливых» периодов была произведена для нескольких сотен метеостанций, расположенных по всему миру [413]. В частности, для 22 европейских объектов (см. раздел 6.1.6 и рисунок 6.9) за период 1904–1999 гг. получены изменения параметров отрицательного биномиального распределения за период в один год, пять, десять и двенадцать лет. Дополнительно исследовалась эволюция аналогичных параметров для «сухих» периодов, а также математические ожидания и дисперсии соответствующих распределений. Соответствующая процедура представлена в алгоритме 6.4.

Алгоритм 6.4. Анализ характеристик распределений «дождливых» и «сухих» периодов для набора метеостанций

1: **function** StationsNB(Stations, Periods $= [1, 5, 10, 12]$)
2: **for all** (Stations) **do**
3: **for all** (Period) **do**
4: *// Вычисление параметров и моментов*
5: $[\mathrm{r}_{i,j}^{(W)}, \mathrm{p}_{i,j}^{(W)}, \mathrm{r}_{i,j}^{(D)}, \mathrm{p}_{i,j}^{(D)}] \leftarrow$ NBParams(Stations$_i$, Periods$_j$);
6: $[\mathbb{E}_{i,j}^{(W)}, \mathbb{D}_{i,j}^{(W)}, \mathbb{E}_{i,j}^{(D)}, \mathbb{D}_{i,j}^{(D)}] \leftarrow$ Moments$(\mathrm{r}_{i,j}^{(W)}, \mathrm{p}_{i,j}^{(W)}, \mathrm{r}_{i,j}^{(D)}, \mathrm{p}_{i,j}^{(D)})$;
7: j++; *// Выбор следующего размера окна*
8: i++; *// Выбор следующей метеостанции*
9: PlotParams$(\mathrm{r}^{(W)}, \mathrm{p}^{(W)}, \mathrm{r}^{(D)}, \mathrm{p}^{(D)}, \mathbb{E}^{(W)}, \mathbb{D}^{(W)}, \mathbb{E}^{(D)}, \mathbb{D}^{(D)})$;
10: Save2CSV$(\mathrm{r}^{(W)}, \mathrm{p}^{(W)}, \mathrm{r}^{(D)}, \mathrm{p}^{(D)}, \mathbb{E}^{(W)}, \mathbb{D}^{(W)}, \mathbb{E}^{(D)}, \mathbb{D}^{(D)})$;
11: **return** ;

Примеры для станций с номерами 5 (Франция) и 6 (Германия) приведены на рисунках 6.16–6.19. Отмечено, что для окон меньшего размера (1 год или 5 лет) параметр формы r может принимать значения больше единицы. Однако с увеличением размера окна и для «дождливых», и для «сухих» периодов параметр формы $r \leqslant 1$, при этом отдельные отклонения наблюдаются для объектов, расположенных в горах (например, метеостанция номер 21 в Швейцарии, см. рисунок 6.9). В целом, отрицательная биномиальная модель с параметром формы меньшим единицы может считаться корректной для большинства регионов на умеренных,

но достаточно длительных временных горизонтах.

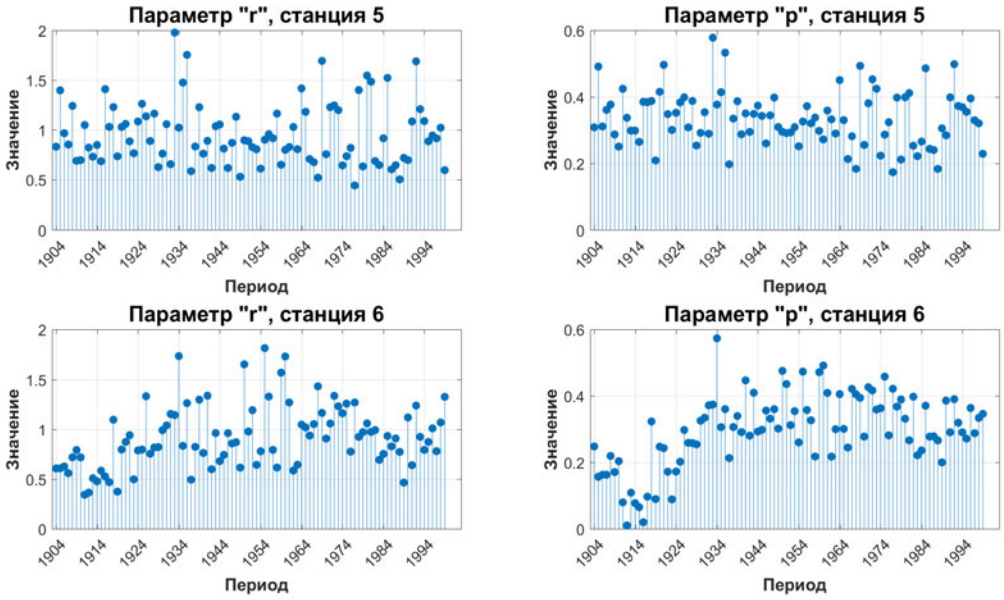

Рис. 6.16. «Дождливые» периоды, станции 5 и 6, окно 1 год

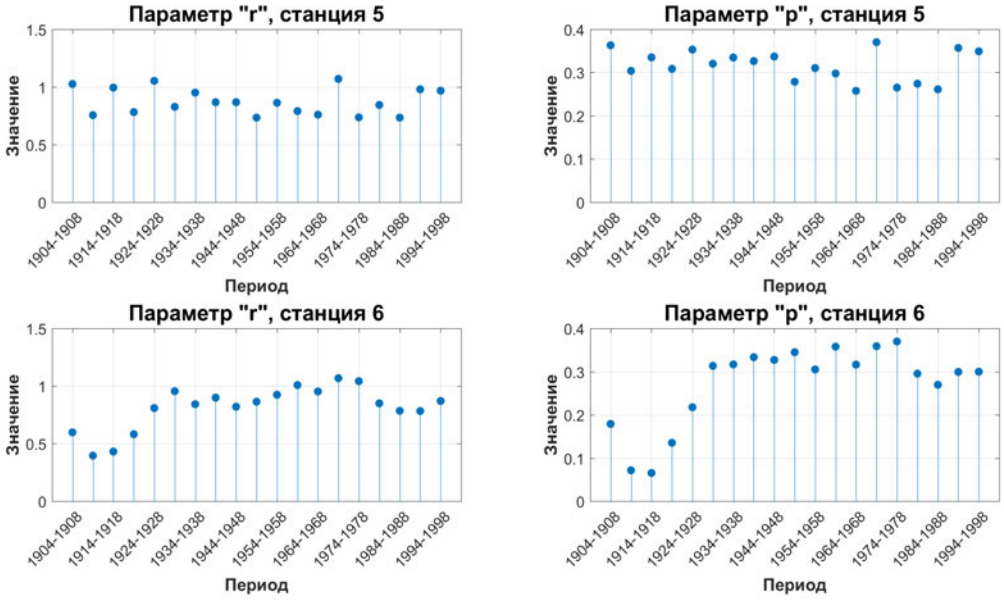

Рис. 6.17. «Дождливые» периоды, станции 5 и 6, окно 5 лет

Длительные периоды позволяют более четко выявлять долгосрочные тенденции в изменении соответствующих характеристик распределений, при этом есть некоторые различия в характере локальных и глобальных трендов. Полученные результаты могут быть использованы непосредственно для анализа изменений климата в различных регионах и прогнозирования явлений, связанных с осадками. Кроме того, высокое согласие между распределением длительностей «дождливых» периодов и отрицательной биномиальной моделью будут использованы при статистическом определении экстремальных объемов осадков.

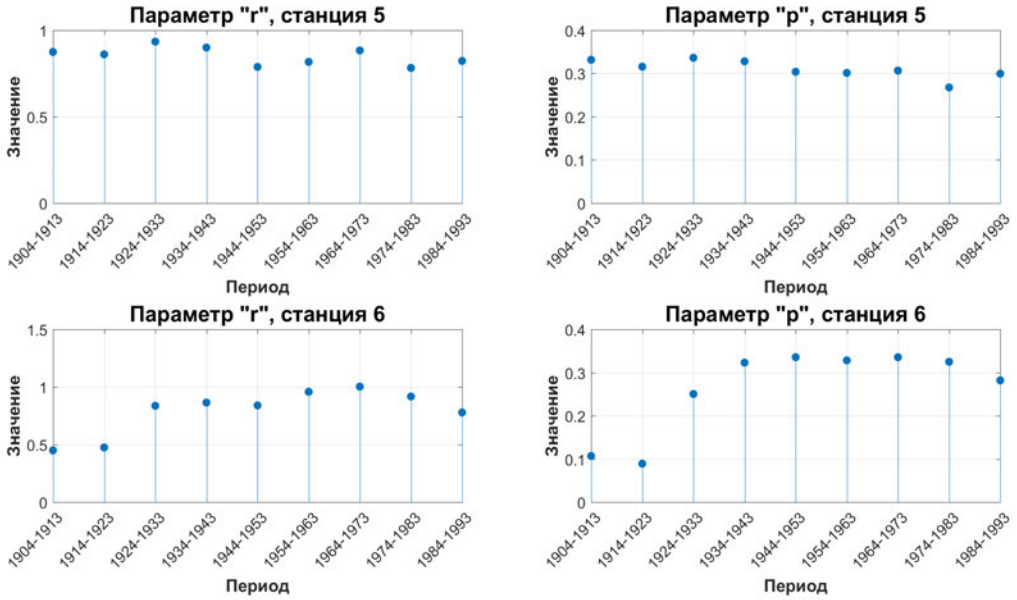

Рис. 6.18. «Дождливые» периоды, станции 5 и 6, окно 10 лет

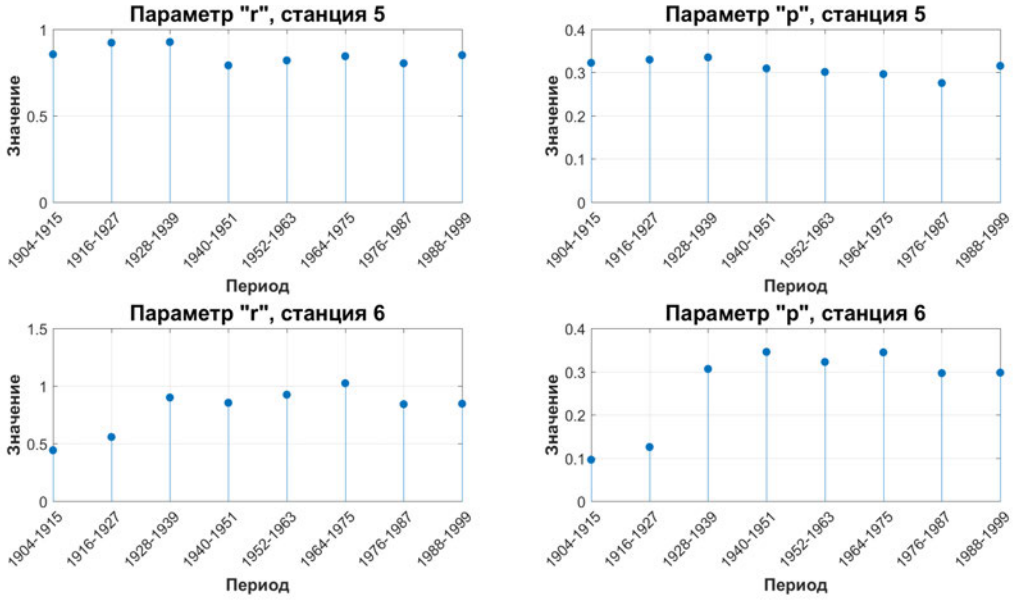

Рис. 6.19. «Дождливые» периоды, станции 5 и 6, окно 12 лет

6.3.2 Распределение объемов осадков

В этом разделе рассмотрим аппроксимацию выборочных распределений для величин суточных осадков и суммарных объемов за «дождливые» периоды. Эмпирически было установлено, что есть высокое соответствие между данными и гамма- и обобщенными распределениями Парето, функция распределения которого имеет вид ($\xi \in \mathbb{R}$ – параметр формы, $\mu \in \mathbb{R}$ – сдвига, $\sigma > 0$ – масштаба):

$$F_{\xi,\sigma,\mu}(y) = \begin{cases} 1 - \left(1 + \frac{\xi(y-\mu)}{\sigma}\right)^{-1/\xi}, & \text{если } \xi \neq 0, \\ 1 - e^{-\frac{y-\mu}{\sigma}} & \text{иначе.} \end{cases} \tag{6.3}$$

На рисунках 6.20– 6.23 продемонстрирована аппроксимация распределений величин осадков и суточных объемов за «дождливые» периоды в Потсдаме и Элисте.

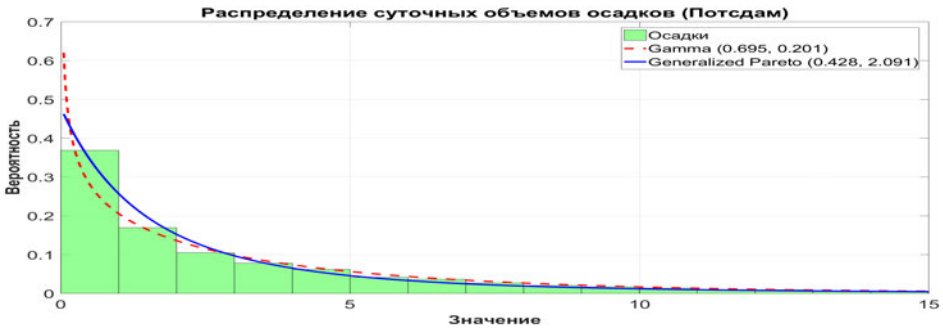

Рис. 6.20. Распределение суточных объемов осадков, Потсдам

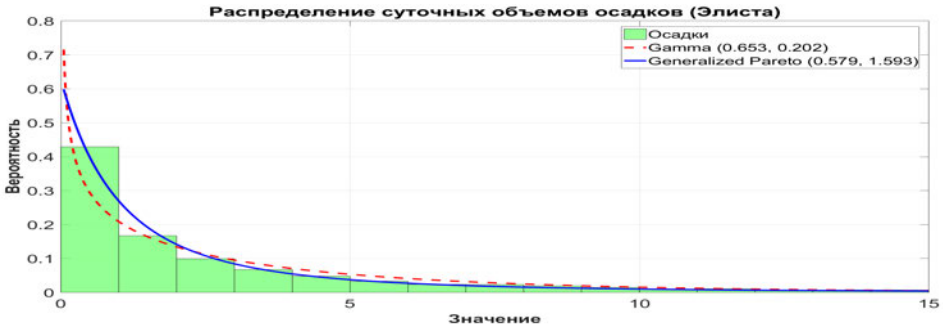

Рис. 6.21. Распределение суточных объемов осадков, Элиста

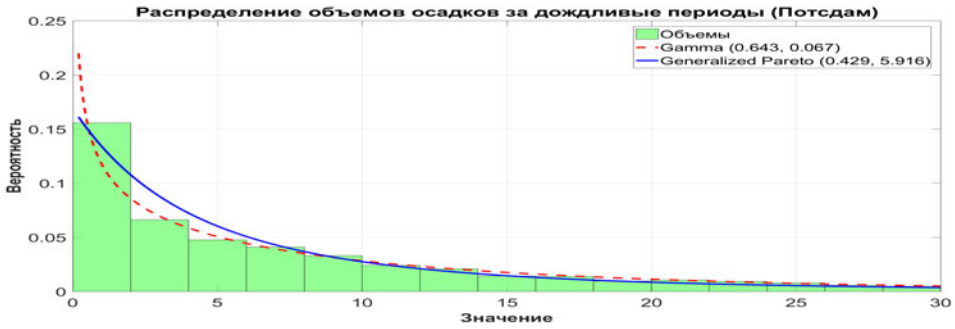

Рис. 6.22. Распределение объемов осадков за периоды, Потсдам

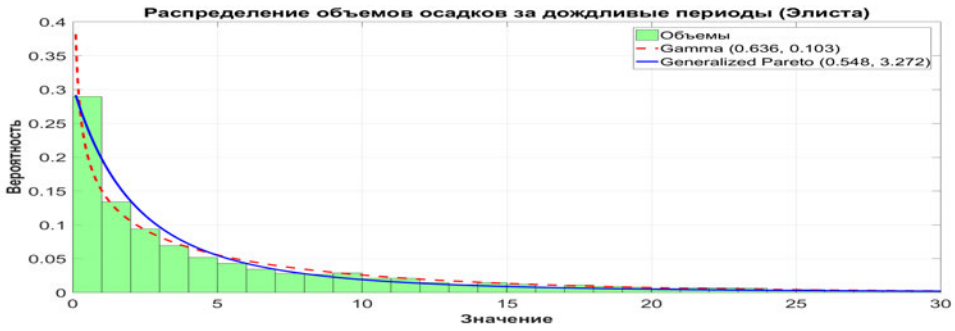

Рис. 6.23. Распределение объемов осадков за периоды, Элиста

Можно показать, что масштабная смесь гамма-распределений со смешивающим экспоненциально законом является распределением типа Парето. Действительно, имеем

$$\int_0^\infty \lambda e^{-\lambda x}\frac{\mu^s}{\Gamma(s)}\lambda^{s-1}e^{-\mu\lambda}d\lambda = \frac{s\mu^s}{(x+\mu)^{1+s}}, \quad x>0,$$

то есть для любых $s>0$, $\mu>0$ распределение Парето с указанной плотностью является смешанным экспоненциальным. Более того, рассмотрим следующие равенства (для $x>0$):

$$\frac{x^{r-1}}{\Gamma(r)}\int_0^\infty \lambda^r e^{-\lambda x}\frac{\mu^s}{\Gamma(s)}\lambda^{s-1}e^{-\mu\lambda}d\lambda = \frac{x^{r-1}\mu^s}{\Gamma(r)\Gamma(s)}\int_0^\infty \lambda^{r+s-1}e^{-\lambda(x+\mu)}d\lambda =$$

$$= \frac{x^{r-1}\mu^s}{\Gamma(r)\Gamma(s)(x+\mu)^{r+s}}\int_0^\infty (x+\mu)^{r+s-1}\lambda^{r+s-1}e^{-\lambda(x+\mu)}d\lambda(x+\mu) =$$

$$= \frac{\Gamma(r+s)\mu^s}{\Gamma(r)\Gamma(s)}\frac{x^{r-1}}{(x+\mu)^{r+s}}.$$

То есть произвольная масштабная смесь гамма-распределений со смешивающим гамма является распределением типа Парето. В случае, если параметр формы удовлетворяет условию $s\leqslant 1$, тогда такое распределение Парето также является смешанным экспоненциальным. Как известно, экспоненциальное распределение обладает максимальной дифференциальной энтропией среди всех, сосредоточенных на неотрицательной полуоси и имеющих конечный первый момент. Таким образом, использование подобной модели позволяет статистически корректно учесть влияние случайных факторов на данные об осадках.

6.3.3 Функциональный подход к оцениванию параметров обобщенных отрицательных биномиальных распределений

Задача статистического оценивания параметров обобщенных отрицательных биномиальных и гамма-распределений является достаточно сложной. Известны подходы на основе классических методов моментов и максимального правдоподобия [271] для обобщенного гамма-распределения, которые ведут к требовательным к ресурсам вычислительным процедурам (например, при аппроксимации второй и третьей производных полигамма-функции [439] или независимые от масштаба

уравнениях оценки формы [386]). Более того, они существенным образом зависят от размера выборки: например, метод максимального правдоподобия ведет к лучшим результатам для больших наборов. Также возможно использование так называемых сеточных методов [88, 234], которые являются достаточно эффективными, но только в случае оптимального выбора многомерной параметрической сетки – данная задача в общем виде до сих пор не решена. Кроме того, оценки сеточного метода являются состоятельными только в случае, если измельчение параметрического пространства растет с объемом выборки, а соответствующие условия достаточно сложно проверяемы на практике. В этом и следующем разделах будет предложен функциональный метод оценивания параметров, основанный на минимизации расстояний ℓ^1-, ℓ^2- и ℓ^∞ для обобщенного отрицательного биномиального распределения и метрик L^1-, L^2- и L^∞ для обобщенного гамма. Данный подход основывается только на решении оптимизационных задач без привлечения каких-либо иных сложных вычислительных процедур. Кроме того, получаемые таким методом оценки являютс не точечными, а функциональными – для распределения (плотности). Таким образом, множеством решений является не Евклидово, а функциональное пространство.

Пусть построена гистограмма для исходных данных – длительностей «дождливых» периодов. Они могут принимать только целочисленные значения, что учитывается при разбиении интервала возможных значений (столбцы располагаются в целых точках). Пусть N_b – число столбцов одинаковой единичной ширины, $\mathbf{h}$ – вектор их высот, причем каждая компонента $h_i \in [0,1]$ для всех номеров $i = \overline{1, N_b}$. Величины h_i определяются как отношение числа наблюдений, попавших в соответствующий интервал, к общему числу элементов в выборке, поэтому сумма площадей под столбиками равна 1. Как и ранее (см. раздел 1.3), плотность обобщенного гамма-распределения обозначается как $f^{GG}_{r,\gamma,\mu}(x)$ (1.10), а обобщенное отрицательное биномиальное распределение для целочисленных k задается величинами $\mathbb{P}(N_{r,\gamma,\mu} = k)$ (1.12).

Предложение 6.1. *Для оценивание параметров обобщенного отрицательного биномиального распределения* (1.12) *необходимо решить одну из следующих оптимизационных задач:*

– *в метрике* ℓ^1*:*

$$(\widehat{r}, \widehat{\gamma}, \widehat{\mu}) = \arg\min_{r,\gamma,\mu} \sum_{k=1}^{N_b} \left| \frac{1}{k!} \int_0^\infty e^{-z} z^k f^{GG}_{r,\gamma,\mu}(z) dz - h_k \right|; \tag{6.4}$$

– *в метрике* ℓ^2*:*

$$(\widehat{r},\widehat{\gamma},\widehat{\mu}) = \arg\min_{r,\gamma,\mu} \sqrt{\sum_{k=1}^{N_b}\left(\frac{1}{k!}\int_0^\infty e^{-z}z^k f^{GG}_{r,\gamma,\mu}(z)dz - h_k\right)^2}; \tag{6.5}$$

– *в метрике* ℓ^∞*:*

$$(\widehat{r},\widehat{\gamma},\widehat{\mu}) = \arg\min_{r,\gamma,\mu} \max_{k=\overline{1,N_b}}\left|\frac{1}{k!}\int_0^\infty e^{-z}z^k f^{GG}_{r,\gamma,\mu}(z)dz - h_k\right|. \tag{6.6}$$

Отметим, что поскольку классическое отрицательное биномиальное распределение является частным случаем обобщенного (при $\gamma = 1$), для оценивания его параметров может быть использована подобная процедура, хотя с вычислительной точки зрения в данной ситуации значительно проще найти оценки методом моментов или максимального правдоподобия. Кроме того, в ряде моделей параметр r может считаться известным (и совпадать с аналогичным для классического распределения). В случае такого фиксированного значения в оптимизационных задачах (6.4)–(6.6) ищутся только оценки $\widehat{\gamma}$ и $\widehat{\mu}$ при заданном значении r в формулах в правой части.

На рисунках 6.24 и 6.25 для 3323 и 2937 «дождливых» периодов за шестидесятилетнюю историю наблюдений в Потсдаме и Элисте приведена функциональная аппроксимация эмпирического распределения длительностей классическим и обобщенным отрицательным биномиальном законом для метрики ℓ^1 при фиксированном параметре r (его значение было определено выше – 0,847 для Потсдама и 0,876 для Элисты). Видно высокое визуальное соответствие обоих законов распределения, при этом для Потсдама (с точностью до округления) величина ошибки в каждой из метрик для каждого из распределений сопоставима, а для Элисты обобщенный закон дает меньшую ошибку во всех случаях.

Дополнительные графики и описание программного решения, разработанного для проведения указанной аппроксимации, рассмотрены в разделе 7.2. В таблице 6.18 приведены результаты аппроксимации при решении оптимизационной задачи сразу для всех трех параметров для Потсдама. Очевидно, что во всех случаях обобщенное распределение продемонстрировало лучшие результаты (они выделены жирным шрифтом). Таким образом, его использование целесообразно на практике, несмотря на отсутствие в стандартных статистических пакетах методов для анализа его параметров.

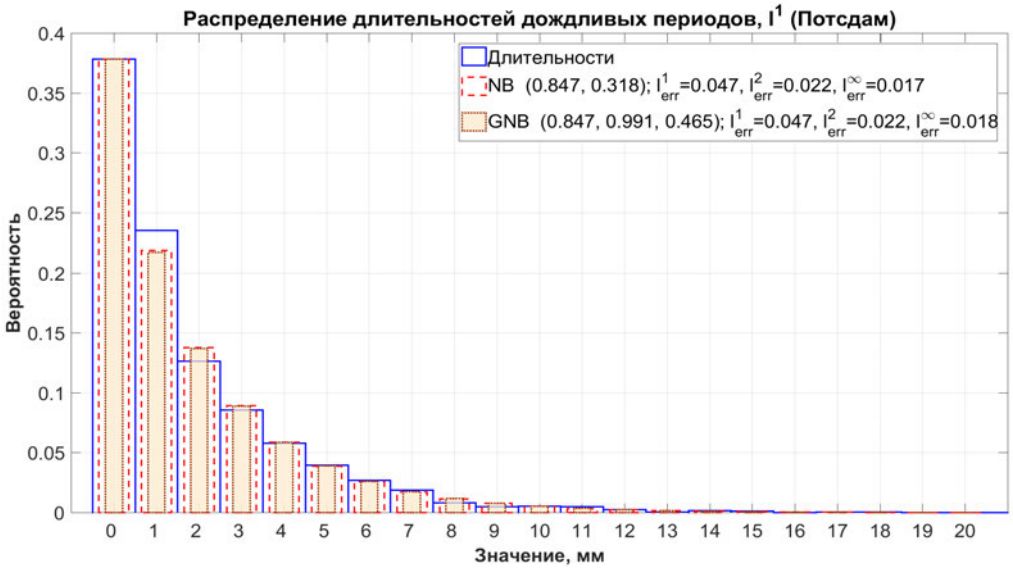

Рис. 6.24. Распределение длительностей «дождливых» периодов в Потсдаме, пример аппроксимации в ℓ^1

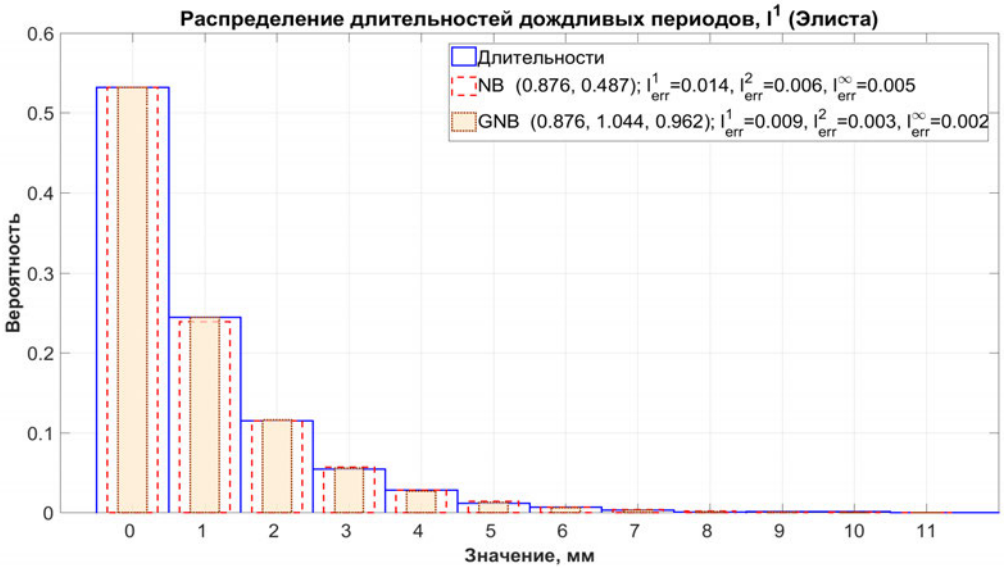

Рис. 6.25. Распределение длительностей «дождливых» периодов в Элисте, пример аппроксимации в ℓ^1

Таблица 6.18. Ошибки функциональной аппроксимации длительностей «дождливых» периодов в различных метриках, Потсдам

Приближающее распределение	**Ошибка**		
	ℓ^1	ℓ^2	ℓ^∞
NB (ℓ^1-оптимизация)	0,047	0,022	0,018
GNB (ℓ^1-оптимизация)	**0,043**	0,018	0,014
NB (ℓ^2-оптимизация)	0,051	0,0195	0,015
GNB (ℓ^2-оптимизация)	0,05	**0,015**	0,0097
NB (ℓ^∞-оптимизация)	0,061	0,022	0,012
GNB (ℓ^∞-оптимизация)	0,061	0,017	**0,007**

Процедура функциональной оптимизации параметров классического и обобщенного отрицательных биномиальных распределений представлена в алгоритме 6.5.

Алгоритм 6.5. Функциональная оптимизация для оценивания параметров обобщенного отрицательного биномиального распределения

function GNBApprox(Data, Metrics=$[\ell^1, \ell^2, \ell^\infty]$, α=0.05, $\mathrm{r}_{fix}=\emptyset$)
  // *Определение длительностей «дождливых» периодов*
  Lengths $\leftarrow$ WetPeriodsLengths(Data);
  if ($\mathrm{r}_{fix} \neq \emptyset$) **then**
    // *Функциональная оптимизация, если* $\alpha \equiv 0$
    r_{NB}=NBFit(WetP-1, α);
  for all (Metrics) **do** // *Решение задач* (6.4)–(6.6)
    // *Если* $r_{fix} = \emptyset$, *то параметр* r_i=r_{NB}
    $[r_i, \gamma_i, \mu_i, err_i] \leftarrow$ GNBEval(Lengths, $\mathrm{Metrics}_i$, r_{fix});
  PlotGNB(r, γ, μ, err); // *Визуализация результатов*
  return $[r, \gamma, \mu, err]$;

6.3.4 Функциональный подход к оцениванию параметров обобщенных гамма-распределений

В данном разделе рассмотрим решения для функционального оценивания параметров обобщенных гамма-распределений. Пусть построена гистограмма для исходных данных – объемов осадков за «дождливые» периоды. Для разбиения на интервала при построении гистограммы используется правило Фридмана–Дьякониса [203], хорошо подходящее для распределений с тяжелыми хвостами:

$$2\,\frac{x_{0,75} - x_{0,25}}{\sqrt[3]{n}}, \tag{6.7}$$

где $x_{0,25}$, $x_{0,75}$ – квантили уровней 0,25 и 0,75, числитель дроби в (6.7) представляет собой интерквартильный размах, а через n обозначен объем выборки. Пусть $\mathbf{b}$ – вектор границ полученных разбиений, а величины N_b и $\mathbf{h}$ определены также, как и в предыдущем пункте (с поправкой на тот факт, что столбцы больше не имею единичной ширины).

Предложение 6.2. *Для оценивание параметров обобщенного гамма-распределения с плотностью* $f^{GG}_{r,\gamma,\mu}(x)$ (1.10) *необходимо решить одну из следующих оптимизационных задач:*

– *в метрике* L^1*:*

$$(\widehat{r},\widehat{\gamma},\widehat{\mu}) = \arg\min_{r,\gamma,\mu} \sum_{k=1}^{N_b-1} \int_{b_k}^{b_{k+1}} \left| f^{GG}_{r,\gamma,\mu}(z) - h_k \right| dz; \tag{6.8}$$

– *в метрике* L^2*:*

$$(\widehat{r},\widehat{\gamma},\widehat{\mu}) = \arg\min_{r,\gamma,\mu} \sqrt{\sum_{k=1}^{N_b-1} \int_{b_k}^{b_{k+1}} \left(f^{GG}_{r,\gamma,\mu}(z) - h_k \right)^2 dz}; \tag{6.9}$$

– *в метрике* L^∞*:*

$$(\widehat{r},\widehat{\gamma},\widehat{\mu}) = \arg\min_{r,\gamma,\mu} \max_{k\in[1,N_b-1]} \int_{b_k}^{b_{k+1}} \left| f^{GG}_{r,\gamma,\mu}(z) - h_k \right| dz. \tag{6.10}$$

Как было отмечено для случая отрицательного биномиального распределения, этот подход может быть использован и для оценивания параметров классического гамма-распределения, а также допускается возможность зафиксировать параметр r. На рисунках 6.26 и 6.27 приведена функциональная аппроксимация эмпирического распределения объемов осадков за «дождливые» периоды, рассмотренные в предудущем разделе, классическим и обобщенным гамма-распределениями для метрики L^2 как при фиксированном параметре r (синяя пунктирная линия на графиках), так и для случая оценивания сразу всех трех параметров.

Очевидно высокое визуальное соответствие данных и приведенных законов, при этом случай «свободного» r дает меньшую ошибку и для Потсдама, и для Элисты. Дополнительные графики и описание программного решения, разработанного для проведения указанной аппроксимации, рассмотрены в разделе **??**. В таблице 6.19 приведены результаты аппроксимации при решении оптимизационной задачи сразу для всех трех параметров для Потсдама. Очевидно, что во всех случаях обобщенное распределение продемонстрировало лучшие результаты (они выделены жирным шрифтом). Безусловно, подобное усложнение моделей влечет необходимость реализации отличающихся от стандартных методов вычислительных процедур, однако за счет этого может быть достигнут дополнительный эффект с точки зрения качества аппроксимации реальных данных. Таким образом, разработанные программные решения (см. раздел **??**) необходимы для проведения более тонкого анализа пространственно-временных наблюдений.

Рис. 6.26. Распределение объемов осадков за «дождливые» периоды в Потсдаме, пример аппроксимации в L^2

Рис. 6.27. Распределение объемов осадков за «дождливые» периоды в Элисте, пример аппроксимации в L^2

Таблица 6.19. Ошибки функциональной аппроксимации объемов осадков за «дождливые» периодов в различных метриках, Потсдам

Приближающее распределение	**Ошибка**		
	L^1	L^2	L^∞
Гамма (L^1-оптимизация)	0,18	0,1158	0,1035
GG (L^1-оптимизация)	**0,1516**	0,0541	0,0541
Гамма (L^2-оптимизация)	0,2653	0,068	0,0594
GG (L^2-оптимизация)	0,1599	**0,0517**	0,0473
Гамма (L^∞-оптимизация)	0,2789	0,0685	0,0521
GG (L^∞-оптимизация)	0,0412	0,0531	**0,0412**

Процедура функциональной оптимизации параметров обобщенного гамма-распределения представлена в алгоритме 6.6.

Алгоритм 6.6. Функциональная оптимизация для оценивания параметров обобщенного гамма-распределения

function GGApprox(Data, Metrics=$[L^1, L^2, L^\infty]$, α=0.05, $\mathrm{r}_{fix}=\emptyset$)
  // *Определение объемов осадков за «дождливые» периоды*
  Vols $\leftarrow$ WetPeriodsVol(Data);
  if ($\mathrm{r}_{fix} \neq \emptyset$) **then**
    // *Функциональная оптимизация, если* $\alpha \equiv 0$
    r_γ=GammaFit(Vols, α);
  for all (Metrics) **do** // *Решение задач* (6.8)–(6.10)
    // *Если* $r_{fix} = \emptyset$, *то параметр* r_i=r_γ
    $[r_i, \gamma_i, \mu_i, err_i] \leftarrow$ GGEval(Vols, $\mathrm{Metrics}_i$, r_{fix});
  PlotGG(r, γ, μ, err); // *Визуализация результатов*
  return $[r, \gamma, \mu, err]$;

6.3.5 Стабилизация суммарных осадков за «дождливые» периоды

В процессе изучения реальных данных было установлено, что для объемов осадков за «дождливые» периоды $X_1, X_2, \ldots$ не выполняется классический закон больших чисел, при котором усредненные суммы $\frac{1}{n}(X_1 + \ldots + X_n)$ должны были бы сходиться почти наверное к некоторой константе a при $n \to \infty$. Действительно, на рисунке 6.28 видно, что для Потсдама наблюдается медленный убывающий тренд, в то время как для Элисты – возрастающий.

Для учета наличия трендов для случайных величин $X_1, X_2, \ldots$ (не обязательно независимых и одинаково распределенных) потребуем выполнение условия (1.32).

Предложение 6.3. *Предположим, что* $X_1, X_2, \ldots, X_n$ *– наблюдаемые значения последовательных дней с ненулевыми объемами осадков,* $n \in \mathbb{N}$ *– общее количество доступных наблюдений. Обозначим*

$$s_k = X_1 + \ldots + X_k$$

для всех натуральных $k = 1, \ldots, n$. *Если выполнено условие* (1.32), *то для достаточно большого* k $(1 \leqslant m \leqslant k \leqslant n)$, *могут быть использованы следующие оценки параметров* a *и* β:

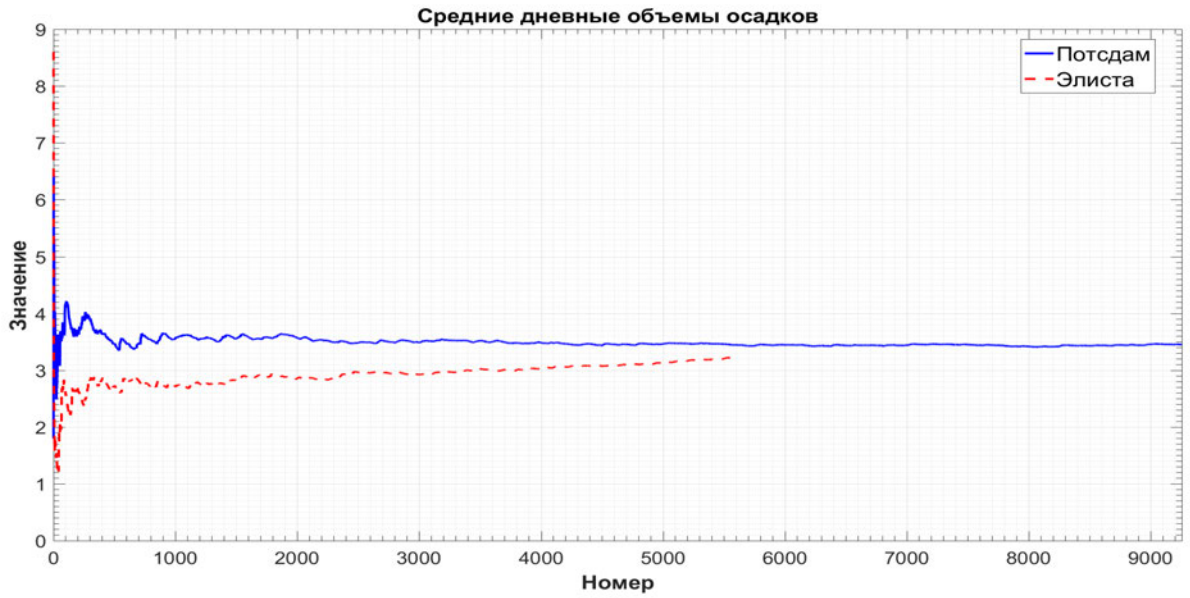

Рис. 6.28. Нарушение классического закона больших чисел для осадков

$$\widehat{a} = \exp\left\{ \frac{\sum\limits_{k=m}^{n} \log s_k \cdot \sum\limits_{k=m}^{n} (\log k)^2 - \sum\limits_{k=m}^{n} \log k \cdot \sum\limits_{k=m}^{n} \left(\log k \cdot \log s_k\right)}{(n-m+1)\sum\limits_{k=m}^{n} (\log k)^2 - \left(\sum\limits_{k=m}^{n} \log k\right)^2} \right\}, \tag{6.11}$$

$$\widehat{\beta} = \frac{\sum\limits_{k=m}^{n} \log s_k - (n-m+1)\log \widehat{a}}{\sum\limits_{k=m}^{n} \log k}. \tag{6.12}$$

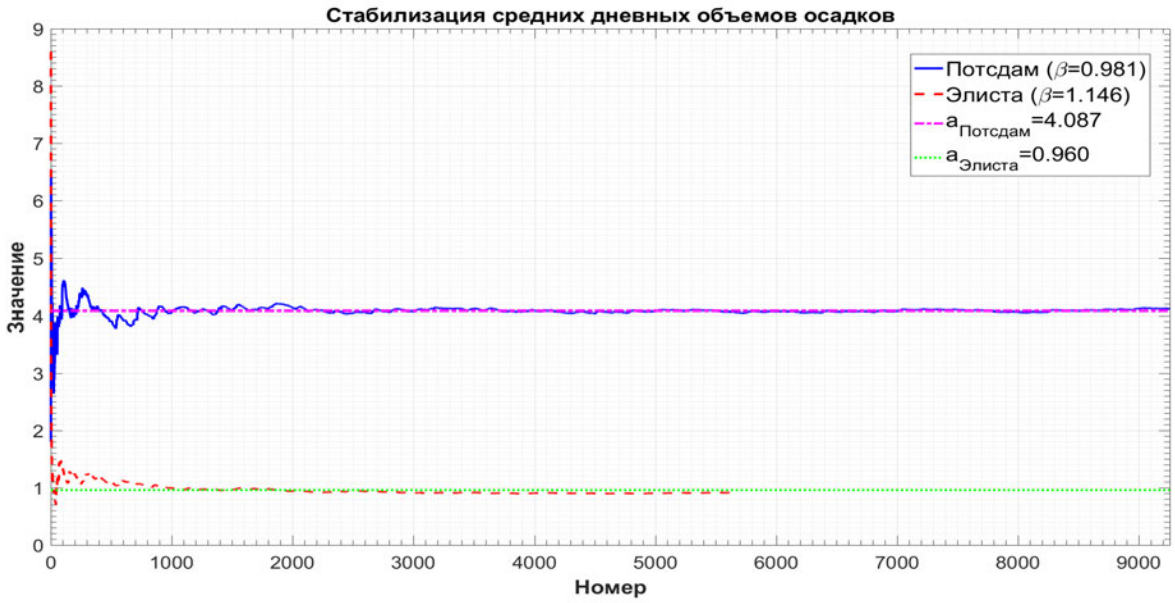

Рис. 6.29. Стабилизация суммарных осадков за «дождливые» периоды

Результаты подбора параметров, компенсирующих наличие трендов,

представлены на рисунке 6.29. Получено, что для Потсдама $a = 4.087$ и $\beta = 0.981$, в то время как для Элисты $a = 0.96$ и $\beta = 1.146$. Для получения оценок неизвестных параметров a и β можно воспользоваться методом наименьших квадратов (МНК). Действительно, если выполнено условие (1.32), то можно записать следующее приближенное равенство:

$$\frac{s_k}{k^\beta} \approx a \Longleftrightarrow -\beta \log k + \log s_k \approx \log a.$$

Тогда использование МНК для решения задачи

$$(\widehat{a}, \widehat{\beta}) = \arg\min_{\beta,\, \log a} \sum_{k=m}^{n} (\log s_k - \beta \log k - \log a)^2$$

ведет непосредственно к формулам (6.11) и (6.12). Данная процедура представлена в алгоритме 6.7.

Алгоритм 6.7. Определение параметров стабилизации усредненных объемов

```
1: function StabParams(Data)
2:     m ← 3000;                          // Величина выбирается эмпирически
3:     Data+ ← find(Data>0);
4:     // Реализация МНК, см. формулы (6.11) и (6.12)
5:     [CumSum, a, β] ← LS(Data+);
6:     PlotAverages(CumSum, a, β);                          // Визуализация
7:     return [CumSum, a, β];
```

С учетом результатов теоремы 1.8, обобщающей классические результаты Реньи для случая отрицательных биномиальных сумм, которые являются адекватными аппроксимациями для распределений длительностей «дождливых» периодов, обобщенные гамма-распределения с умеренными значениями μ могут рассматриваться как адекватная и теоретически обоснованная модель для осадков за достаточно длительные интервалы.

6.4 Статистические методы определения экстремальности осадков

Как уже было отмечено во введении к данной главе, оценки закономерности и трендов в экстремальных осадках крайне важны для понимания процессов изменения климата на различных временных горизонтах. Однако выводы по суточным или усредненным по «дождливым»

периодам объемам могут существенно отличаться [445]. Это может происходить как из-за большей чувствительность ежедневных наблюдений к пропущенным значениям, так и из-за использования не вполне подходящих математических моделей. Наконец, нет однозначного подхода, какие же наблюдения должны считаться аномальными в силу того, что подобные события являются редкими.

Существует несколько способов определения экстремальности объемов осадков. Во-первых, это могут быть все наблюдения, которые за некоторый выбранный период времени превышают 95%-ю выборочную квантиль, при этом дни без осадков также учитываются. Во-вторых, возможно рассмотрение данных только за «дождливые» периоды, при этом определение аномальности и соответствующие статистические характеристики будут отличаться от предыдущего случая [443]. Достаточно популярным [148, 306, 369] является и подход, основанный на классической теории экстремальных значений, называемый `Peaks over Threshold` (`PoT`) ключевой недостаток которого заключается в необходимости определения распределения, превышение квантилей заданного уровня которого и считается индикатором экстремальности наблюдений.

В данном разделе разработаны несколько методов, лишенные указанных недостатков. Во-первых, будут предложены непараметрические критерии на основе модификации `PoT`-методов с учетом полученных в разделе 1.3 обобщений классической теоремы Реньи для редеющих потоков [217] и теоремы Пикандса–Балкемы–Де Хаана [141, 353]. Отметим также, что с использованием результатов данной теоремы возможно построение эффективных вероятностных прогнозов различных катастрофических событий, например, землетрясений в Арктике [92]. На основе модели обобщенных гамма-распределений для объемов и интенсивностей осадков будут предложены параметрические статистические критерии для определения типа аномальности каждого наблюдения. С использованием темперированного распределения Снедекора-Фишера (см. раздел 1.3) будет продемонстрирован квантильный подход к определению экстремальных наблюдений.

6.4.1 Модифицированный метод превышения порогового значения

Рассмотрим подход на основе двух фундаментальных результатов – теорем Реньи и Пикандса–Балкемы–Де Хаана, которые позволяют из-

бегать априорных предположений о распределении данных для выбора порогового значения. Из указанных выше теорем следует, что распределение разностей моментов превышения порогового значения должно соответствовать экспоненциальному закону, а величины превышений данного порога – обобщенному распределению Парето (6.3). Модифицированный PoT-метод представлен в алгоритме 6.8.

Алгоритм 6.8. Метод превышения порогового значения

function PoT(Data, γ=1, step=0.01, α=0.01, dir =↓)
  // *Реализация восходящего (↑) и нисходящего (↓) методов*
  i ← 1;
  $\mathrm{LVL}_i \leftarrow \mathrm{MIN}(\mathrm{Data})\cdot\mathcal{I}_{\uparrow}(dir)+\mathrm{MAX}(\mathrm{Data})\cdot\mathcal{I}_{\downarrow}(dir)$;
  repeat
    Ind← FIND(Data>LVL); // *Моменты превышения порога*
    dInd← DIFF(Ind);
    // *Проверка гипотез о распределениях*
    $\mathrm{p}^{(W)}_{val,i} \leftarrow$ FitDist(dInd, 'Weibull', γ, α); // *Вейбулловость*
    // *Паретовость превышений*
    $\mathrm{p}^{(GP)}_{val,i} \leftarrow$ FitDist(Data(dInd)-LVL_i, 'GeneralizedPareto', α);
    // *Смещение уровня в зависимости от типа метода*
    i++;
    $\mathrm{LVL}_i \leftarrow \mathrm{LVL}_{i-1}+\mathrm{step}\cdot(\mathcal{I}_{\downarrow}(dir) - \mathcal{I}_{\uparrow}(dir))$;
    // *Критерий останова определяется типом метода*
  until $(\mathrm{LVL}_i \leqslant \mathrm{MAX}(\mathrm{Data}))\cdot\mathcal{I}_{\uparrow}(dir)$ **or** $(\mathrm{LVL}_i \geqslant \mathrm{MIN}(\mathrm{Data}))\cdot\mathcal{I}_{\downarrow}(dir)$
  return $[\mathrm{p}^{(W)}_{val}, \mathrm{p}^{(GP)}_{val}, \mathrm{LVL}]$;

Пороговое значение может быть определено в рамках статистической процедуры, в которой для каждого уровня, начиная с некоторого значения, например минимума в данных, с заранее заданным шагом должны проверяться последовательно две гипотезы об экспоненциальности и паретовости описанных выше объектов. В случае принятия обеих текущий уровень может считаться экстремальным пороговым значением. Назовем данный метод восходящим – в соответствии с направлением сдвига порогового значения в процессе анализа данных. Очевидно, что данная процедура может быть реализована и в обратном направлении (с точки зрения смещения уровня в процессе анализа данных). Для этого необходимо задать верхнюю границу (например, совпадающую с максимумом наблюдений) и последовательно проверять гипотезы об экспоненциальности и

паретовости. Необходимо учитывать, что в данном случае на начальных этапах работы алгоритма выборка превышений будет иметь малый объем, поэтому корректная проверка гипотез затруднительна. Необходимо обеспечить минимально приемлемый объем соответствующей выборки. Такой метод назовем нисходящим. Важно отметить, что обе процедуры могут применяться к данным любого знака, независимо от их распределения – теория экстремальных значений гарантирует корректность результатов при выполнении условий регулярности.

Продемонстрируем применение описанного алгоритма на примере наблюдений в Потсдаме и Элисте (см. рисунки 6.30 и 6.31 соответственно).

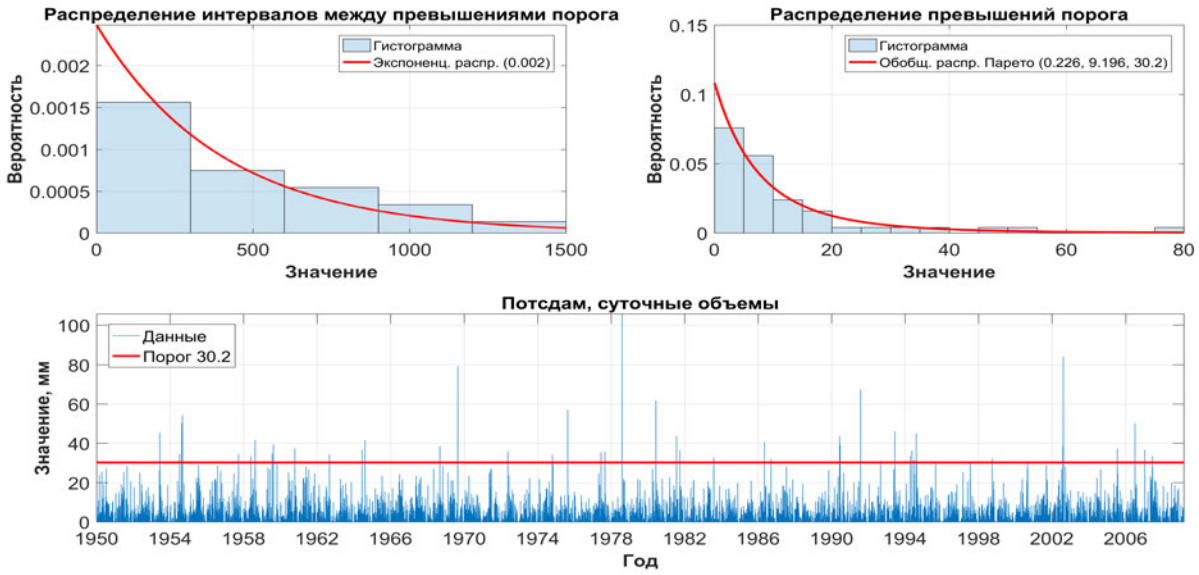

Рис. 6.30. Пороговый уровень для суточных объемов осадков, Потсдам

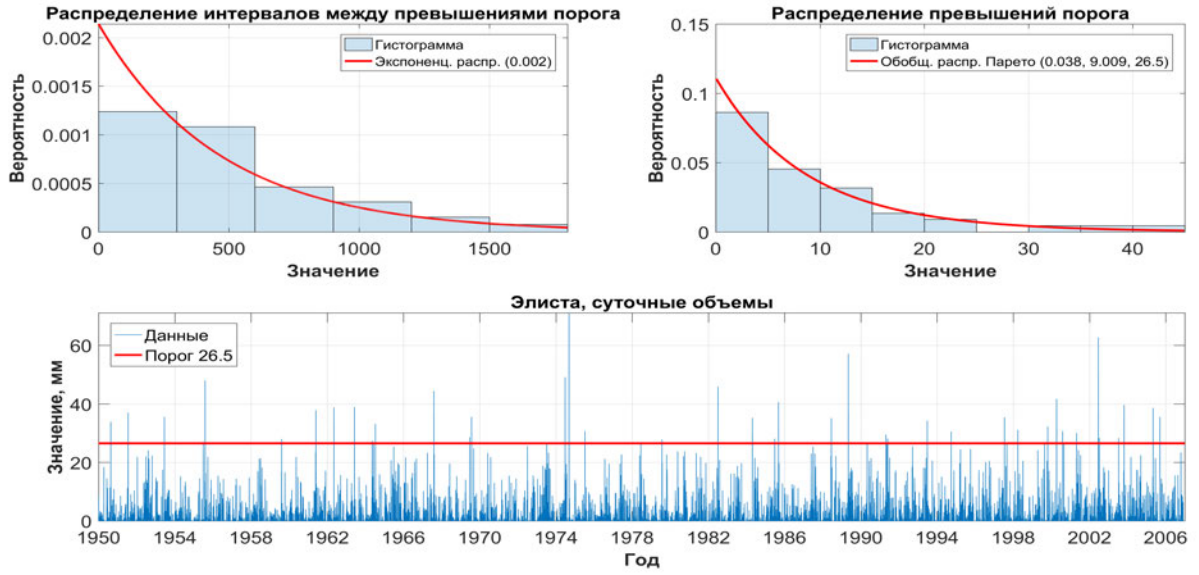

Рис. 6.31. Пороговый уровень для суточных объемов осадков, Элиста

Отметим, что в алгоритме 6.8 для промежутков между превышениями пороговых значений указан вызов функции `FitDist` статистической проверки соответствия данных и распределения Вейбулла. В случае проверки гипотез об экспоненциальности параметр формы γ полагается равным единице – такое значении указано в алгоритме 6.8 по умолчанию. Любое иное положительное значение интерпретируется как необходимость оценивать параметр в показателе распределения Вейбулла. Кроме того, можно заметить, что для каждого проверяемого уровня сохраняются P-значения для проверки гипотез о вейбулловости и паретовости. В разделе 6.4.4 эта особенность реализации алгоритма будет продемонстрирована отдельно.

Для получения результатов, представленных на рисунках 6.30 и 6.31, использован модифицированный `PoT`-метод в восходящем жадном варианте – в качестве порогового уровня используется первый, для которого сразу обе гипотезы не отвергаются. Продемонстрировано высокое (в том числе, и визуальное) соответствие между экспериментальными данными и подогнанными распределениями (см. гистограммы и аппроксимирующие кривые на верхних графиках). На нижнем графике на том же рисунке продемонстрировано, что полученное пороговое значение превышается относительно небольшое число раз за весь период наблюдений, таким образом, данные пики могут рассматриваться как потенциально экстремальные.

Для Потсдама критический уровень составляет 30,2 мм, шаг изменения уровня – 0,01 мм, статистическая значимость критерия χ^2 $\alpha = 0{,}01$, при проверке на экспоненциальность $P_{\text{знач}} = 0{,}07$ (параметр оценивается значением 0,002), для обобщенного Парето $P_{\text{знач}} = 0{,}29$ (параметры: 0,226, 9,196 и 30,2). Для Элисты критический уровень составляет 26,5 мм, при проверке на экспоненциальность $P_{\text{знач}} = 0{,}84$ (параметр оценивается значением 0,002), для обобщенного Парето $P_{\text{знач}} = 0{,}6$ (параметры: 0,038, 9,009 и 36,5).

6.4.2 Статистическая проверка гипотез об экстремальности наблюдений в скользящем режиме

Задача определения экстремальных объемов осадков и их интенсивностей является важной, однако не существует единого критерия для их определения, поскольку экстремальные для засушливых регионов осадки являются обычными для умеренного или тропического климата. В

данном разделе будет предложен статистический алгоритм выявления аномальных наблюдений в данных, который основан на сравнении текущего выпавшего объема с рядом из нескольких других объемов в том же географическом месте. При этом для построения статистических критериев будет использовано только предположение о том, что объемы осадков или их интенсивности имет классическое или обобщенное гамма-распределение. При этом выше уже была продемонстрирована высокая согласованность этих моделей и реальных данных.

Задача построения статистического теста для проверки того, является ли конкретный объем аномальным относительно других, формализуется следующим образом. Пусть задана выборка $X_1, X_2, \ldots, X_M$ из $M \geqslant 2$ положительных наблюдений. Предположим, что $X_1 \geqslant X_i$, $i = 2, \ldots, M$, то есть является «экстремальным» относительно остальных наблюдений. Тогда возможны два случая: X_1 является обычным наблюдением и его «экстремальный» характер в указанном выше смысле объясняется сугубо стохастическими причинами, или X_1 является аномальным выбросом, появившимся из-за влияния некоторых внешних факторов.

Основываясь на результатах раздела 6.3, будем считать, что каждое из указанных наблюдений X_i, $i = \overline{1, M}$ имеет обобщенное гамма-распределение с некоторыми параметрами $r > 0$, $\gamma > 0$ и $\mu > 0$. Для удобства в дальнейшем будем обозначать соответствующие независимые случайные величины (с. в.) как $\overline{G}^{(1)}_{r,\gamma,\mu}, \overline{G}^{(2)}_{r,\gamma,\mu}, \ldots, \overline{G}^{(M)}_{r,\gamma,\mu}$ для всех допустимых $M \in \mathbb{N}$. Аналогичная последовательность независимых гамма-распределенных (то есть $\gamma = 1$) величин будет обозначаться как $G^{(1)}_{r,\mu}, G^{(2)}_{r,\mu}, \ldots, G^{(M)}_{r,\mu}$. Очевидным образом из условия независимости последовательности с. в. $\overline{G}^{(1)}_{r,\gamma,\mu}, \overline{G}^{(2)}_{r,\gamma,\mu}, \ldots, \overline{G}^{(M)}_{r,\gamma,\mu}$ (то есть однородности составленной из них выборки) следует однородность выборки $\left(\overline{G}^{(1)}_{r,\gamma,\mu}\right)^{\gamma}, \left(\overline{G}^{(2)}_{r,\gamma,\mu}\right)^{\gamma}, \ldots, \left(\overline{G}^{(M)}_{r,\gamma,\mu}\right)^{\gamma}$. Рассмотрим относительный вклад случайной величины $\left(\overline{G}^{(1)}_{r,\gamma,\mu}\right)^{\gamma}$ в общую сумму $\left(\overline{G}^{(1)}_{r,\gamma,\mu}\right)^{\gamma} + \left(\overline{G}^{(2)}_{r,\gamma,\mu}\right)^{\gamma} + \ldots + \left(\overline{G}^{(M)}_{r,\gamma,\mu}\right)^{\gamma}$:

$$R = \frac{\left(\overline{G}^{(1)}_{r,\gamma,\mu}\right)^{\gamma}}{\left(\overline{G}^{(1)}_{r,\gamma,\mu}\right)^{\gamma} + \left(\overline{G}^{(2)}_{r,\gamma,\mu}\right)^{\gamma} + \ldots + \left(\overline{G}^{(m)}_{r,\gamma,\mu}\right)^{\gamma}}.$$

Из соотношения (1.11) следует, что

$$R \stackrel{d}{=} \frac{G^{(1)}_{r,\mu}}{G^{(1)}_{r,\mu} + G^{(2)}_{r,\mu} + \ldots + G^{(M)}_{r,\mu}} \stackrel{d}{=} \frac{G^{(1)}_{r,1}}{G^{(1)}_{r,1} + G^{(2)}_{r,1} + \ldots + G^{(M)}_{r,1}}.$$

Таким образом, случайная величина R характеризует вклад одного объема осадков за достаточно длительный временной интервал в суммарный объем за M «дождливых» периодов. Отметим, что

$$R = \left(1 + \frac{1}{G^{(1)}_{r,\mu}}(G^{(2)}_{r,\mu} + \ldots + G^{(m)}_{r,\mu})\right)^{-1} \stackrel{d}{=} \left(1 + \frac{G_{(M-1)r,\mu}}{G_{r,\mu}}\right)^{-1},$$

где гамма-распределенные случайные величины в правой части независимы. Обозначая $k = (M-1)r$, имеем:

$$\frac{G_{k,\mu}}{G_{r,\mu}} = \frac{k}{r} \cdot \left(\frac{r}{k} \cdot \frac{G_{k,\mu}}{G_{r,\mu}}\right) \stackrel{d}{=} \frac{k}{r} \cdot Q_{k,r},$$

где последнее равенство вытекает из соотношения $Q_{k,r} \stackrel{d}{=} rG_{k,\mu}(kG_{r,\mu})^{-1}$ при независимых с. в.ах в правой части [275], при этом случайная величина $Q_{k,r}$ имеет распределение Снедекора-Фишера с параметрами $k > 0$, $r > 0$ относительно меры Лебега

$$f_{k,r}(x) = \frac{\Gamma(k+r)}{\Gamma(k)\Gamma(r)}\left(\frac{k}{r}\right)^k \frac{x^{k-1}}{(1+\frac{k}{r}x)^{k+r}}, \quad x \geqslant 0.$$

Тогда плотность случайной величины R представима в виде

$$p(x;k,r) = \frac{\Gamma(k+r)}{\Gamma(r)\Gamma(k)}(1-x)^{k-1}x^{r-1}, \quad 0 \leqslant x \leqslant 1,$$

то есть она имеет бета-распределение с параметрами k и r. Поэтому для статистической проверки однородности выборки из M независимых наблюдений, имеющих одинаковое обобщенное гамма-распределение, может быть использована следущая процедура.

Предложение 6.4. *Пусть $V_1, \ldots, V_M$ – суммарные объемы осадков за M «дождливых» периодов, и, кроме того, $V_1 \geqslant V_j$ для всех допустимых $j \geqslant 2$. Для проверки гипотезы H_0: «объем осадков V_1 не является аномально большим относительно $V_1 + \ldots + V_M$» может быть использована статистика*

$$SR = \frac{V_1^\gamma}{V_1^\gamma + \ldots + V_M^\gamma},$$

которая в случае ее справедливости имеет бета-распределение с параметрами $k = (M-1)r$ и r. В случае, если $SR > \beta_{k,r}(1-\alpha)$, где $\beta_{k,r}(1-\alpha)$ – квантиль уровня $(1-\alpha)$, $\alpha \in (0,1)$, соответствующего бета-распределения, гипотеза H_0 отвергается, а объем V_1 должен быть признан экстремально большим. Уровень значимости данного критерия равен α.

Рассмотрим случайную величину R_{SF}, связанную с R соотношением

$$R_{SF} = \frac{(M-1)R}{1-R} = \frac{(M-1)\left(\overline{G}^{(1)}_{r,\gamma,\mu}\right)^{\gamma}}{\left(\overline{G}^{(2)}_{r,\gamma,\mu}\right)^{\gamma} + \ldots + \left(\overline{G}^{(M)}_{r,\gamma,\mu}\right)^{\gamma}} \stackrel{d}{=}$$
$$\stackrel{d}{=} \frac{(M-1)G^{(1)}_{r,\mu}}{G^{(2)}_{r,\mu} + \ldots + G^{(m)}_{r,\mu}} \stackrel{d}{=} \frac{k}{r}\frac{G_{r,\mu}}{G_{k,\mu}} \stackrel{d}{=} Q_{r,k}.$$

Таким образом, описанный выше тест на однородность может быть основан и на статистике с распределением Снедекора-Фишера с параметрами r и $k = (M-1)r$.

Предложение 6.5. *Пусть $V_1, \ldots, V_M$ – суммарные объемы осадков за M «дождливых» периодов, и, кроме того, $V_1 \geqslant V_j$ для всех допустимых $j \geqslant 2$. Для проверки гипотезы H_0: «объем осадков V_1 не является аномально большим относительно $V_2 + \ldots + V_M$» может быть использована статистика*

$$SR_{SF} = \frac{(M-1)V_1^{\gamma}}{V_2^{\gamma} + \ldots + V_M^{\gamma}},$$

которая в случае ее справедливости имеет распределение Снедекора-Фишера с параметрами r и $k = (M-1)r$. В случае, если $SR_{SF} > q_{r,k}(1-\alpha)$, где $q_{r,k}(1-\alpha)$ – квантиль уровня $(1-\alpha)$, $\alpha \in (0,1)$, соответствующего распределения Снедекора-Фишера, гипотеза H_0 отвергается, а объем V_1 должен быть признан экстремально большим. Уровень значимости данного критерия равен α.

Данная процедура может быть модифицирована для проверки экстремальности нескольких объемов с произвольными номерами относительно остальных. Действительно, рассмотрим некоторое число $l \in \mathbb{N}$, $1 \leqslant l < M$, и некоторую подпоследовательность номеров $i_1, i_2, \ldots, i_l \subset [1, M]$. Обозначим

$$T_l^{\gamma} = V_{i_1}^{\gamma} + V_{i_2}^{\gamma} + \ldots + V_{i_l}^{\gamma}, \quad T^{\gamma} = V_1^{\gamma} + V_2^{\gamma} + \ldots + V_M^{\gamma}.$$

Предложение 6.6. *Пусть $V_1, \ldots, V_M$ – суммарные объемы осадков за M «дождливых» периодов. Для проверки гипотезы H_0: «объемы осадков $V_{i_1}, V_{i_2}, \ldots, V_{i_l}$ не являются аномально большим относительно $V_1 + \ldots + V_M$» может быть использована статистика*

$$SR_{GG} = \frac{(M-l)T_l^{\gamma}}{l(T^{\gamma} - T_l^{\gamma})}, \tag{6.13}$$

которая в случае ее справедливости имеет распределение Снедекора-Фишера с параметрами lr *и* $(M-l)r$*. В случае, если* $SR_{GG} > q_{lr,(M-1)r}(1-\alpha)$*, где* $q_{lr,(M-1)r}(1-\alpha)$ *– квантиль уровня* $(1-\alpha)$*,* $\alpha \in (0,1)$*, соответствующего распределения Снедекора-Фишера, гипотеза* H_0 *отвергается, а суммарный вклад величин* $V_{i_1}, V_{i_2}, \ldots, V_{i_l}$ *должен быть признан экстремально большим. Уровень значимости данного критерия равен* α*.*

Описанная процедура может быть дополнительно модифицирована за счет применения метода скользящего окна. Задавая его ширину равной $m \leqslant M$ и сдвигая каждый раз на один элемент в направлении астрономического времени, с помощью статистики SR_{GG} 6.13, полагая в ней $l = 1$, можно последовательно проверить экстремальность каждого объема относительно остальных в описанном выше смысле. Данная процедура может быть полезна в ситуации, когда в одно окно попадают достаточно близкие по абсолютной величине объемы, мало отличающиеся от максимального. Таким образом, в рамках данного подхода каждый элемент (начиная с номера m и до $M - m + 1$) проверяется в точности m раз. Тогда каждое наблюдение считается:

– абсолютно экстремальным (в дальнейшем будем обозначать как `abs`), если оказывается аномальным во всех m случаях;

– промежуточным экстремумом (`int`), если признается аномальным более чем в половине случаев (то есть не меньше чем на $\lceil \frac{m}{2} \rceil$ положениях окна);

– относительно экстремальным (`rel`), если оказывается аномальным хотя бы один раз, но не более, чем в половине случаев;

– стандартным, если не было распознано как экстремальное ни на одном из положений окна.

В алгоритме 6.9 представлена реализация данной процедуры. Экстремумы с индексами GG и без них отличаются значением параметра γ, используемом в статистике (6.13). В первом случае оно определяется как оценка соответствующего параметра обобщенного гамма-распределения для объемов (см. алгоритм 6.6), а во втором предполагается, что распределение объемов описывается классическим гамма-распределением, поэтому величина $\gamma \equiv 1$.

Для установления соответствия между астрономическим временем и номерами наблюдений в «дождливых» периодах используется функция `Days2Observs`, которая определяет их среднюю продолжительности на

Алгоритм 6.9. Статистическая проверка аномальности наблюдений

```
1: function GGExtremes(Data, Days=[30 90 180 365], α=0.01)
2:     // Преобразование размера окна в днях к наблюдениям за периоды
3:     Windows← Days2Observs(Days); // Размеры скользящего окна
4:     Vols ← Volumes(Data);     // Объемы за «дождливые» периоды
5:     γ ← GGApprox(Vols, L^2, α);                    // См. алгоритм 6.6
6:     i ← 1;
7:     for all (Windows) do
8:         m ← Windows_i;
9:         for j=1, M − m + 1 do
10:            // Решения по всем элементам из окна на основе (6.13)
11:            ExtrInd_{j:j+m−1} ← SR_GG(Vols_{j:j+m−1}, γ, l=1);
12:        // Классификация всех наблюдений
13:        [abs, abs_GG, int, int_GG, rel, rel_GG] ← Identification(ExtrInd);
14:        PlotExtr(abs, abs_GG, int, int_GG, rel, rel_GG); // Визуализация
15:    return γ;
```

основе отрицательной биномиальной модели (см. раздел 6.3):

$$E = r_{wet} \cdot \frac{1 - p_{wet}}{p_{wet}} + r_{dry} \cdot \frac{1 - p_{dry}}{p_{dry}},$$

где величины с индексами wet соответствуют «дождливым», а с dry – «сухим» периодам. Умножение данной величины на размер окна в днях и округление до ближайшего целого приводят к нужному результату.

Итак, процедура идентификации аномальных значений может быть описана следующим образом (см. рисунок 6.32). К данным последовательно применяются восходящий, нисходящий (см. алгоритме 6.8) и параметрический методы поиска экстремальных значений (см. алгоритм 6.9 и рисунок 6.33). Для первых двух из них задается начальное значение уровня, а затем последовательно проверяются условия теорем Реньи и Пикандса–Балкемы–Де Хаана при заданном уровне значимости α. В результате можно сравнить решения каждого из этих методов. Стоит отметить, что восходящий и нисходящий метод могут быть использованы для любых данных, а параметрический подход ориентирован на ряд дополнительных предположений о распределениях характеристик наблюдений (классическую или обобщенную отрицательную биномиальную модели распределений длительностей изучаемых явлений), и их строгую положительность.

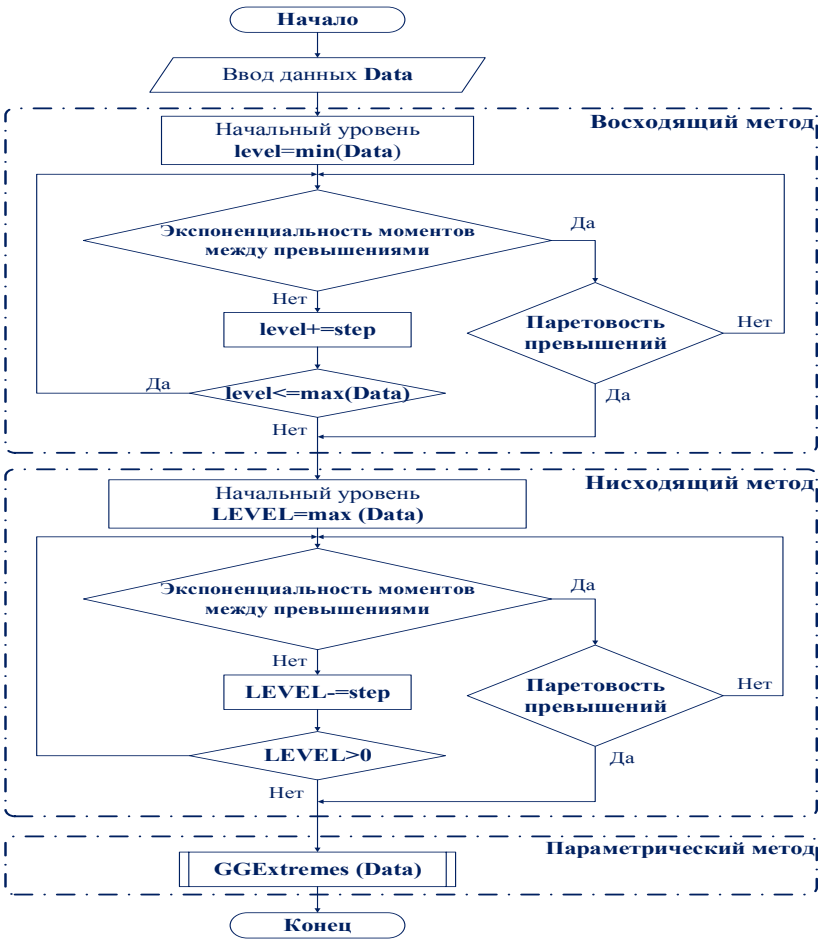

Рис. 6.32. Процедура идентификации экстремальных наблюдений

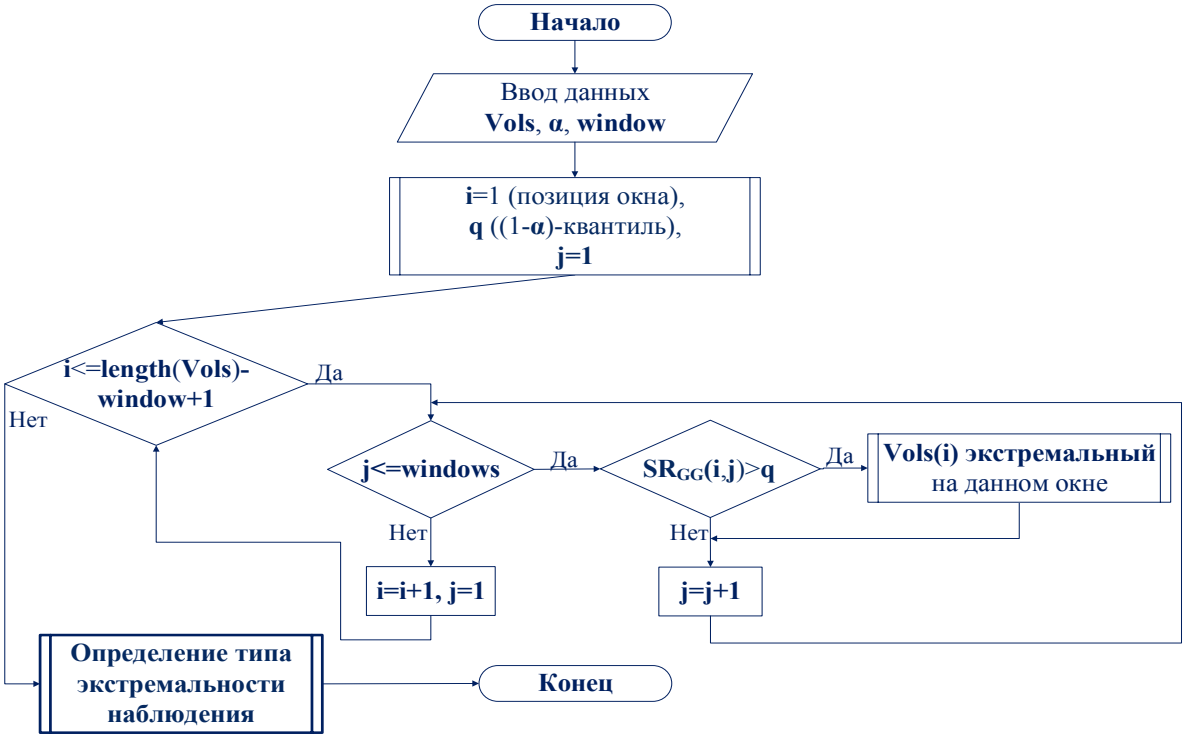

Рис. 6.33. Параметрический тест, режим скользящего окна

6.4.3 Анализ экстремальности объемов осадков

В этом разделе рассмотрим примеры применения разработанных выше методов к реальным данным. Во-первых, эмпирически было установлено, что при [293], что при небольших размерах скользящего окна, соответствующих 30 или 90 дням, тесты являются крайне чувствительными, и относят хотя бы к одному из перечисленных в предыдущем разделе типов аномальности достаточно много наблюдений. При этом к абсолютно экстремальным не относятся явные выбросы в данных (особенно подобная ситуация характерна для уровня значимости $\alpha = 0{,}01$; несколько лучше дела обстоят для $\alpha = 0{,}05$). Поэтому для демонстрации работы методов приведены окна, ширина которых соотвествует одному году.

На рисунках 6.34 и 6.35 продемонстрировано сравнение результатов различных методов для Потсдама и Элисты.

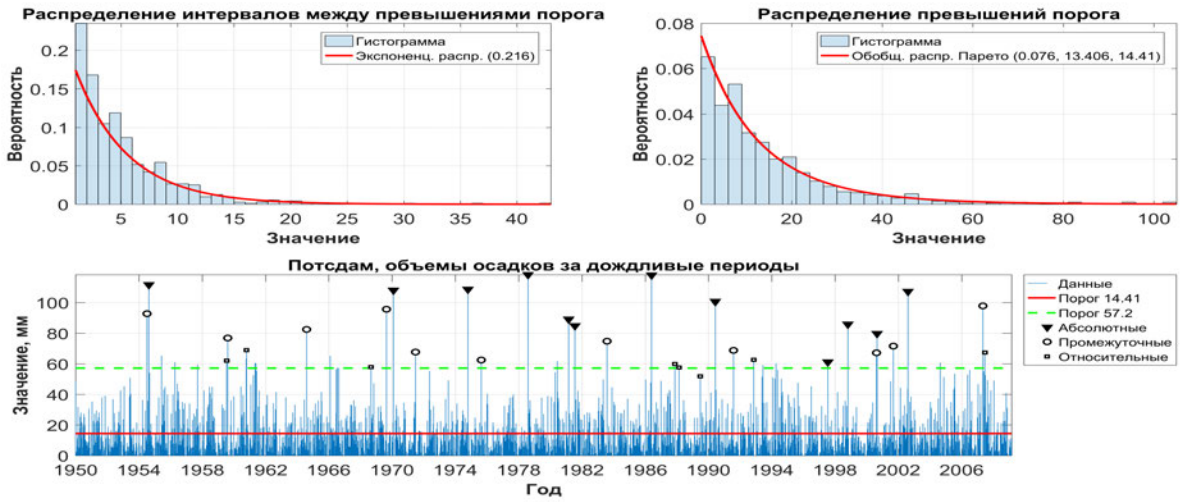

Рис. 6.34. Сравнение методов определения экстремальности, Потсдам

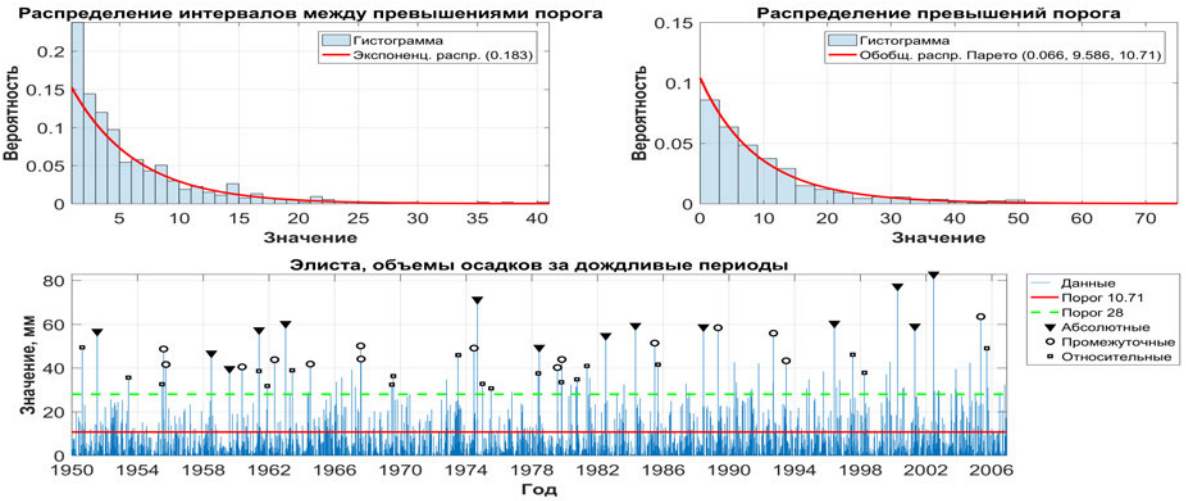

Рис. 6.35. Сравнение методов определения экстремальности, Элиста

На верхних графиках на рисунках 6.34 и 6.35 приведены данные для восходящего метода, а на нижних – пороговые уровни для восходящего и нисходящего (см. алгоритм 6.8) методов, а также результаты для параметрического подхода (см. алгоритм 6.9) для скользящего окна в 360 дней и предположения о классическом гамма-распределении для объемов. Треугольниками, кругами и квадратами обозначены абсолютные, промежуточные и относительные экстремумы, соответственно.

Для объемов осадков за «дождливые» периоды в Потсдаме верхний критический уровень (получен нисходящим методом) составляет 57,2 мм (шаг изменения – 0,01 мм). Для экспоненциального распределения $P_{\text{знач}} = 0{,}1$ (параметр оценивается значением 0,014), для обобщенного Парето $P_{\text{знач}} = 0{,}03$ (параметры: 0,097, 16,95 и 57,2). Нижний критический уровень (получен восходящим методом) составляет 14,41 мм. Для экспоненциального распределения $P_{\text{знач}} = 0{,}058$ (параметр оценивается значением 0,216), для обобщенного Парето $P_{\text{знач}} = 0{,}29$ (параметры – 0,076; 13,406 и 14,41).

Для объемов осадков за «дождливые» периоды в Элисте верхний критический уровень (получен нисходящим методом) составляет 28 мм. Для экспоненциального распределения $p_{\text{знач}} = 0{,}082$ (параметр оценивается значением 0,029), для обобщенного Парето $p_{\text{знач}} = 0{,}44$ (параметры: $-0{,}095$, 13,66 и 28). Нижний критический уровень (получен восходящим методом) составляет 10,71 мм. Для экспоненциального распределения $p_{\text{знач}} = 0{,}062$ (параметр оценивается значением 0,183), для обобщенного Парето $p_{\text{знач}} = 0{,}21$ (параметры – 0,066, 9,586 и 10,71).

Очевидно, что обеих ситуациях для накопленных данных восходящий метод устанавливает критическую планку слишком низко – и в рассмотрении оказывается избыточное количество пиков. Параметрический метод выделяет порядка 12–14 объемов за период наблюдений почти в 60 лет, которые могут рассматриваться как аномальные. Очевидно, что такая методология позволяет более аккуратно анализировать реальные данные, однако, как было отмечено выше, это достигается за счет дополнительного использования разработанных вероятностно-статистических моделей для метеорологических явлений.

Сравним работу методов на основе классического и обобщенного гамма-распределений (см. алгоритм 6.9) для Потсдама и Элисты. Величина $\alpha = 0{,}01$ выбрана в качестве уровня значимости, рассматриваются два типа окон – относительно короткое (90 дней) и достаточно большое (365 дней). Результаты представлены на рисунках 6.36–6.39.

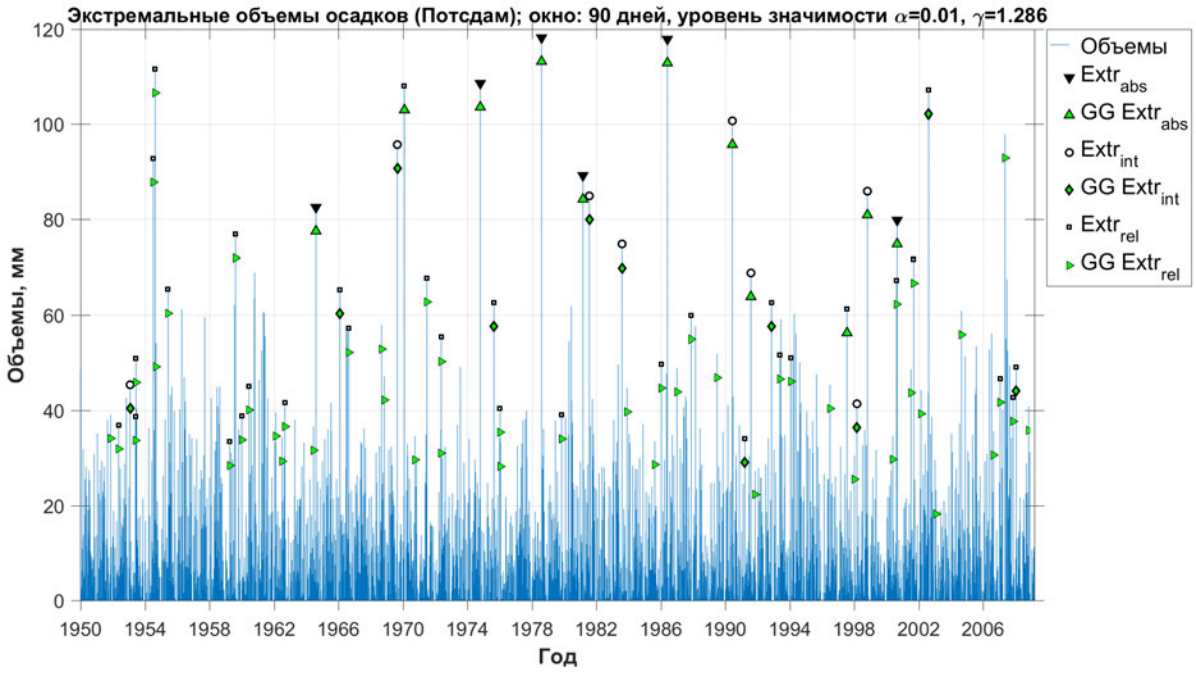

Рис. 6.36. Аномальные объемы осадков, окно 90 дней (Потсдам)

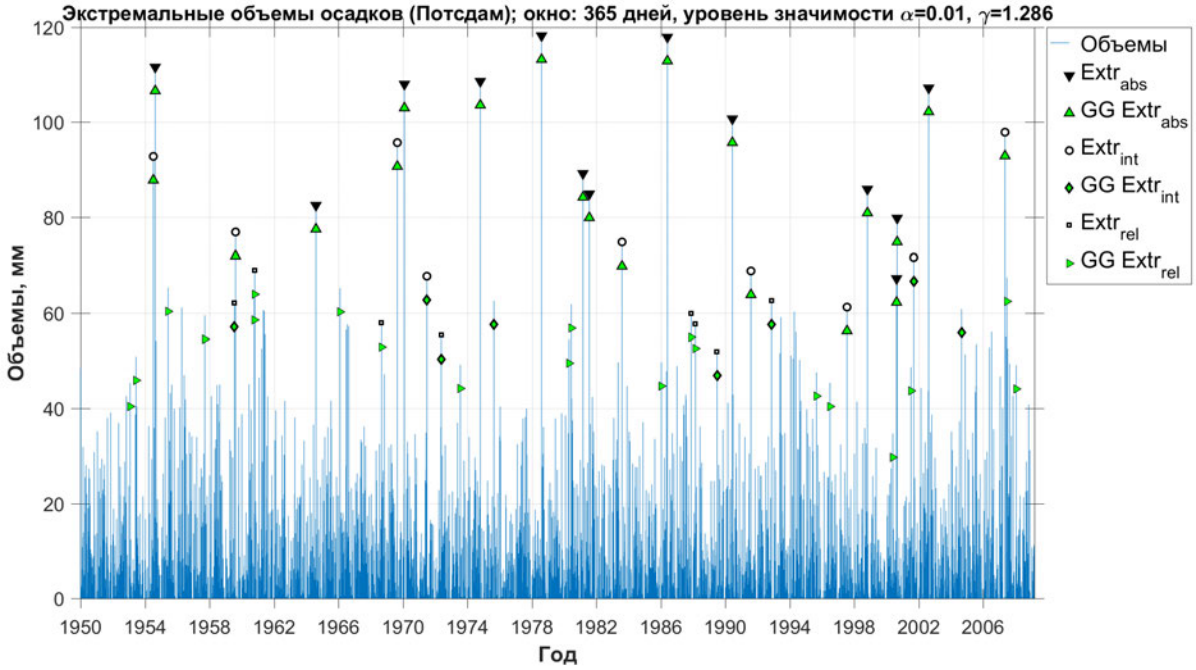

Рис. 6.37. Аномальные объемы осадков, окно 365 дней (Потсдам)

Треугольниками, кругами и квадратами, по-прежнему, обозначены абсолютные, промежуточные и относительные экстремумы, соответственно, для статистики SR_{GG} (6.13) с $\gamma = 1$. Для результатов обобщенного гамма-теста использованы следующие маркеры для аналогичных величин: зеленые перевернутые треугольники, ромбы и развернутые вправо треугольники, соответственно. В этом случае параметр γ имеет нетривальное значение в обоих случаях, а именно: $\gamma = 1{,}286$ для

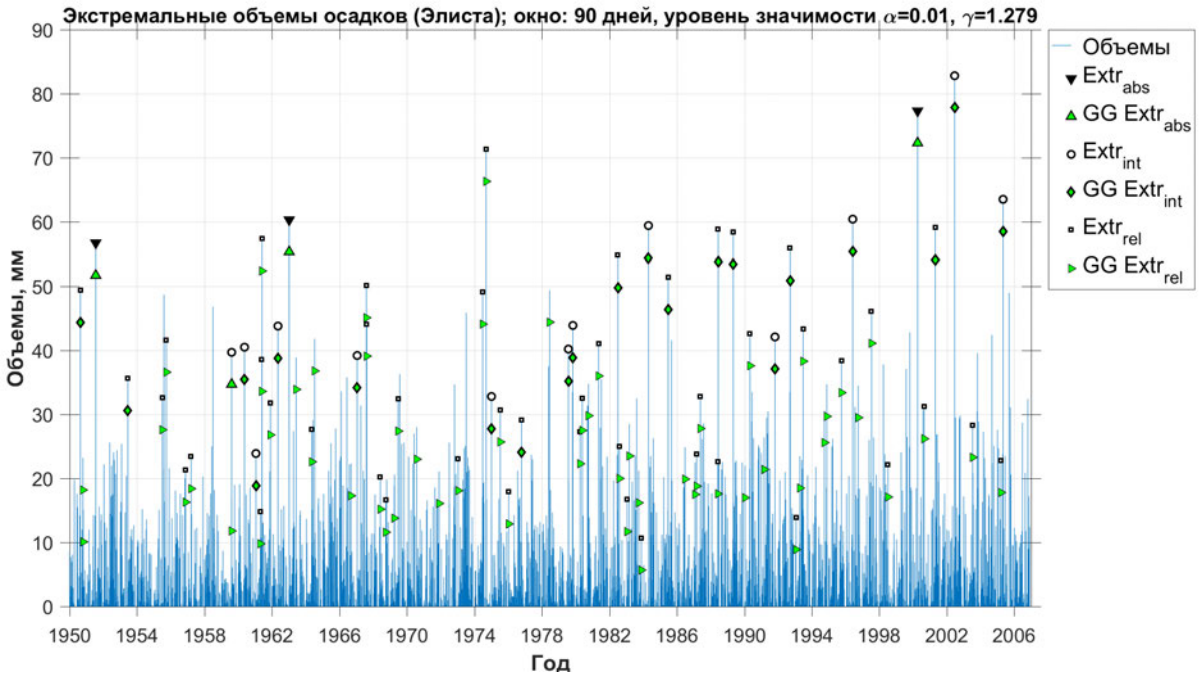

Рис. 6.38. Аномальные объемы осадков, окно 90 дней (Элиста)

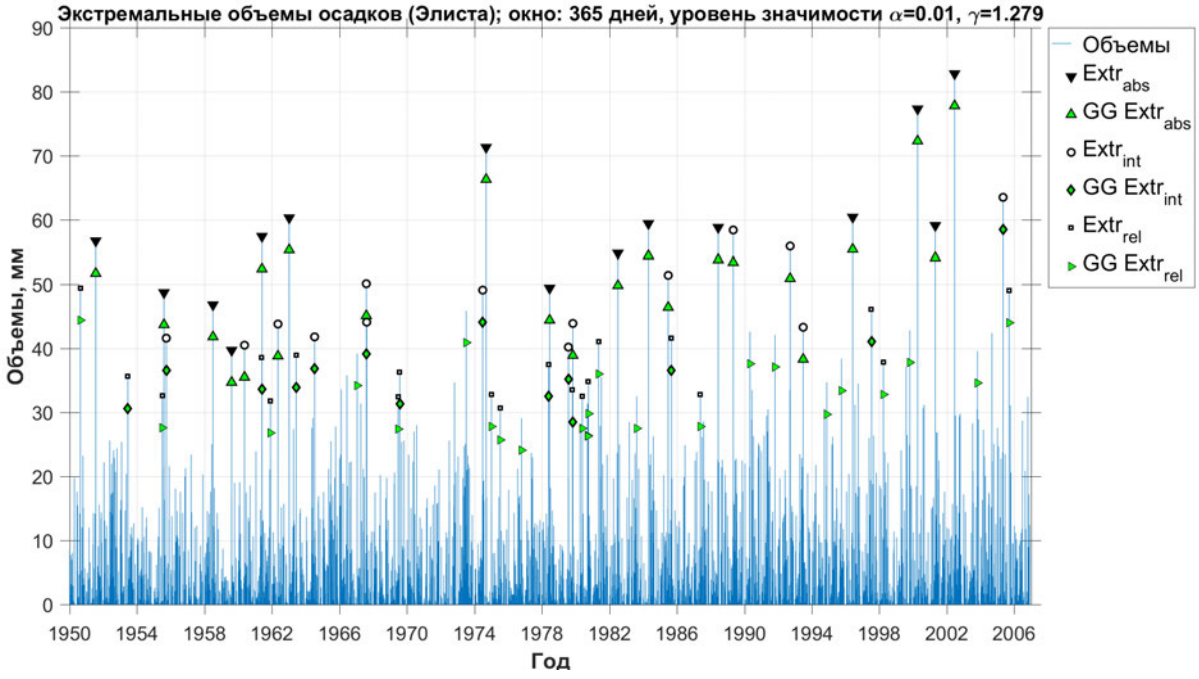

Рис. 6.39. Аномальные объемы осадков, окно 365 дней (Элиста)

Потсдама и $\gamma = 1{,}279$ для Элисты. Несмотря на то,что результаты обоих тестов можно считать достаточно близкими, в случае использования обобщенного гамма-распределения большее число явных выбросов (которые могут и не быть максимальными по сравнению вообще со всем периодом наблюдений) размечены как абсолютно экстремальные. При этом их количество не увеличивается так же сильно, как в результате применения `PoT` метода, поэтому данное обстоятельство может рассмат-

риваться как достоинство, а не недостаток метода. Аналогичные выводы остаются справделивыми и для других размеров окон.

6.4.4 Анализ экстремальности интенсивностей

Всюду выше рассматривались только объемы осадков за «дождливые» периоды, но не их интенсивности. При этом именно последние, определяемые как отношение суммарного объема осадков за период к его продолжительности, являются важными величинами, экстремальность которых может приводить к наводнениям, оползням и селевым потокам. Известны работы (см., например, статью [331]), в которых приводится теоретическое обоснование корректности использования гамма-моделей для интенсивностей осадков. Таким образом, разработанные для объемов статистические методы идентификации экстремальности могут быть применены и в данной ситуации. Процедура анализа интенсивностей основана на использовании созданных решений с необходимыми настройками параметров и представлена в алгоритме 6.10.

Алгоритм 6.10. Статистический анализ интенсивностей

```
1: function IntensitiesExtremes(Data, Days=[30 90 180 365], α=0.01)
2:     Ints ← Intensities(Data);
3:     PoT(Ints);                          // См. алгоритм 6.8
4:     γ ← GGExtremes(Ints);               // См. алгоритм 6.9
5:     PoT(Ints, γ);
6:     return ;
```

В данном случае используется проверка и экспонциальности, и вейбулловости моментов между превышениями порогового значения, а также осуществляется вывод P-значения для каждого положения уровня. Результаты для Потсдама и Элисты продемонстрированы на рисунках 6.40 и 6.41. Для выбранного уровня значимости $\alpha = 0{,}01$ соответствующая гипотеза не отвергается для всех пороговых уровней, расположенных справа от красной линии на верхних графиках. Нижние графики на рисунках 6.40 и 6.41 содержат оцененные значения параметров распределения Вейбулла.

На рисунках 6.42 и 6.43 приведены результаты сравнения параметрического теста на основе обобщенного гамма-распределения (см. алгоритм 6.9) и нисходящего `PoT`-метода (см. алгоритм 6.8), причем для последнего изображены сразу несколько пороговых уровней.

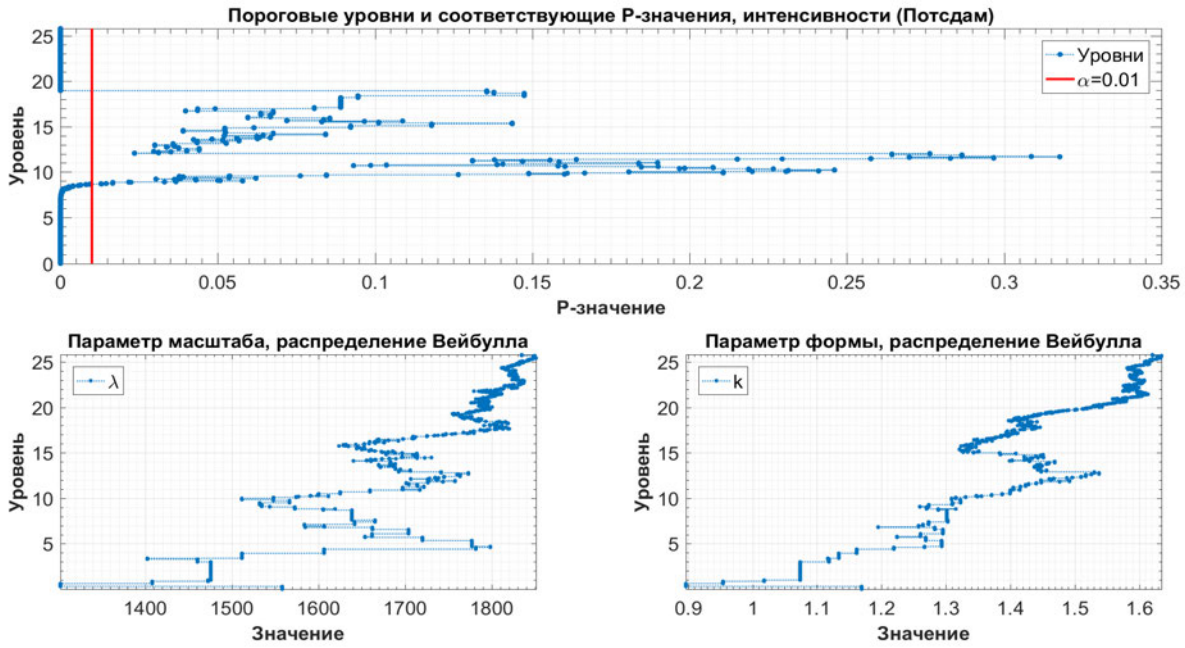

Рис. 6.40. Пороговые уровни и соответствующие P-значения, Потсдам

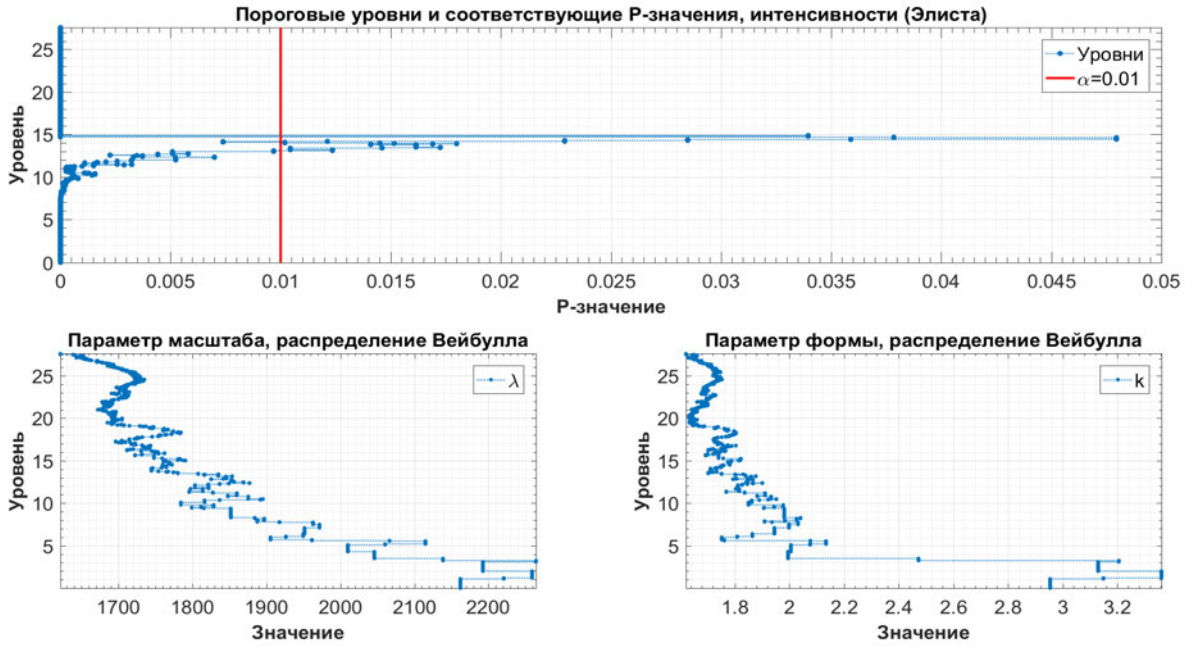

Рис. 6.41. Пороговые уровни и соответствующие P-значения, Элиста

А именно, пороговые значения с индексами:

– `low`: соотвествуют минимальному уровню, на котором гипотеза об экспоненциальности или вейбулловости не отвергается (для распределения Вейбулла – самая нижняя точка на верхних графиках на рисунках 6.40 and 6.41, которая лежит справа от красной линии);

– `maxval`: соотвествуют максимальному P-значению (самая правая точка на упомянутых графиках);

– `high`: соотвествуют самому высокому положению порога, при котором гипотеза о распределении не отвергается (наивысшая точка справа от красной линии на графиках).

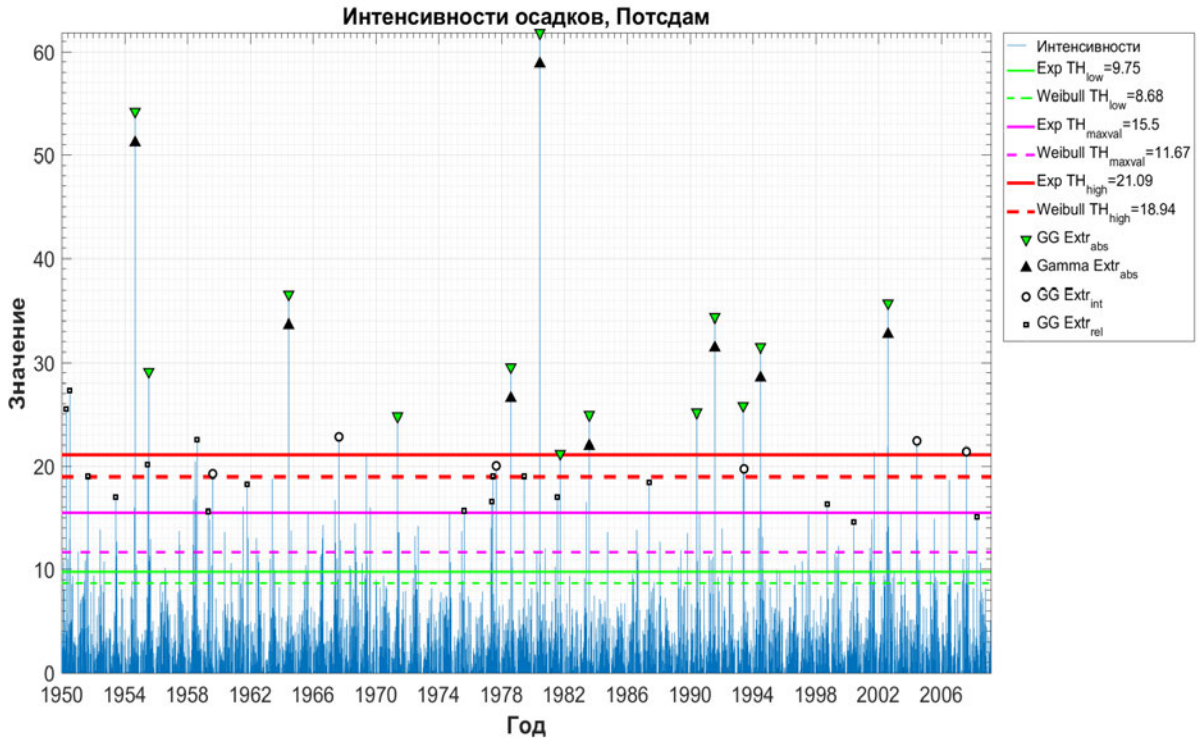

Рис. 6.42. Аномальные интенсивности осадков, окно 365 дней (Потсдам)

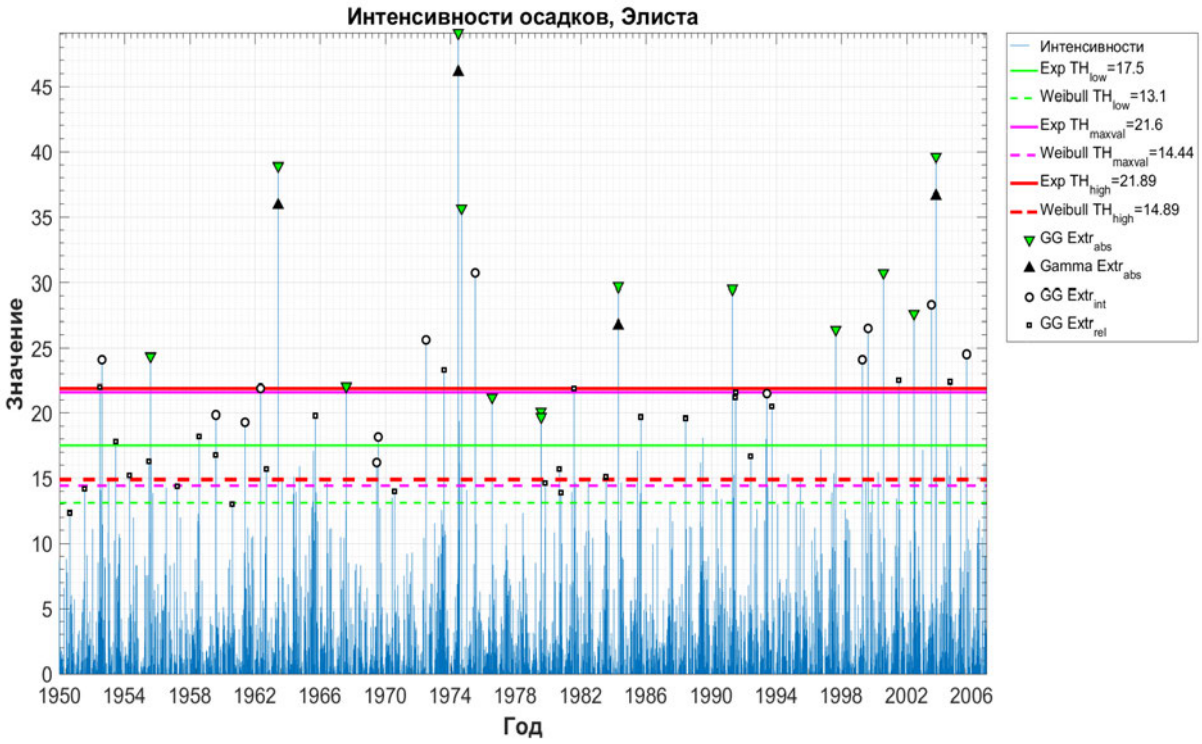

Рис. 6.43. Аномальные интенсивности осадков, окно 365 дней (Элиста)

На рисунках 6.42 и 6.43 нанесены маркеры для всех типов экстремальности для обобщенного гамма-теста (зеленые треугольники, круги и квадраты), и абсолютно аномальные – для теста на основе классического варианта распределений (черные треугольники). Для обобщенного теста значения параметра γ составляют 1,0775 для Потсдама 1,1257 для Элисты, соответственно.

Для Потсдама результаты обоих параметрических тестов являются близкими и представляются достаточно корректными. При этом наибольший уровень отсечения для `PoT`-метода получен с помощью экспоненциального порога с индексом `high`, а «лучший» порог распределения Вейбулла уровня (при этом с индексом `maxval`) ниже на 2,15 единиц, поэтому между ними расположены несколько промежуточных экстремумов параметрического теста.

Для Элисты решения экспоненциального и вейбулловского методов отличаются всего на 0,29 – и в обеих ситуациях это оказывается порог с индексом `maxval`. При этом статистический тест на основе классического гамма-распределения отнес к абсолютно аномальным всего 4 выброса в данных.

6.4.5 Темперированное распределение Снедекора-Фишера как модель экстремальных объемов

В данном разделе с использованием модели отрицательного биномиального распределения для продолжительностей «дождливых» периодов и полученных в разделе 1.3 результатов для темперированного распределения Снедекора-Фишера, которое является асимптотическим для экстремальных наблюдений в рамках сделанных предположений (см. теорему 1.2), будет продемонстрирован подход к определению экстремальных суточных объемов как превышающих квантили выбранных уровней данного распределения. Для этого будут предложены несколько методов оценивания неизвестных параметров, проведено их сравнение. Кроме того, полученные результаты будут сопоставлены с известными в метеорологии подходами [443] на одних и тех же реальных данных.

Сначала рассмотрим процедуры для статистического оценивания неизвестных параметров r, μ и λ асимптотического распределения для экстремальных наблюдений в выборке, представляющее собой распределение положительной степени случайной величины с распределением Снедекора-Фишера (1.26). Пусть последовательность $\{X_{i,j}\}$, $i = \overline{1,m}$,

$j = \overline{1, m_i}$, представляет объем осадков в j-й день «дождливого» периода с номером i. Рассмотрим соответствующие порядковые статистики $X^*_{(1)}, \ldots, X^*_{(m)}$, построенные по выборке $X^*_1, \ldots, X^*_m$, где каждая величина определяется как $X^*_k = \max\{X_{k,1}, \ldots, X_{k,m_k}\}$. Воспользуемся методом квантилей для оценивания неизвестных параметров.

Выпишем в явном виде квантиль $x(\alpha; \lambda, \mu, r)$ функции распределения $F_{\lambda,\mu,r}(x)$ (1.26) уровня $\alpha \in (0, 1)$, как решение уравнения $F_{\lambda,\mu,r}(x) = \alpha$ относительно x. Имеем:

$$x(\alpha; \lambda, \mu, r) = \left(\frac{\alpha^{1/r}}{\mu - \mu\epsilon^{1/r}} \right)^{1/\lambda}.$$

Фиксируя три произвольных числа $0 < p_1 < p_2 < p_3 < 1$, получим систему уравнений относительно неизвестных параметров r, μ, λ:

$$X^*_{([mp_k])} = \left(\frac{p_k^{1/r}}{\mu - \mu p_k^{1/r}} \right)^{1/\lambda}, \quad k = 1, 2, 3 \tag{6.14}$$

(здесь через $[\cdot]$ обозначается целая часть аргумента). Величина r может быть найдена численно:

$$r = \left(\log \frac{X^*_{([mp_1])}}{X^*_{([mp_3])}} \log \frac{p_1}{p_2} - \log \frac{X^*_{([mp_1])}}{X^*_{([mp_2])}} \log \frac{p_1}{p_3} \right) \times$$
$$\times \left(\log \frac{1 - p_3^s}{1 - p_1^s} \log \frac{X^*_{([mp_1])}}{X^*_{([mp_2])}} - \log \frac{1 - p_2^s}{1 - p_1^s} \log \frac{X^*_{([mp_1])}}{X^*_{([mp_3])}} \right)^{-1}, \tag{6.15}$$

и в дальнейшем считается известной.

Предложение 6.7. *Пусть параметр r функции распределения $F_{\lambda,\mu,r}(x)$ (1.26) оценен численно по формуле (6.15) или исходя из отрицательной биномиальной модели. Тогда решение системы (6.14) относительно параметров λ и μ имеет вид*

$$\widehat{\mu}_q = \frac{p_2^{\frac{1}{r}}}{(1 - p_2^{\frac{1}{r}})(X^*_{([mp_2])})^{\lambda}}, \tag{6.16}$$

$$\widehat{\lambda}_q = \frac{\frac{1}{r}(\log p_1 - \log p_3) + \log(1 - p_3^{\frac{1}{r}}) - \log(1 - p_1^{\frac{1}{r}})}{\log X^*_{([mp_1])} - \log X^*_{([mp_3])}}. \tag{6.17}$$

Величины p_1, p_2 and p_3 могут быть выбраны как $p_1 = \tau$, $p_2 = \frac{1}{2}$, $p_3 = 1 - \tau$ для некоторого значения $\tau \in (0, \frac{1}{4}]$.

В случае заданного значения параметра r, более точные оценки μ и λ могут быть найдены минимизацией расстояния между эмпирической и модельной функциями методом наименьших квадратов. Из теоремы Гливенко следует, что

$$\left(\frac{\mu(X^*_{(i)})^\lambda}{1+\mu(X^*_{(i)})^\lambda}\right)^r \approx \frac{i}{m}. \tag{6.18}$$

При этом предполагается, что все элементы вариационного ряда различны. Тогда для всех $i = \overline{1, m-1}$ имеют место следующие импликации:

$$(6.18) \iff \left\{\frac{\mu(X^*_{(i)})^\lambda}{1+\mu(X^*_{(i)})^\lambda} \approx \left(\frac{i}{m}\right)^{1/r}\right\} \iff$$

$$\iff \left\{\mu(X^*_{(i)})^\lambda \approx \frac{i^{1/r}}{m^{1/r}-i^{1/r}}\right\} \iff$$

$$\iff \left\{\log\mu + \lambda\log X^*_{(i)} \approx \log\frac{i^{1/r}}{m^{1/r}-i^{1/r}}\right\}.$$

Поэтому оценки $\widehat{\mu}_{LS}$ и $\widehat{\lambda}_{LS}$ могут быть найдены решением следующей задачи методом наименьших квадратов:

$$(\widehat{\mu}_{LS},\, \widehat{\lambda}_{LS}) = \arg\min_{\log\mu,\lambda} \sum_{i=1}^{m-1}\left(\log\mu + \lambda\log X^*_{(i)} - \log\frac{i^{1/r}}{m^{1/r}-i^{1/r}}\right)^2,$$

Предложение 6.8. *При известном значении параметра r оценки методом наименьших квадратов величин λ и μ имеют следующий вид:*

$$\widehat{\mu}_{LS} = \exp\left\{\frac{1}{m-1}\Big(\sum_{j=1}^{m-1}\log\frac{j^{1/r}}{m^{1/r}-j^{1/r}} - \widehat{\lambda}_{LS}\sum\nolimits_{j=1}^{m-1}\log X^*_{(j)}\Big)\right\}, \tag{6.19}$$

$$\widehat{\lambda}_{LS} = \sum_{j=1}^{m-1}\log X^*_{(j)}\left(\left(\log\frac{j^{1/r}}{m^{1/r}-j^{1/r}}\right)^{m-1} - \sum_{k=1}^{m-1}\log\frac{k^{1/r}}{m^{1/r}-k^{1/r}}\right)\times$$

$$\times\left((m-1)\sum_{j=1}^{m-1}\left(\log X^*_{(j)}\right)^2 - \left(\sum_{j=1}^{m-1}\log X^*_{(j)}\right)^2\right)^{-1}. \tag{6.20}$$

Сравним указанные методы оценивания параметров на примере реальных данных для Потсдама и Элисты. Для этого воспользуемся невязками D_q и D_{LS} между эмпирической $\widehat{F}(x)$ и аппроксимирующими функциями темперированного распределения Снедекора-Фишера $F^{(q)}_{SF}(x)$ и

$F_{SF}^{(LS)}(x)$ для квантильного подхода и метода наименьших квадратов:

$$D_q = \max_{x \in \mathbf{X}} \left| \widehat{F}(x) - F_{SF}^{(q)}(x) \right|, \quad D_{LS} = \max_{x \in \mathbf{X}} \left| \widehat{F}(x) - F_{SF}^{(LS)}(x) \right|,$$

где $\mathbf{X} = (X_1, \ldots, X_n)$ – рассматриваемая выборка. Данные величины и полученные оценки параметров приведены в таблицах 6.20 и 6.21.

Таблица 6.20. Параметры экстремального распределения, Потсдам

Минимальная длительность периода (дни)	**Объем тестовой выборки**	$\mathbf{D_q}$	$\mathbf{D_{LS}}$	$\widehat{\boldsymbol{\mu}}_{\mathbf{q}}$	$\widehat{\boldsymbol{\mu}}_{\mathbf{LS}}$	$\widehat{\boldsymbol{\lambda}}_{\mathbf{q}}$	$\widehat{\boldsymbol{\lambda}}_{\mathbf{LS}}$
1	3323	0,09	0,092	0,169	0,212	1,18	1,29
2	2066	0,045	0,065	0,0383	0,054	1,755	1,71
3	1282	0,031	0,041	0,01	0,013	2,261	2,183
4	862	0,026	0,027	0,0049	0,0045	2,449	2,524
6	384	**0,025**	**0,026**	0,0015	0,0012	2,822	2,949
8	163	0,04	0,045	0,0007	0,0005	3,174	3,255
10	73	0,041	0,042	0,0003	0,0003	3,385	3,352
15	12	0,13	0,09	0,0014	0,0009	2,67	2,973

Таблица 6.21. Параметры экстремального распределения, Элиста

Минимальная длительность периода (дни)	**Объем тестовой выборки**	$\mathbf{D_q}$	$\mathbf{D_{LS}}$	$\widehat{\boldsymbol{\mu}}_{\mathbf{q}}$	$\widehat{\boldsymbol{\mu}}_{\mathbf{LS}}$	$\widehat{\boldsymbol{\lambda}}_{\mathbf{q}}$	$\widehat{\boldsymbol{\lambda}}_{\mathbf{LS}}$
1	2937	0,06	0,06	0,361	0,349	1,053	1,263
2	1374	0,049	0,055	0,108	0,101	1,424	1,574
3	656	**0,041**	**0,045**	0,0454	0,0376	1,707	1,9
4	319	0,051	0,06	0,0234	0,0273	1,891	1,94
6	77	0,07	0,075	0,0181	0,0144	2,011	2,186
7	42	0,15	0,01	0,0197	0,0207	1,983	2,179
8	22	0,12	0,14	0,014	0,0358	2,01	1,764
10	10	0,17	0,16	0,0136	0,0375	2,163	1,802

Поскольку темперированное распределение Снедекора-Фишера является асимптотическим, поэтому производится дополнительное прореживание данных: используется минимальная продолжительность «дождливого» периода (первая колонка в таблицах), а также число соответствующих наблюдений. Из таблиц 6.20 и 6.21 следует, что наилучшие

результаты в смысле невязок получаются для периодов длительностью не менее трех дней в Элисте и шести – в Потсдаме.

На рисунках 6.44 и 6.45 продемонстрированы примеры близости эмпирического и модельных распределений, оценки которых получены описанными выше способами, для Потсдама и Элисты. Видно, что графики близки даже для минимальной возможной продолжительности «дождливого» периода (больше примеров приведено в статье [233]).

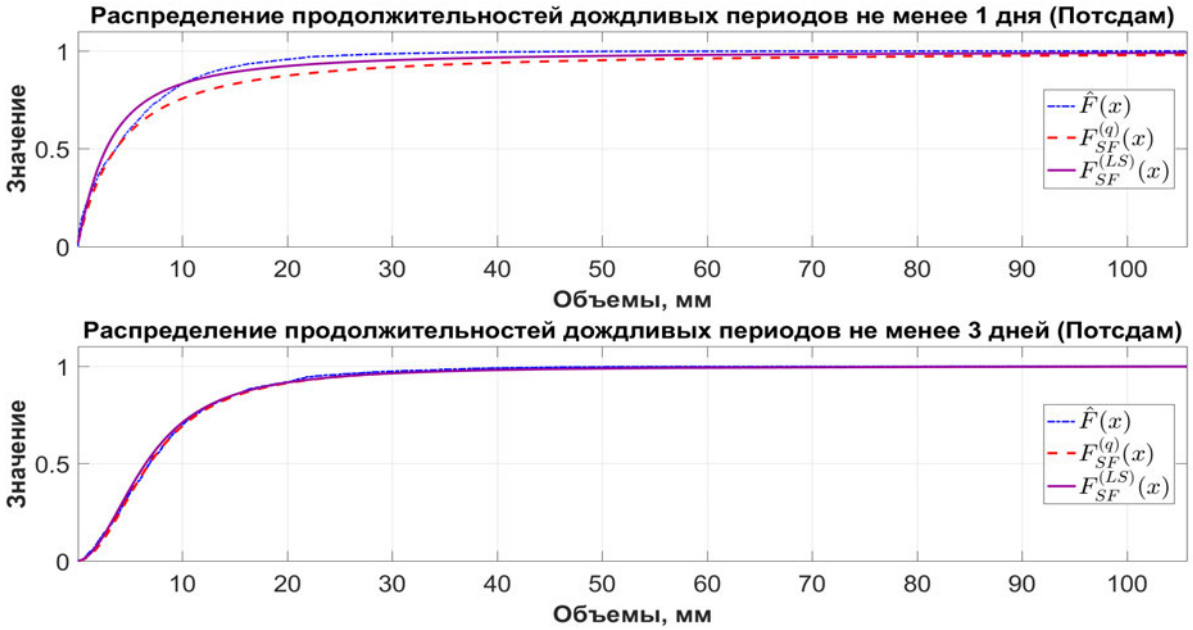

Рис. 6.44. Аппроксимация асимптотическим распределением, Потсдам

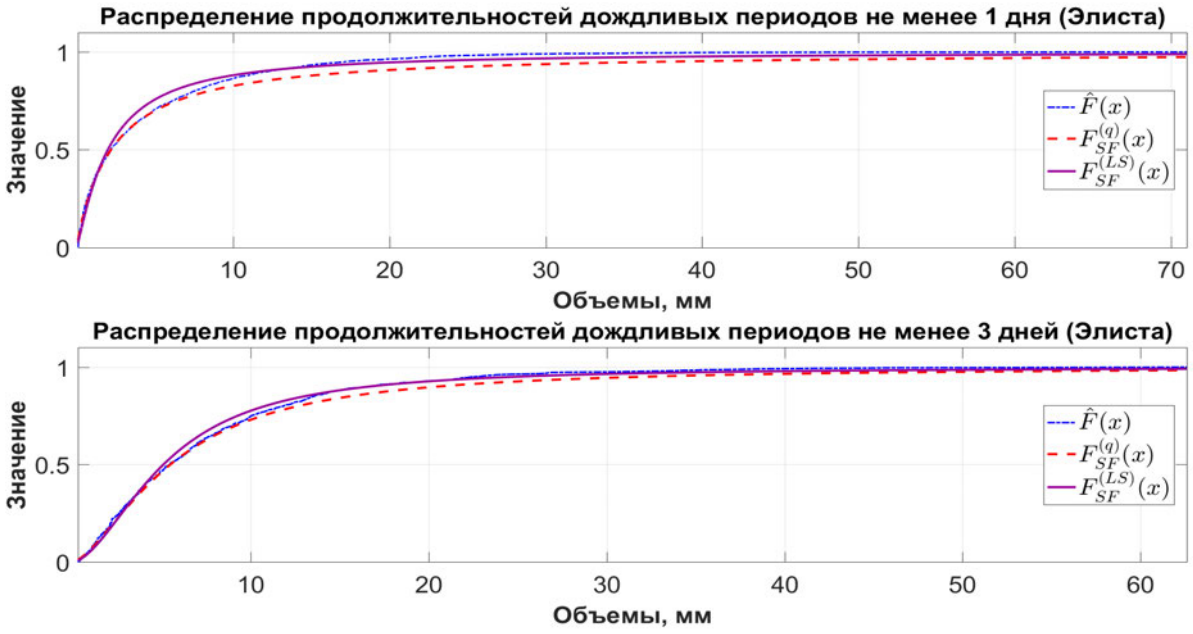

Рис. 6.45. Аппроксимация асимптотическим распределением, Элиста

Очевидно, что метод наименьших квадратов (см. формулы (6.19) и (6.20)), соответствующий функции $F_{SF}^{(LS)}(x)$ на графиках, ведет к более точным приближениям эмпирической функции распределения по сравнению с квантильным методом (см. формулы (6.16) и (6.17)) с функцией распределения $F_{SF}^{(q)}(x)$.

В прикладных задачах, имеющих дело с экстремальными значениями, зачастую предполагается, что они соответствуют одному из трех типов функций распределений. Однако данные результаты носят асимптотический характер – и могут нарушаться для выборок умеренного объема. Решение в данном случае основано на подходах работы [197] с учетом использования теоретически обоснованного в диссертации темперированного распределения Снедекора-Фишера в качестве асимптотического (см. раздел 1.3). На рисунке 6.46 представлена процедура идентификации аномальности суточных объемов осадков.

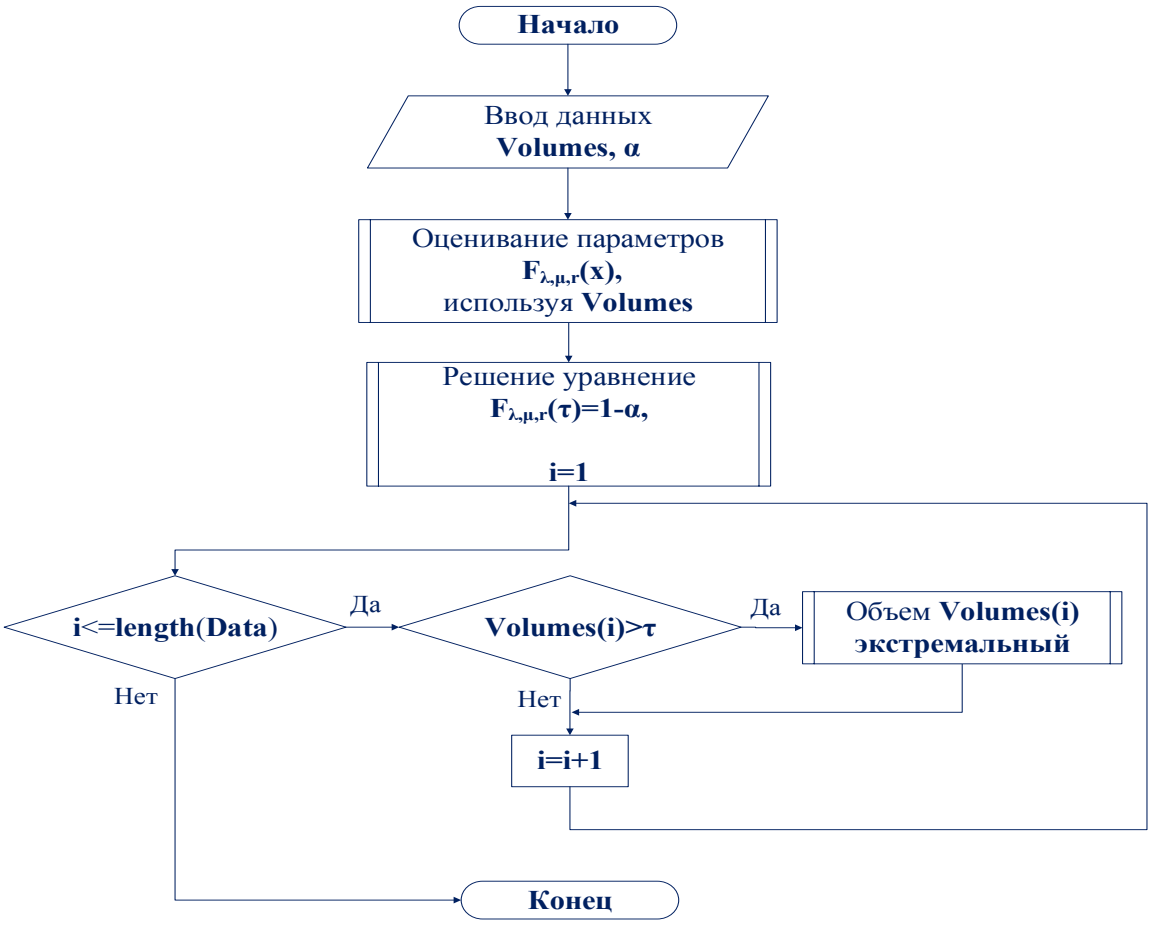

Рис. 6.46. Процедура определения аномальности объемов осадков

Блок «Оценивание параметров» более подробно рассмотрен в алгоритме 6.11, а процедура поиска пороговых уровней – в алгоритме 6.12.

Алгоритм 6.11. Оценивание параметров функции $F_{\lambda,\mu,r}(x)$ (1.26)

```
1: function SFParams(Data, Thresholds=1,10, τ = 1/4)
2:     F ← @(x, λ, μ, r) (μx^λ/(1+μx^λ))^r I(x)_{x⩾0};   // Указатель на функцию
3:     i ← 1; p ← [τ, 1/2, 1 − τ];
4:     for all (Thresholds) do
5:         Sample ← WetSample(Data, Thresholds_i);
6:         r ← rFind(Sample); // По формуле (6.15) или параметр NB
7:         // Квантильный метод, см. формулы (6.16) и (6.17)
8:         [μ̂_q, λ̂_q] ← SF^(q)(F, Sample, r, p);
9:         // МНК при заданном r, см. формулы (6.19) и (6.20)
10:        [μ̂_LS, λ̂_LS] ← SF^(LS)(F, Sample, r, p);
11:        i++;
12:    return [μ̂_q, λ̂_q, μ̂_LS, λ̂_LS];
```

В функции `WetSample` алгоритма 6.11 обеспечивается уникальность элементов вариационного ряда: если есть совпадающие наблюдения, то они заменяются весьма близкими, но различающимися, например за счет искусственного зашумления аддитивной случайной величиной ε с нормальным распределением с параметрами 0 и 0,01.

Алгоритм 6.12. Идентификация экстремальных наблюдений

```
1: function SFThresholds(Data, r, μ̂_LS, λ̂_LS, α=[0.01 0.05 0.1])
2:     F ← @(x, λ, μ, r) (μx^λ/(1+μx^λ))^r I(x)_{x⩾0};   // Указатель на функцию
3:     i ← 1;
4:     for all (α) do
5:         Threshold_{1−α_i} ← fsolve((F(x, r, μ̂_LS, λ̂_LS), 1 − α_i)); i++;
6:     PlotSFThresholds(Data, Thresholds);          // Визуализация
7:     return Thresholds;
```

Пример применения к данным в Потсдаме и Элисте рассмотрен на рисунках 6.47 и 6.48. При этом для наглядности изображения в каждом «дождливом» периоде оставлено единственное наблюдение с максимальным значением. Горизонтальные линии соответствуют квантилям уровней 0,9 (синие точки), 0,95 (пунктирная сиреневая) и 0,99 (сплошная красная) для подогнанного методом наименьших квадратов темперированного распределения Снедекора-Фишера.

За анализируемый период в Потсдаме выделяются 13 «дождливых»

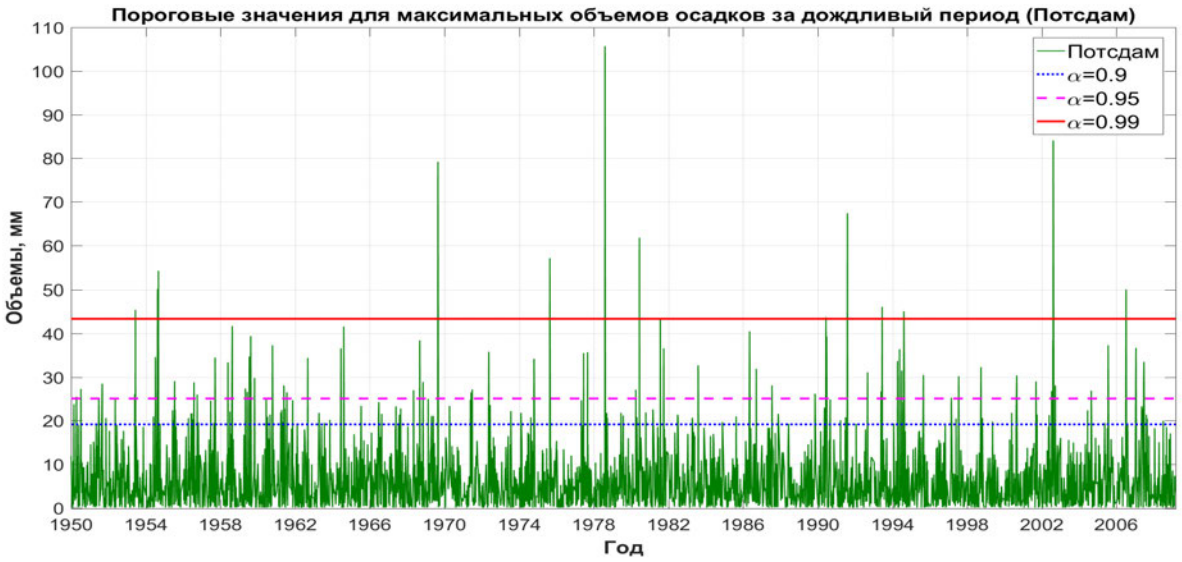

Рис. 6.47. Уровни для аномальных суточных объемов осадков, Потсдам

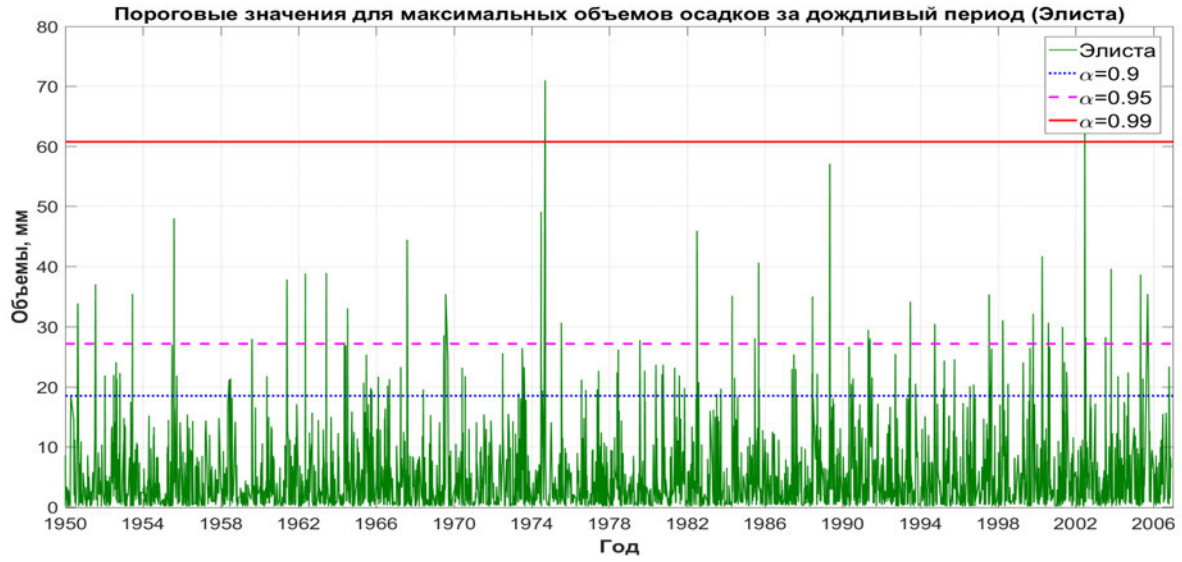

Рис. 6.48. Уровни для аномальных суточных объемов осадков, Элиста

периодов, содержащих аномальные наблюдения, при 99% уровне, и уже 69 – для порога, соответствующего 95% уровню. Остальные локальные экстремумы считаются стандартными. Для Элисты максимальный порог превышен всего в 2 «дождливых» периодах, а для уровня 95%-квантили количество таких интервалов равно 40. Такая разница в числе превышений для двух городов может объясняться тем фактом, что в Элисте экстремальные осадки являются более редкими событиями по сравнению с Потсдамом.

Было проведено сравнение данных тестов с методами, описанными в работе [443]. Установлено [293], что темперированное распределение Снедекора-Фишера обнаруживает экстремальные осадки более точно.

Таким образом, данная процедура подходит для идентификации и прогнозирования аномальных явлений, в том числе позволяя различать существенные, но не экстремально большие осадки.

6.5 Моделирование турбулентных потоков тепла между океаном и атмосферой

Изучение явных (далее обозначаются как `qe`) и скрытых (`qh`) поверхностных турбулентных тепловых потоков между атмосферой и океаном [104], определяющих их взаимодействие и являющихся составляющими теплового баланса, чрезвычайно важно для различных областей наук о Земле. Первый из указанных типов определяется на основе разности температур между средами (вода-воздух) и модуля скорости приводного ветра, умноженного на коэффициент, называемый числом Боуэна, второй связывает насыщенную влажность воздуха в точке с температурой и скоростью ветра. Данные о потоках могут быть получены из нескольких источников, каждый из которых обладает собственными достоинствами и недостатками. Наиболее подробные временные ряды (с периодами наблюдений в 100 и более лет) доступны с помощью программы `Voluntary Observing Ship` (`VOS`) [154], в то время как данные за последние несколько десятилетий собираются с помощью спутников, повторного анализа и комбинированных продуктов (например, `OA-FLUX` [430, 431]) с высоким пространственным и временным разрешением.

Информация об изменчивости поверхностных турбулентных тепловых потоков в большинстве случаев ограничивается первым (в некоторых случаях еще и вторым) моментом вероятностного распределения потоков. Они традиционно вычисляются по временному ряду, соответствующему потоку, и составляют основу для климатологических исследований [158, 277, 385]. Тем не менее детальная оценка характеристик теплового потока, в том числе определение экстремальных значений, требует точного знания вероятностного распределения, а также анализа изменений его параметров во времени и пространстве. Отсутствие подобных знаний при построении океанологических и климатических моделей серьезно снижает качество основанного на них прогнозирования.

Еще одна важная причина для изучения вероятностных распределений турбулентных потоков тепла – необходимость количественного оценивания и минимизации выборочных ошибок в продуктах, построенных по базам `VOS` [259, 260]. Большие по величине ошибки в выборках вли-

яют как на оценки средних величин потоков, так и на характеристики экстремальных потоков.

Попытка выбрать подходящий тип вероятностного распределения для турбулентных потоков была осуществлена в работе [261], в которой показано, что в качестве достаточно хорошей аппроксимации для данных может быть использовано так называемое двухпараметрическое распределение Фишера–Типпета (FT-распределение). В указанной работе были оценены параметры сдвига и масштаба данного распределения, осуществлена проверка гипотез о качестве подбора модели, а также была предложена глобальная климатологическая интерпретация параметров FT-распределения. Кроме того, FT-распределение было использовано для анализа временных рядов значительного объема для поверхностных турбулентных потоков, реконструированных по наблюдениям из базы `VOS` с 1880 года [262]. Тем не менее многие вопросы, связанные с вероятностным распределением поверхностных турбулентных потоков, все еще остаются открытыми. В частности, с помощью FT-распределения часто не удается корректно аппроксимировать экстремальные турбулентные потоки тепла, а также в полной мере учесть случай так называемых «тяжелых хвостов» в распределениях потоков.

Особое внимание будет уделено использованию алгоритма 3.5 на базе CPC-метода для конечных нормальных смесей, предложенного в разделе 3.3, для статистического оценивания случайных коэффициентов в стохастическом дифференциальном уравнении Ланжевена, описывающего турбулентные потоки тепла между океаном и атмосферой. Кроме того, продемонстрировано применение процедур, развитых для осадков, в случае океанологических данных, прежде всего, с целью выявления потенциально экстремальных наблюдений.

6.5.1 Однородность данных

Статистический анализ стохастических закономерностей в наблюдаемых временных рядах традиционно подразумевает работу со всеми имеющимися данными без какой-либо предварительной обработки с целью получения однородных данных. Например, в работе [261] FT-распределение применялось для аппроксимации исходного временного ряда. Однако такой подход, вероятно, не может быть использован для анализа очень длинных временных рядов и эволюции параметров распределения во времени. Действительно, выборка, используемая для статистического анализа, не является однородной, так как отдельные ее эле-

менты не являются независимыми. Чтобы пояснить это обстоятельство, рассмотрим следующий модельный пример.

Предположим, что n – натуральное число, $\xi_1, \xi_2, \ldots, \xi_n$ – независимые одинаково распределенные случайные величины с общей функцией распределения $F(x) = \Phi(x - a)$ (то есть каждая случайная величина ξ_j имеет нормальное распределение со средним a и единичной дисперсией). Определим новый набор случайных величин $\zeta_1, \zeta_2, \ldots, \zeta_n$ следующим образом: $\zeta_k = \xi_1 + \ldots + \xi_k, \quad k = 1, \ldots, n$. Для каждого k элемент ζ_k имеет нормальное распределение со средним ka и дисперсией k, при этом выборка $\zeta_1, \zeta_2, \ldots, \zeta_n$ не является однородной и независимой.

На рис. 6.49 данный эффект проиллюстрирован с помощью гистограмм, построенных по смоделированной выборке $\xi_1, \ldots, \xi_n$ с $n = 1000$ и $a = 2$ (см. верхний график) и соответствующей выборке $\zeta_1, \ldots, \zeta_n$ (см. нижний график). Нижняя гистограмма существенно скошена вправо с малым числом отрицательных значений. Данная картина в точности соответствует форме распределения, предложенного в [261].

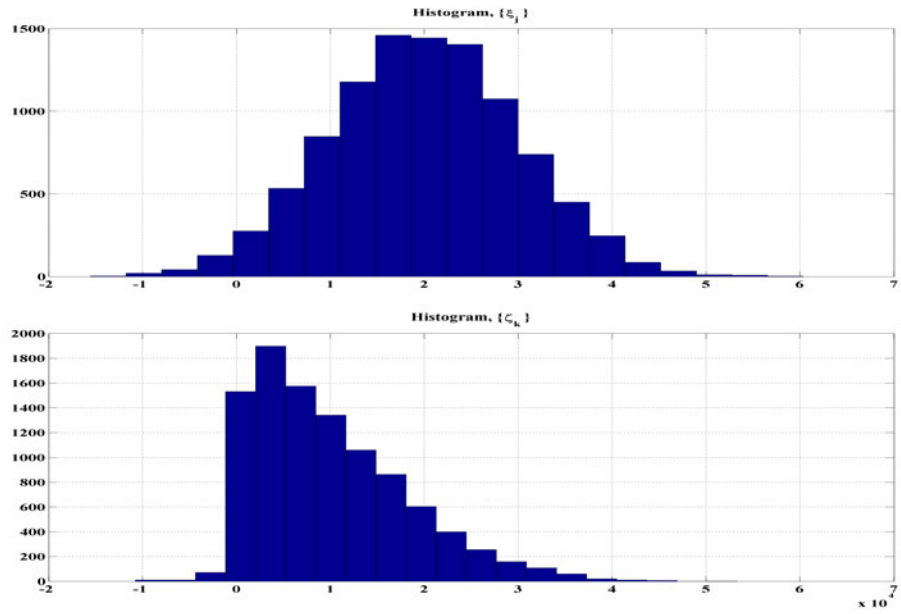

Рис. 6.49. Гистограммы для выборок $\{\xi_n\}$ (вверху) и $\{\zeta_n\}$ (внизу)

Стохастический характер ζ_k в значительной степени определяется суммами $\xi_1 + \ldots + \xi_{k-1}$ и слабо зависит от ξ_k. Чем больше величина k, тем меньший вклад случайной величины ξ_k в ζ_k. Таким образом, любой анализ статистических закономерностей ξ_i, $i = 1, \ldots, n$, непосредственно по выборке $\zeta_1, \zeta_2, \ldots, \zeta_n$ может выполняться только в рамках серьезных дополнительных допущений. Кроме того, с математической точки зрения стандартные статистические процедуры для выборки $\zeta_1, \zeta_2, \ldots, \zeta_n$ не применимы.

Чтобы избежать влияния упомянутых проблем, возникающих при использовании традиционных способов, следует проанализировать преобразованный временной ряд, рассмотрев ряд приращений турбулентных тепловых потоков.

6.5.2 Оценивание неизвестных параметров аппроксимирующих распределений

На рисунке 6.50 представлены гистограммы для приращений исходных данных, построенные на различных окнах (размер каждого – 200 наблюдений), с соответствующей аппроксимацией конечными нормальными смесями (1.7). Очевидно, что параметры приближающего распределения существенным образом изменяются со сдвигом окна, поэтому требуется применение СРС-метода.

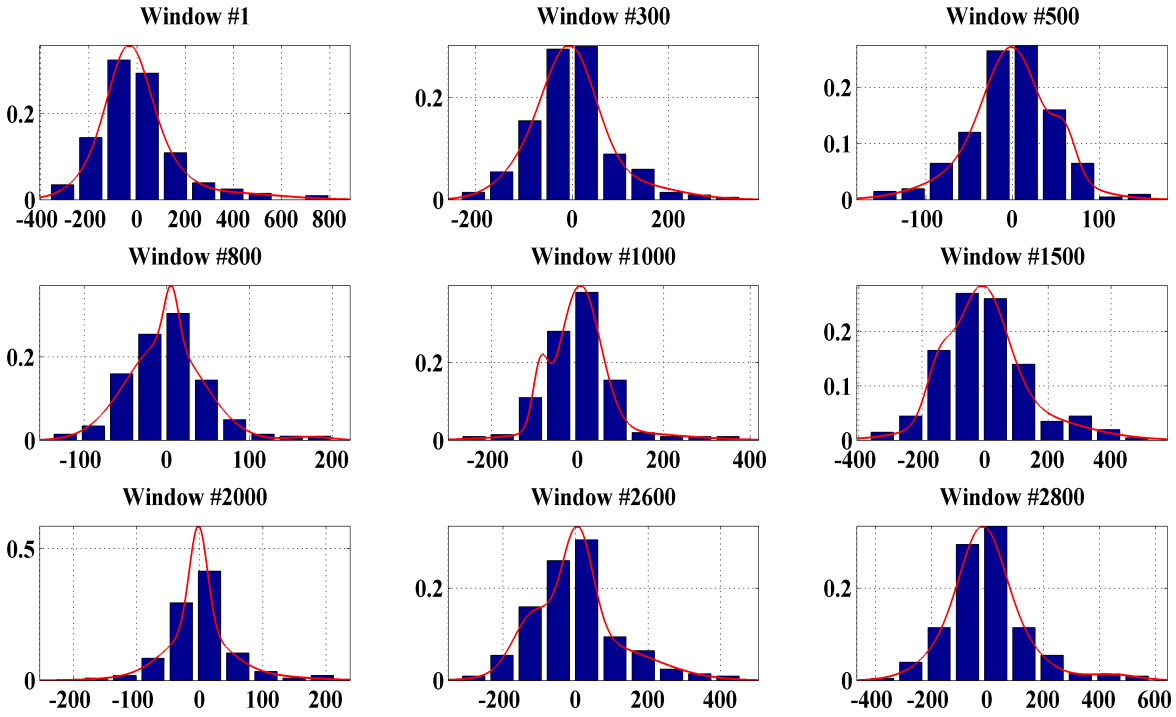

Рис. 6.50. Гистограммы, построенные для различных положений окна, с подгонкой плотностями типа конечных смесей нормальных законов

Для оценивания параметров модели (1.7) в данном случае использованы несколько модификаций EM-алгоритма, ориентированных на преодоление неустойчивости по начальному приближению, а именно:

– метод со случайным выбором начального приближения (`EMRnd`);

– несколько повторных запусков `EMRnd` на одном и том же окне и выбор в качестве итоговой оценки усреднения по всем из них (`EMmean`);

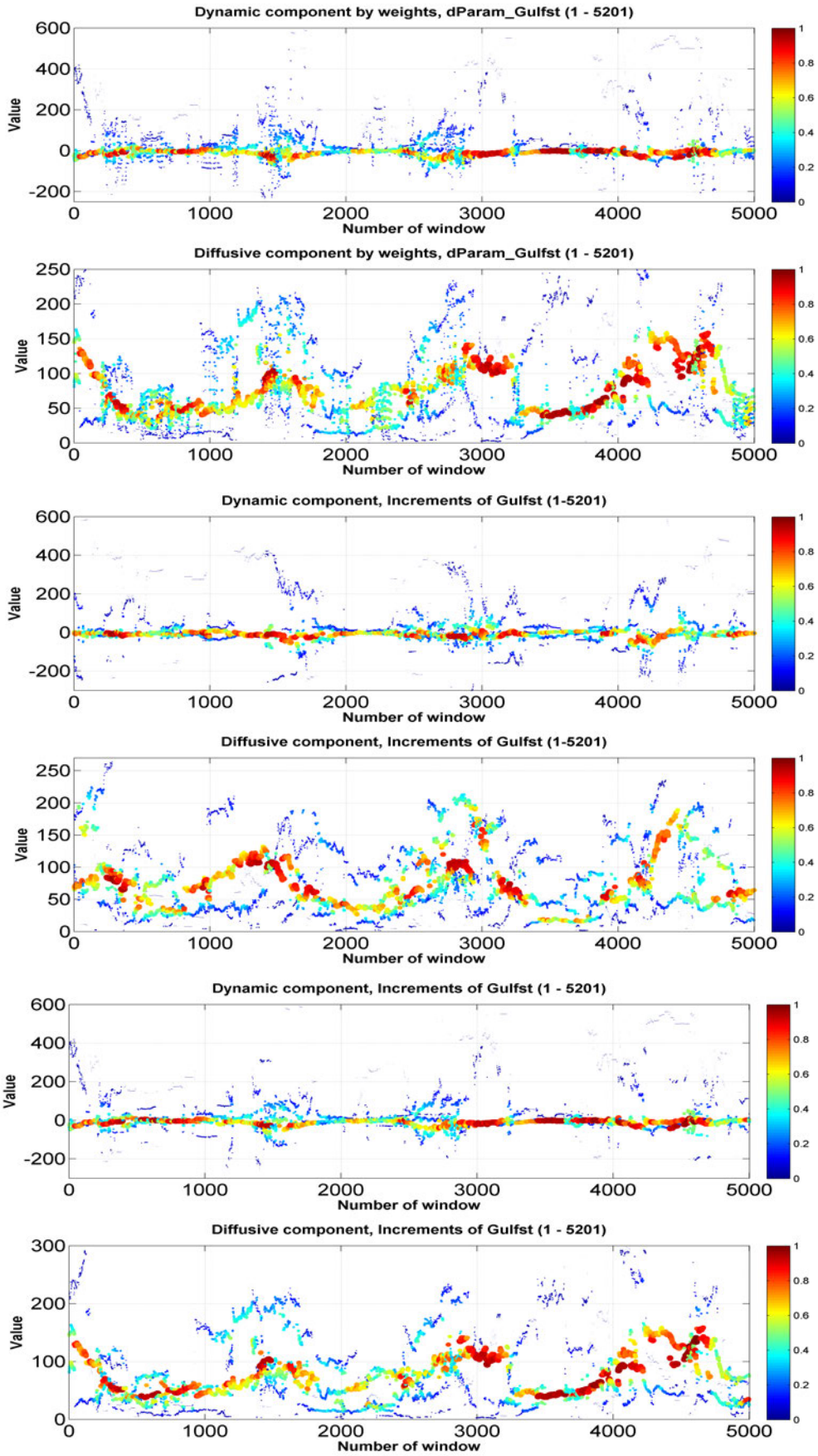

Рис. 6.51. CPC-оценки, полученные тремя модификациями EM-алгоритма

– многократный запуск `EMRnd` на одном и том же окне и выбор в качестве итоговой оценки набора параметров, максимизирующих значение функции правдоподобия (`EMLikelihood`).

Результаты, полученные для трех вышеуказанных способов выбора начального приближения, представлены на рис. 6.51. На верхнем графике представлена эволюция во времени параметров $a_i(t)$ для локальных трендов, оцененных с помощью EM-алгоритма со случайным выбором начальных приближений. На каждом окне расчеты с помощью EM-алгоритма проводятся пять раз, при этом начальные значения выбираются случайным образом для каждого запуска. Результаты усредняются по всему набору значений, полученному при работе алгоритмов на данном окне. Второй график демонстрирует изменение во времени локальных параметров диффузии $\sigma_i^2(t)$, оцениваемых с помощью описанной версии EM-алгоритма. На третьем графике продемонстрирована эволюция во времени параметров $a_i(t)$ для локальных трендов, оцениваемых с помощью «классического» EM-алгоритма со случайным выбором начальных значений для весов. На каждом окне начальные приближения для параметров сдвига и масштаба выбираются единым образом (как среднее и выборочная дисперсия для текущего окна соответственно). На четвертом графике приведены изменения во времени параметров локальных диффузий $\sigma_i^2(t)$, оцениваемых с помощью «классического» EM-алгоритма. На пятом графике приведены изменения во времени параметров $a_i(t)$ для локальных трендов, оцененных с помощью EM-алгоритма с выбором случайного начального приближения на каждом окне. Для каждого положения окна EM-алгоритм запускается пять раз, начальные значения выбираются каждый раз случайным образом. При этом в качестве оценки выбирается набор параметров, соответствующий функции правдоподобия с наибольшим значением среди всех запусков. Шестой график демонстрирует изменение во времени локальных параметров диффузии $\sigma_i^2(t)$, оцениваемых с помощью описанной версии EM-алгоритма.

Можно сделать вывод, что третья версия EM-алгоритма со случайным выбором начальных приближений и поиском набора, соответствующего максимальному значению функции правдоподобия на каждом окне, дает в данном случае наиболее наглядные результаты.

Данная модификация была использована для анализа временной изменчивости параметров распределения приращений тепловых потоков. Квантили различных уровней представлены на рис. 6.52 изотопными

линиями (горизонтальная ось времени соответствует периоду длительностью около $3,5$ лет). На графиках явно выделяется сезонная периодичность, что вполне соответствует физической природе рассматриваемого процесса.

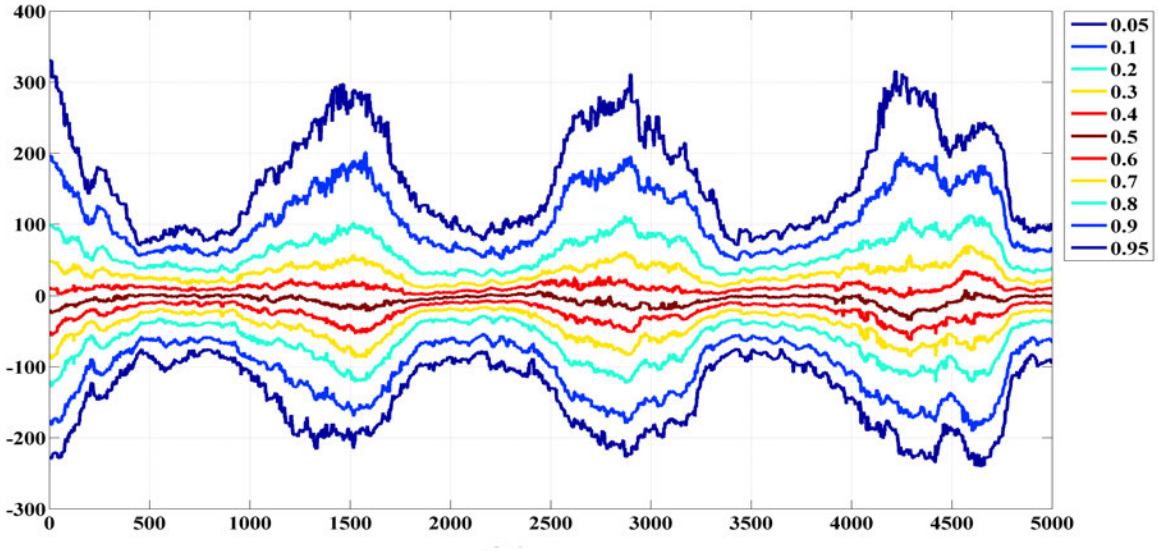

Рис. 6.52. Квантили распределения приращений потоков тепла

На рис. 6.53 представлена эволюция моментных характеристик распределения вероятностей приращений процесса тепловых потоков. Для их построения были использованы результаты, полученные в разделе 3.1.

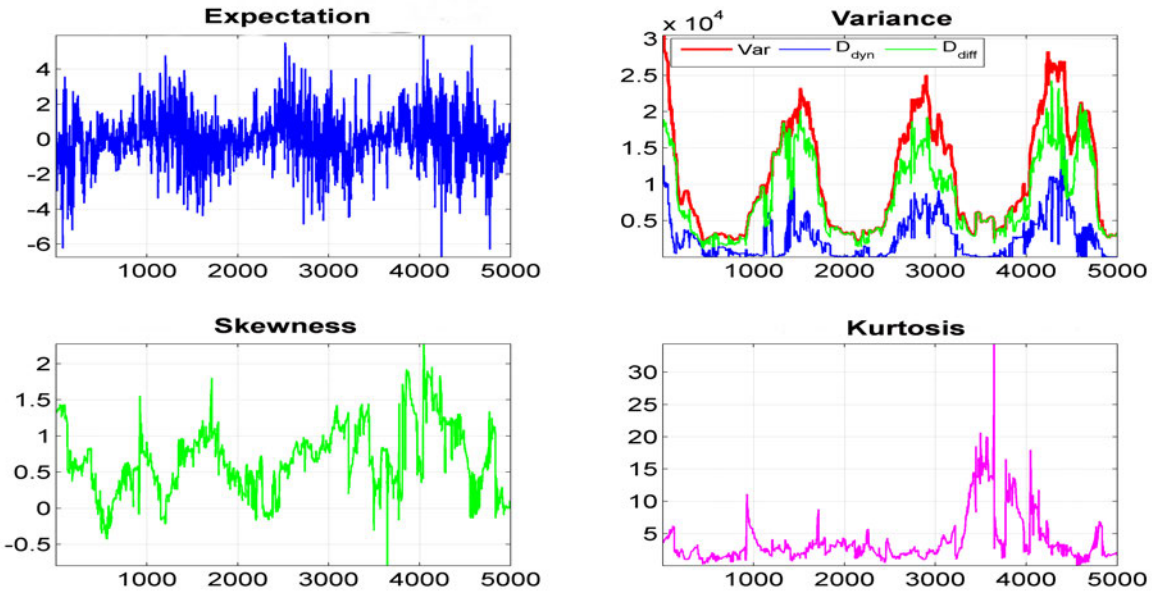

Рис. 6.53. Моментные характеристики вероятностного распределения приращений тепловых потоков

Видно, что математическое ожидание заметно колеблется во времени с изменяющейся амплитудой. Кроме того, для каждого периода амплитуда меньше для периода сезонного увеличения общего среднего по сравнению с амплитудой колебаний для периода сезонного снижения общего среднего значения. Хорошо прослеживается сезонный характер изменения дисперсии. Достаточно интересным представляется и тот факт, что вклад в общую дисперсию чисто стохастической диффузионной компоненты дисперсии – зеленая кривая на правом верхнем графике – больше, чем динамической составляющей, изображенной синей кривой. Можно отметить, что правый хвост распределения приращений тяжелее левого. Кроме того, что эксцесс этого распределения максимален во время периода «спокойствия».

6.5.3 Статистическое оценивание случайных коэффициентов в уравнении Ланжевена

Для статистической оценки коэффициентов в уравнении Ланжевена, используемого для моделирования процессов переноса тепла между океаном и атмосферой [149], воспользуемся описанным в разделе 3.3 алгоритмом 3.5, а также CPC-оценками, полученными при аппроксимации распределений приращений потоков тепла между океаном и атмосферой в нескольких географических точках: Гольфстрим, Лабрадорское море и тропики.

На рисунках 6.54–6.65 верхние графики демонстрируют статистическую структуру процесса теплообмена, а на нижних, содержащих CPC-оценки для параметров сдвига и масштаба, которые формируют динамическую и диффузионную компоненты, приведена эволюция весов, то есть вклад соответствующей структурной составляющей в общее развитие процесса во времени.

Благодаря структурам, возвращаемым функцией `MSMComponents` (см. алгоритм 3.5), можно точно отследить, когда те или иные из компонент существовали, прерывались и возобновлялись, и проанализировать их взаимосвязь с реальными физическими процессами.

Для аппроксимации были использованы четырехкомпонентные нормальные смеси, однако при выбранных настройках жадного алгоритма для всех рядов, за исключением явных потоков в тропиках, были получены пять локальных компонент связности.

Видно, что общее число компонент изменяется не слишком сильно,

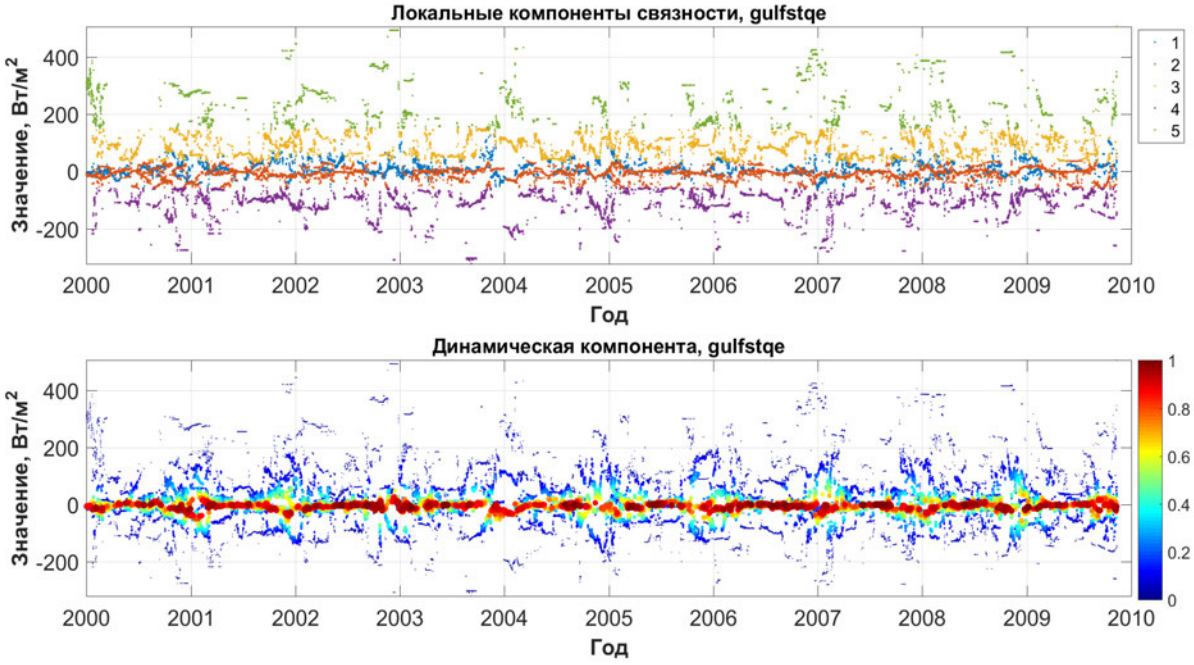

Рис. 6.54. Оценки распределения сдвига (Гольфстрим, явные потоки)

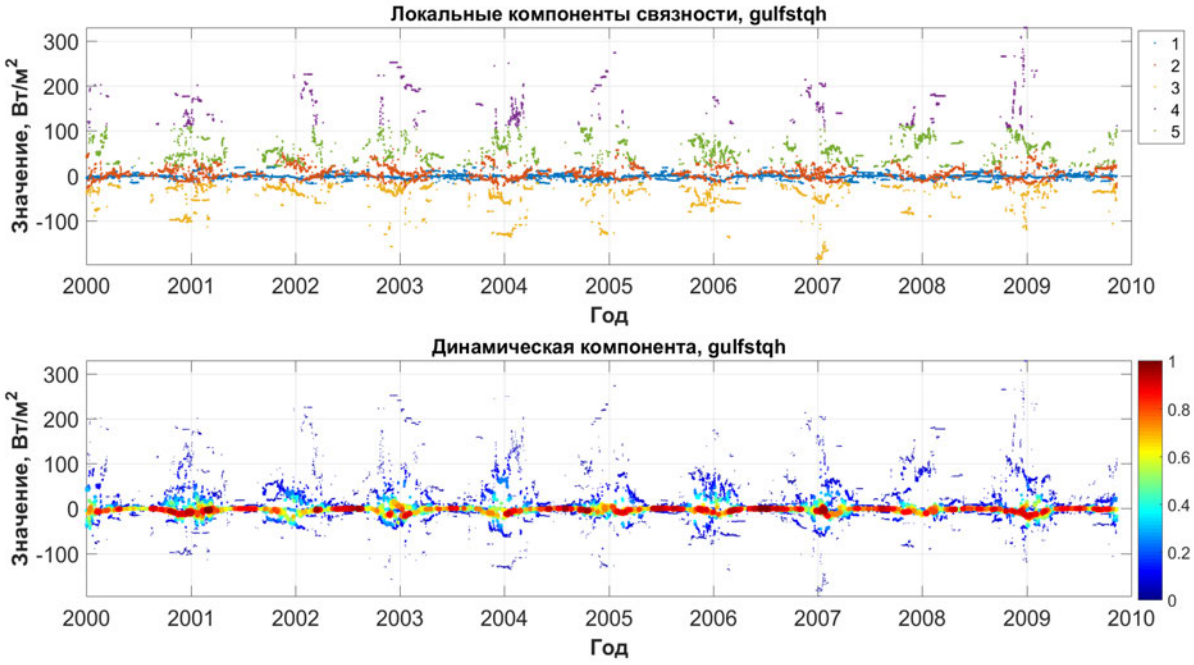

Рис. 6.55. Оценки распределения сдвига (Гольфстрим, скрытые потоки)

поэтому результаты автоматического определения их количества с помощью жадного алгоритма 3.4 варьируются от ряда к ряду не очень существенно. Однако для лучшего учета локальных процессов полученное число компонент (4–5) может быть расширено за счет повышения чувствительности процедуры путем выбора меньшего порогового значения в формуле (3.28).

Таким образом, предложенная процедура (см. алгоритм 3.5) оказа-

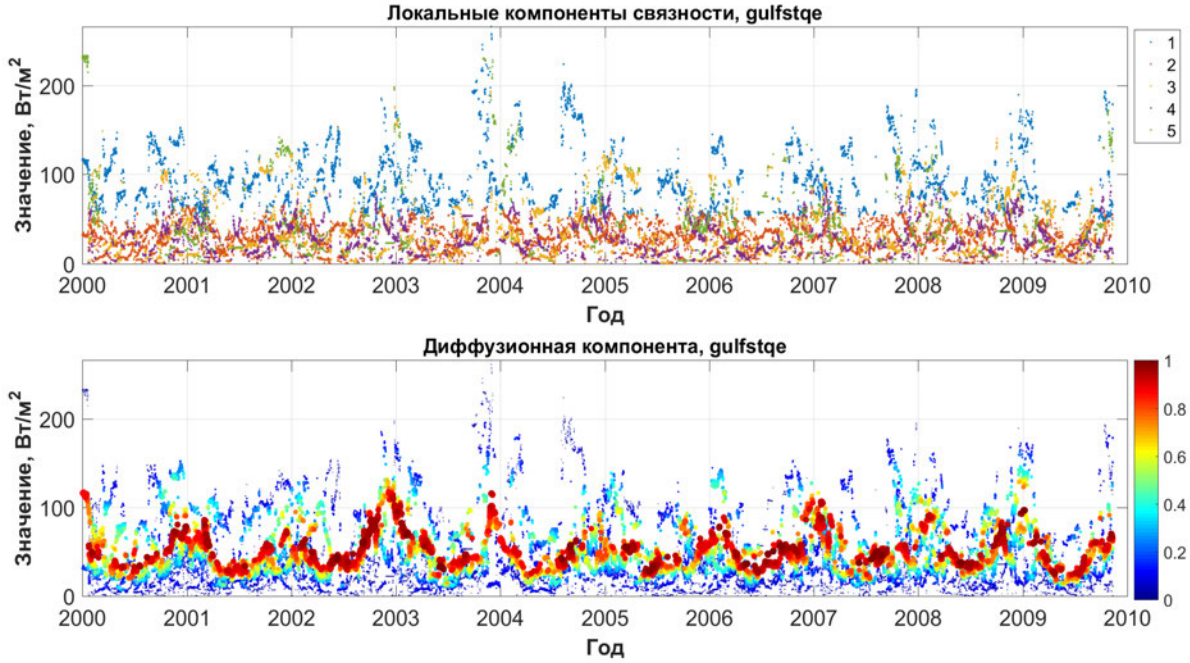

Рис. 6.56. Оценки распределения коэффициента диффузии (Гольфстрим, явные потоки)

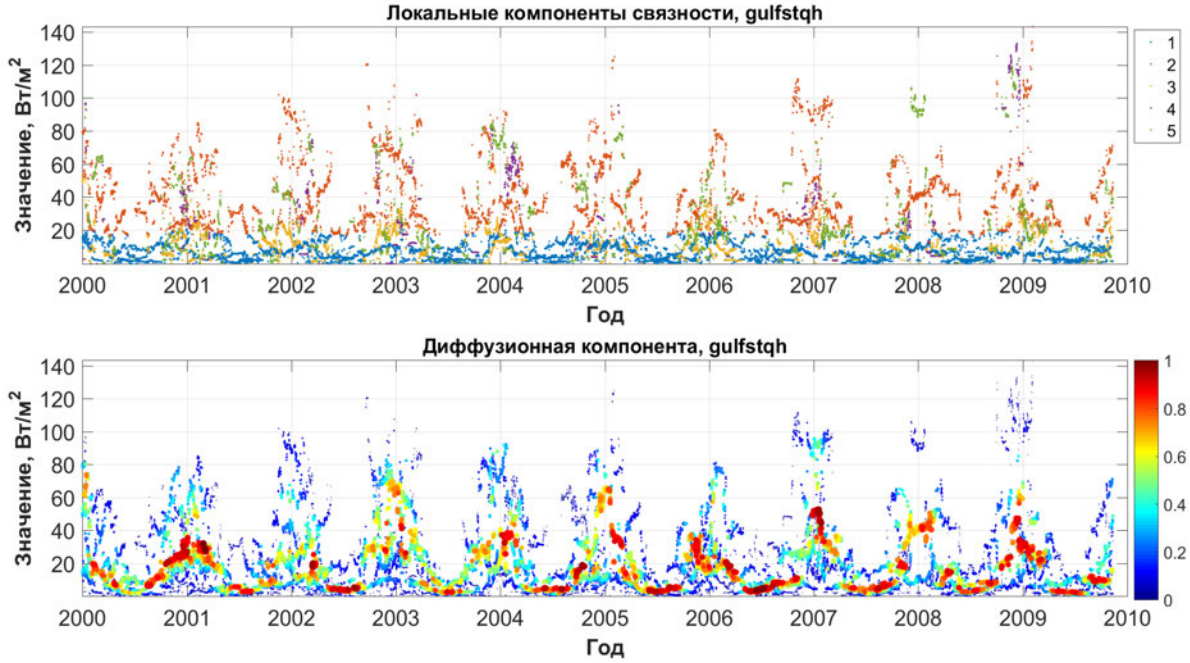

Рис. 6.57. Оценки распределения коэффициента диффузии (Гольфстрим, скрытые потоки)

лась эффективной и для решения задачи определения числа структурных компонент в плазменной турбулентности (см. раздел 5.2), и в рамках статистического оценивания случайных функциональных коэффициентов в уравнении Ланжевена для потоков тепла между океаном и атмо-

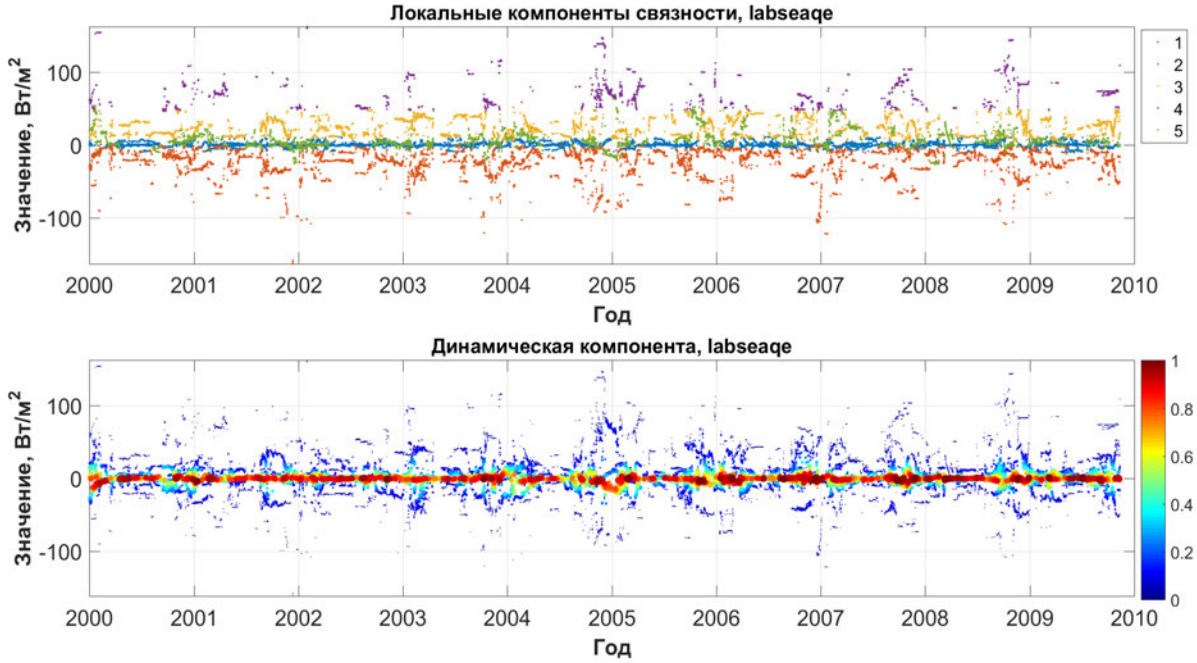

Рис. 6.58. Оценки распределения сдвига (Лабрадорское море, явные потоки)

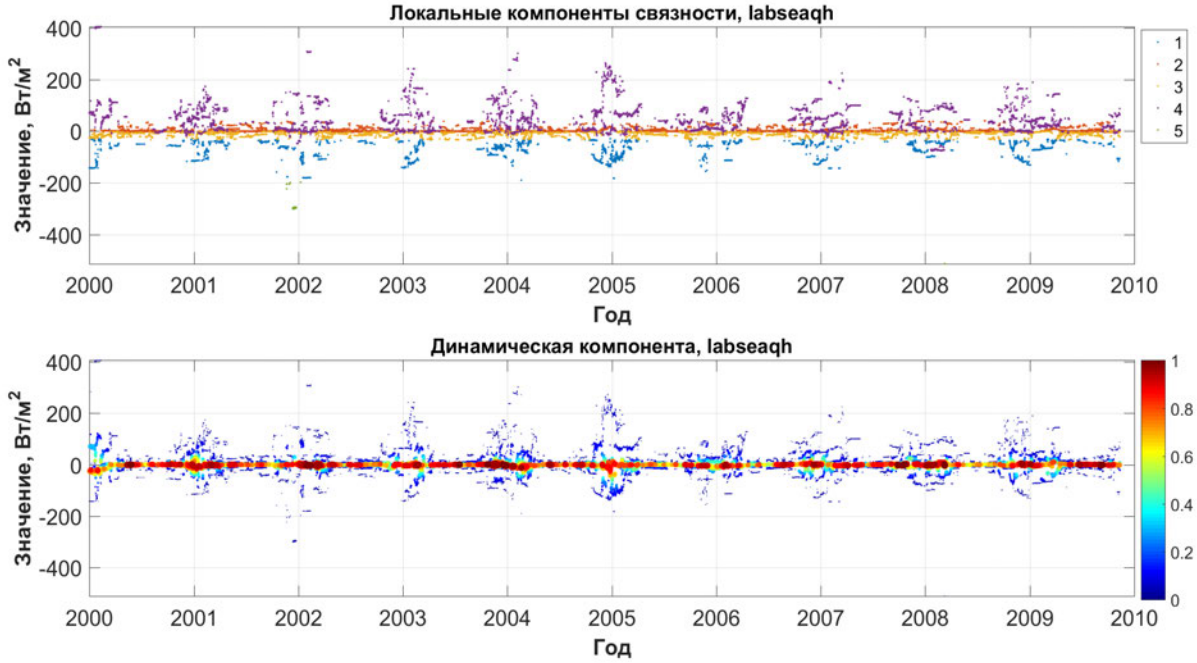

Рис. 6.59. Оценки распределения сдвига (Лабрадорское море, скрытые потоки)

сферой.

Функциональные параметры (компоненты распределения (2.3) как функции времени), полученные в результате описываемых статистических процедур, могут быть использованы при обучении интеллекту-

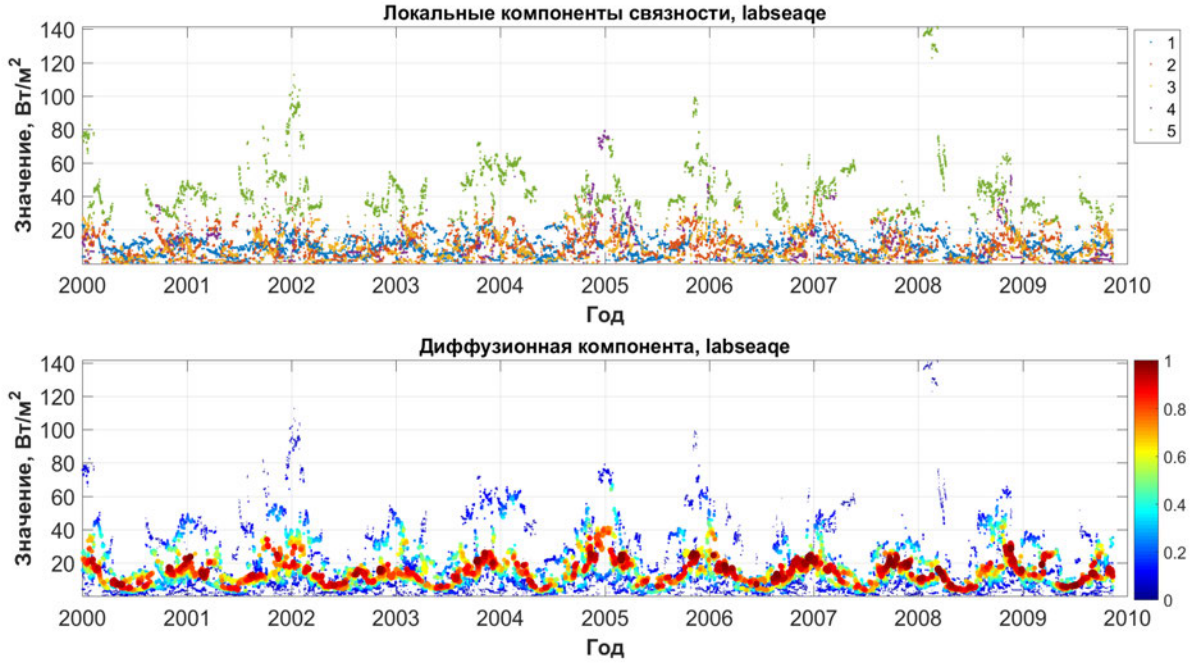

Рис. 6.60. Оценки распределения коэффициента диффузии (Лабрадорское море, явные потоки)

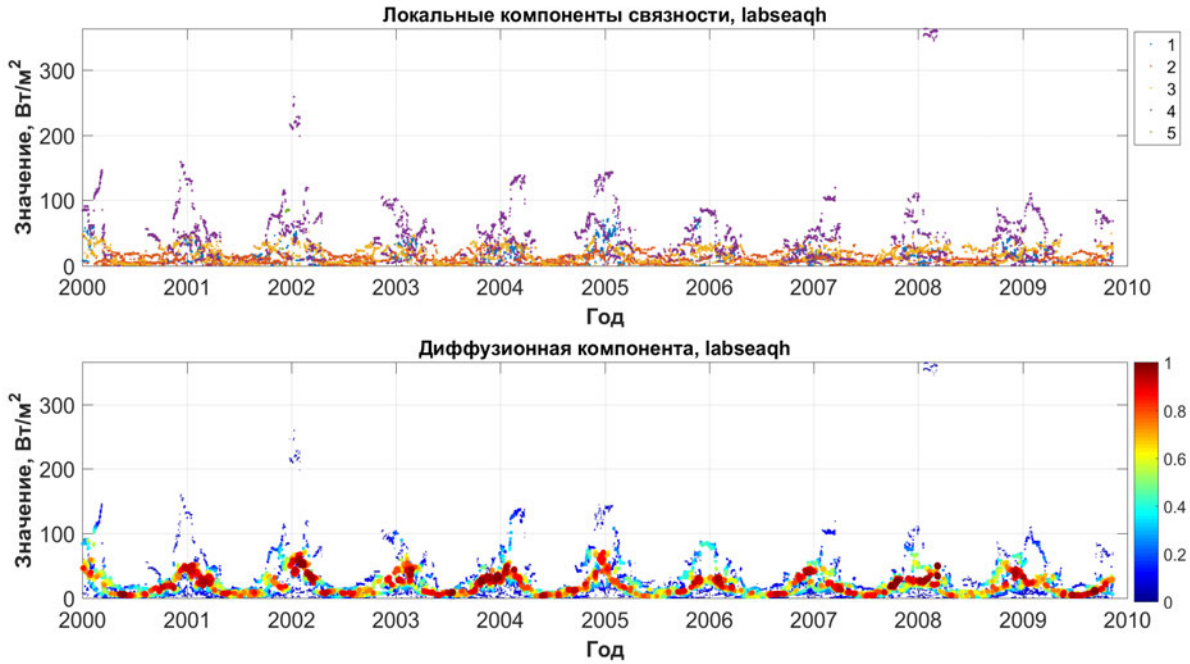

Рис. 6.61. Оценки распределения коэффициента диффузии (Лабрадорское море, скрытые потоки)

альных алгоритмов прогнозирования процесса $X(t)$, удовлетворяющего уравнениям типа (2.1). В частности, для этого могут быть использованы как полученные оценки коэффициентов уравнения, так и построенные по ним моментные характеристики аппроксимирующих нормальных

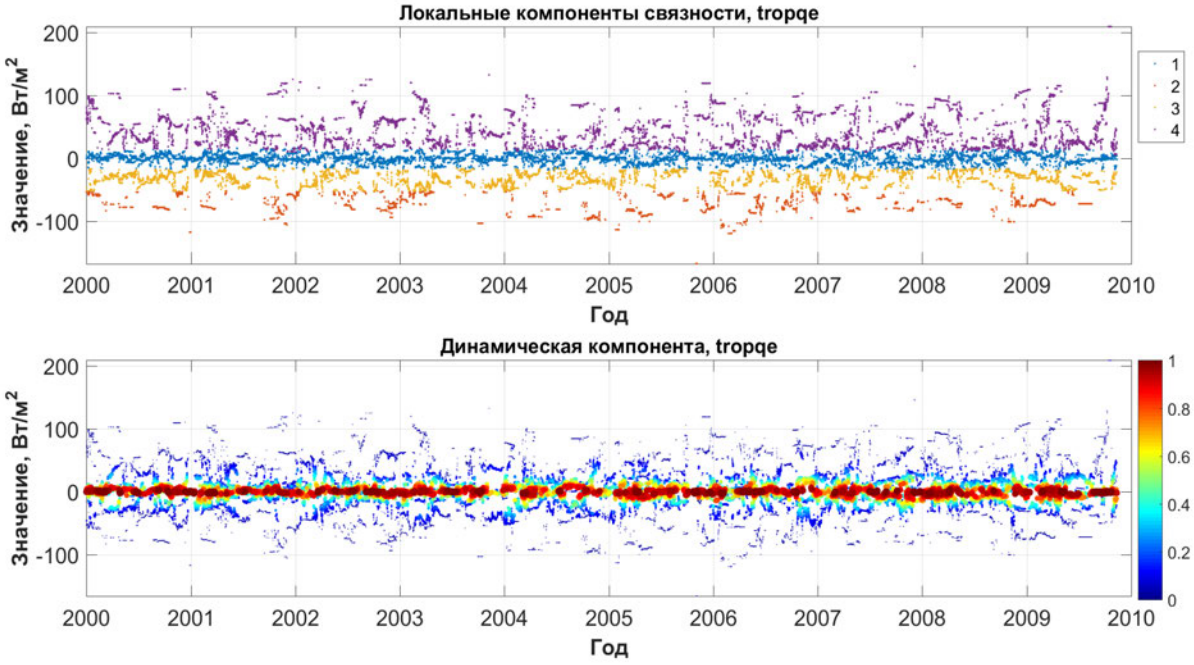

Рис. 6.62. Оценки распределения сдвига (тропики, явные потоки)

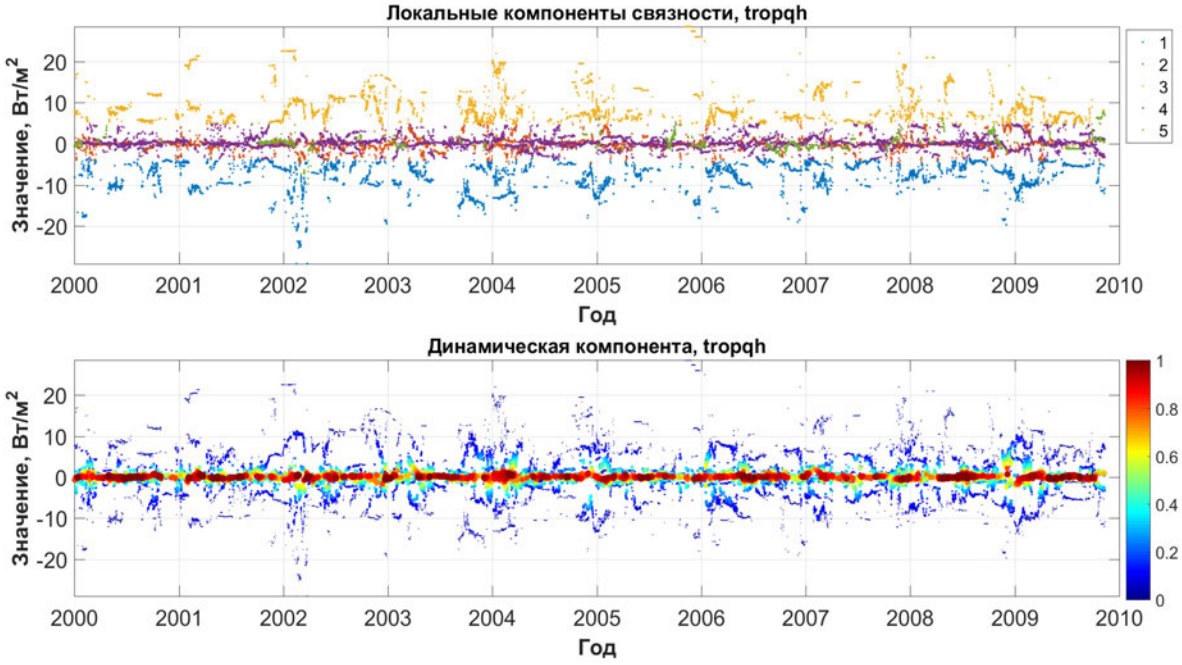

Рис. 6.63. Оценки распределения сдвига (тропики, скрытые потоки)

смесей. Как было показано в разделе 5.2 для физических рядов турбулентной плазмы, расширение признакового пространства нейронных сетей за счет таких величин позволяет повысить точность обучения, причем в ряде случае достаточно существенно, не прибегая к необходимости получения новых объемов данных. Это весьма важное обстоятельство для океанологических пространственно-временных рядов. Все

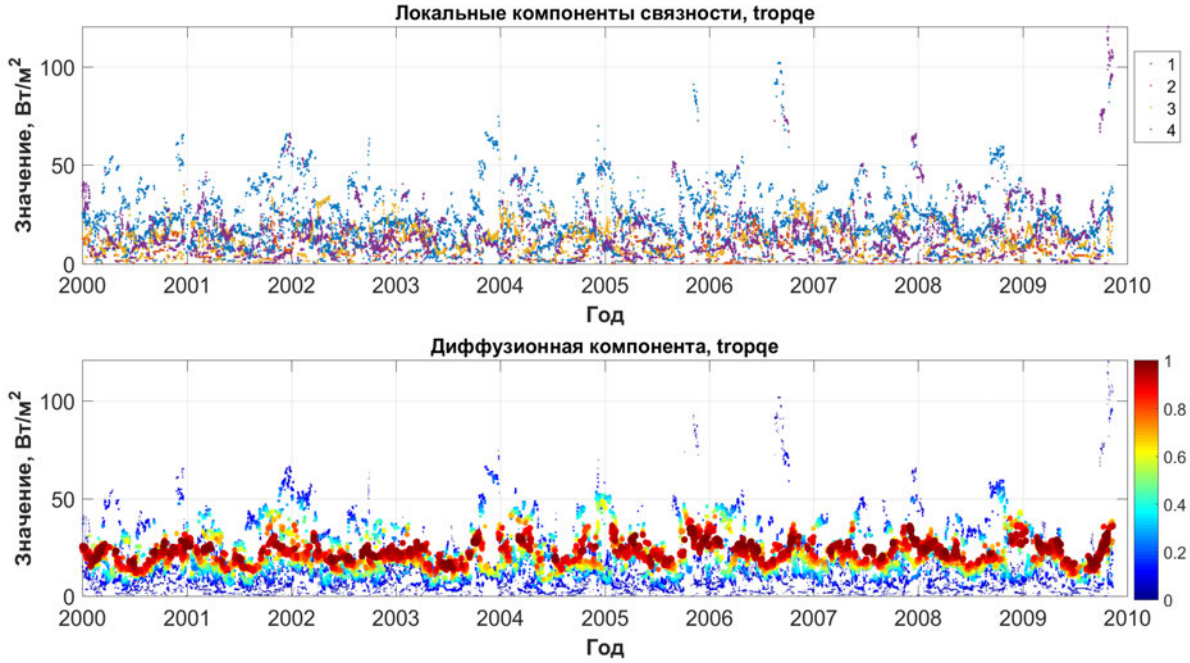

Рис. 6.64. Оценки распределения коэффициента диффузии (тропики, явные потоки)

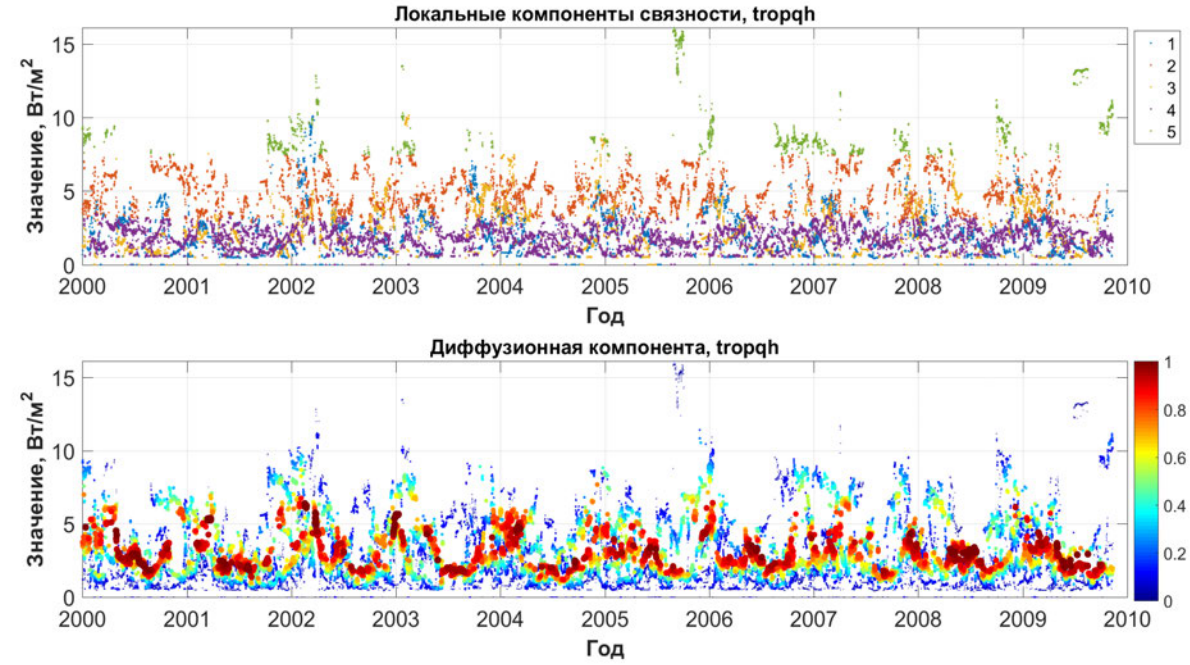

Рис. 6.65. Оценки распределения коэффициента диффузии (тропики, скрытые потоки)

анализируемые выборки для потоков получены с интервалом в шесть часов между соседними наблюдениями. Это значительно меньше, чем время, необходимое для получения CPC-оценок и даже обучения с их учетом нейронных сетей, что свидетельствует в пользу эффективности предлагаемых вычислительных процедур в задачах обработки и анализа подобных данных.

6.5.4 Анализ экстремальных наблюдений

Сначала рассмотрим достаточно наглядный способ определения доли экстремальных наблюдений в исходной выборке на основе CPC-метода. Предположим, что в модели (1.7) параметр k выбирается равным двум, причем как для ускорения вычислений, так и для получения более контрастной общей картины. На рис. 6.66 представлены два графика для рассмотренного ранее ряда.

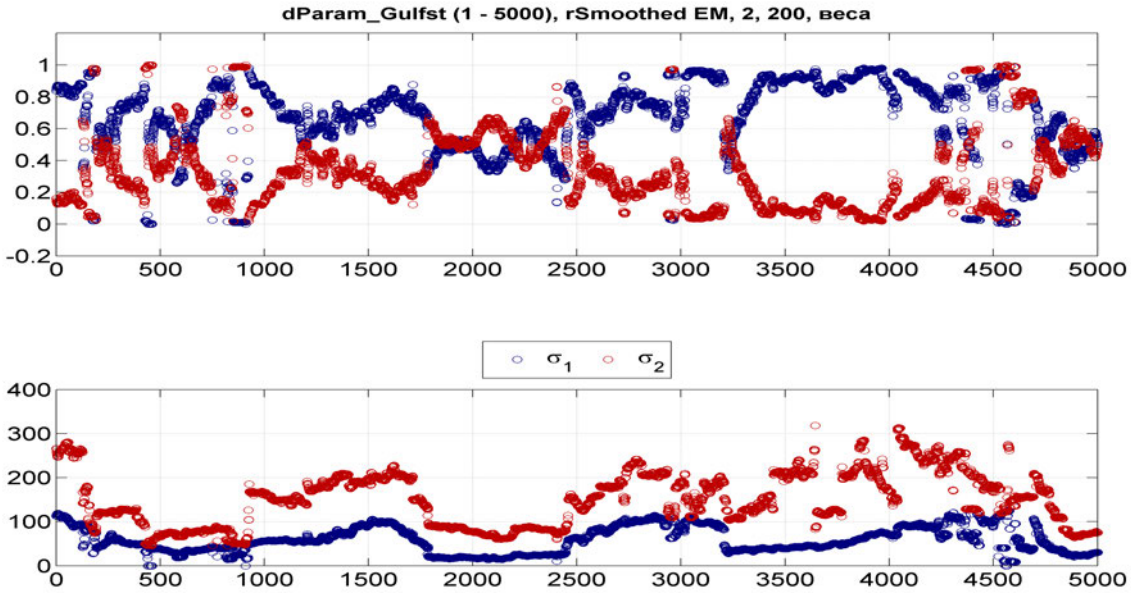

Рис. 6.66. Определение доли экстремальных наблюдений на основе весов (вверху) и среднеквадратических отклонений (внизу)

На верхнем графике изображена эволюция весов каждой из компонент, а на нижнем – соответствующие им среднеквадратические отклонения. При этом справедливо соотношение $\sigma_1 \leqslant \sigma_2$ (отметим, что один параметр превосходит другой минимум в 2 раза для каждого положения скользящего окна), а цвета точек на обоих графиках соответствуют друг другу: красным обозначены графики для компоненты с наибольшей дисперсией на каждом окне, синим – с наименьшей.

Наблюдения, соответствующие компоненте с наибольшей дисперсией, могут быть проинтерпретированы как экстремальные (относительно наблюдений другой компоненты). Стоит отметить, что веса этой компоненты лежат в диапазоне $[0{,}2, 0{,}4]$, то есть можно сказать, что указанные наблюдения составляют примерно треть от общего числа. Отметим, что указанная картина сохраняется для всех сезонов, при этом характер экстремальных наблюдений меняется.

Как было отмечено в разделе 6.4, модифицированный метод определения экстремальных значений может быть использован для данных любой природы. В данном разделе воспользуемся им для анализа скрытых и явных потоков тепла (см. алгоритм 6.13).

Алгоритм 6.13. PoT-метод для потоков тепла

function PoTFluxes(Fluxes, α=0.05)
    j$\leftarrow$ 1;
    for all Fluxes **do**
        // Исходящий PoT-метод для потоков, см. алгоритм 6.8
        [$p_{val}^{(W)}$, $p_{val}^{(GP)}$, LVL] $\leftarrow$ PoT($Flux_j$, 0.1);
        Ind$\leftarrow$ FindFirstIndex($p_{val}^{(W)} > \alpha$ **and** $p_{val}^{(GP)} > \alpha$);
        Threshold$_j$ $\leftarrow$ LVL$_{Ind}$;
        j++;
    return Thresholds;

На рисунке 6.67 представлены пороговые значения, определенные с помощью нисходящего метода (см. алгоритм 6.8) для скрытых и явных потоков тепла для Гольфстрима (примеры статистической обработки данных qe был рассмотрен выше) и тропической зоны за 2000–2010 гг.

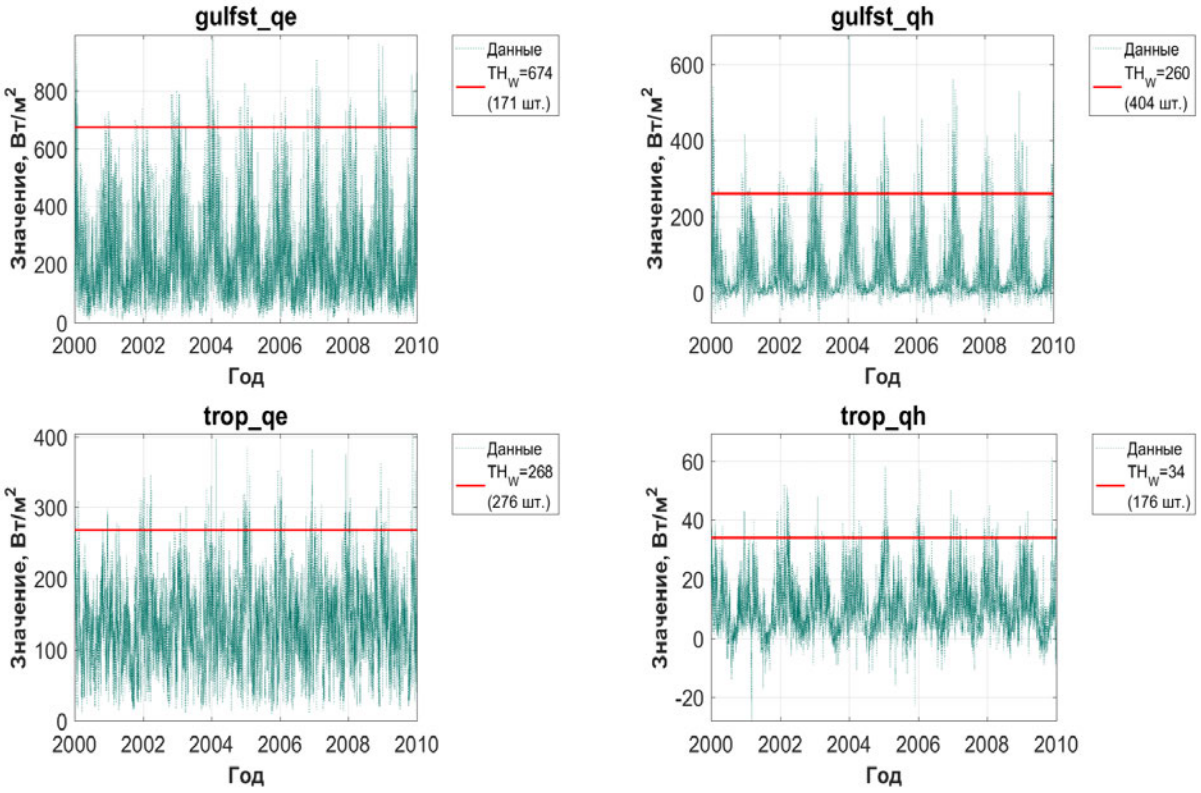

Рис. 6.67. Пороговые значения для потоков тепла

В легенде в явном виде указывается значение порога, полученного на основе проверки гипотезы вейбулловости для моментов времени между его превышениями, при этом в скобках приведено число элементов, превышающих данный уровень. Очевидно, что для каждого из случаев речь идет о сотнях превышений (при этом все пороги расположены достаточно близко к максимуму данных). Кроме того, в приведенных примерах существенную роль играет сезонность. Поэтому, несмотря на продемонстрированную возможность использования PoT-методологии и для таких наблюдений, здесь по аналогии с осадками можно рекомендовать более глубокую разработку аналитического аппарата для повышения качества обработки данных.

6.5.5 Аппроксимация распределений характеристик локальных трендов

В этом разделе рассмотрим не сами потоки тепла, а некоторые производные величины, а именно локальные тренды монотонности в рядах. Проанализируем распределения длительностей промежутков монотонного возрастания и убывания данных, а также величины суммарных потоков за каждый из указанных периодов по аналогии с «дождливыми» периодами и соответствующими объемами осадков (см. раздел 6.3).

Алгоритм 6.14. Аппроксимация распределений характеристик трендов

```
1:  function FluxesTrends(Fluxes)
2:      i←1;
3:      Params ← ∅;
4:      for all Fluxes do
5:          // Формирование рядов для анализа для каждого потока
6:          [Trend↑, Trend↓, Sum↑, Sum↓]←Trends(Fluxes_i);
7:          // Params_i формируется как объединение всех параметров,
        возвращаемых вызываемыми в блоке функциями
8:          procedure Params_i
9:              // Функциональное оценивание (алгоритмы 6.5 и 6.6)
10:             GNBApprox(Trend↑,ℓ²,0,1); GNBApprox(Trend↓,ℓ²,0,1);
11:             GGApprox(Sum↑, L², 0, 1); GGApprox(Sum↑, L², 0, 1);
12:         subplot(Params_i, 2, 2);                    // Визуализация
13:         i++;
14:     return Params;
```

Рассмотрим аппроксимацию эмпирических распределений указанных случайных величин с помощью обобщенных отрицательного биномиального и гамма-распределений (см. алгоритм 6.14).

В данном алгоритме выделяются следующие ключевые этапы:

– определение границ локальных трендов для последующего анализа (см. функцию `Trends` в алгоритме 6.14);

– функциональная аппроксимация распределений к данным (см. процедуру $\texttt{Params}_i$) на основе методов, описаных в разделе 6.3;

– визуализация результатов одновременно для локально возрастающих и убывающих трендов, включая и соответствующие суммы (см. примеры на рисунках 6.68–6.71).

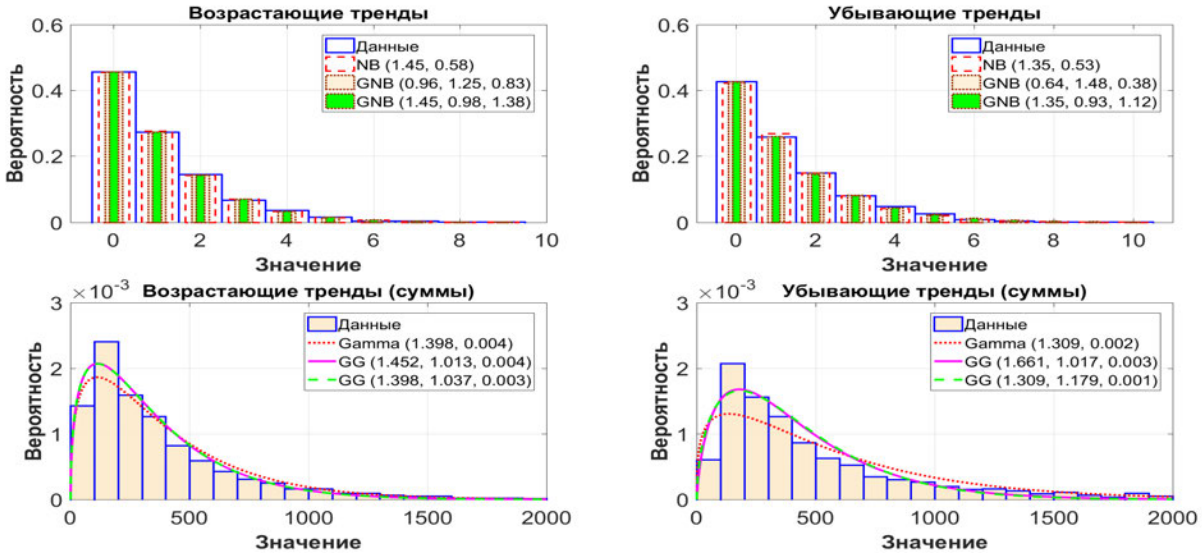

Рис. 6.68. Распределения характеристик (Гольфстрим, явные потоки)

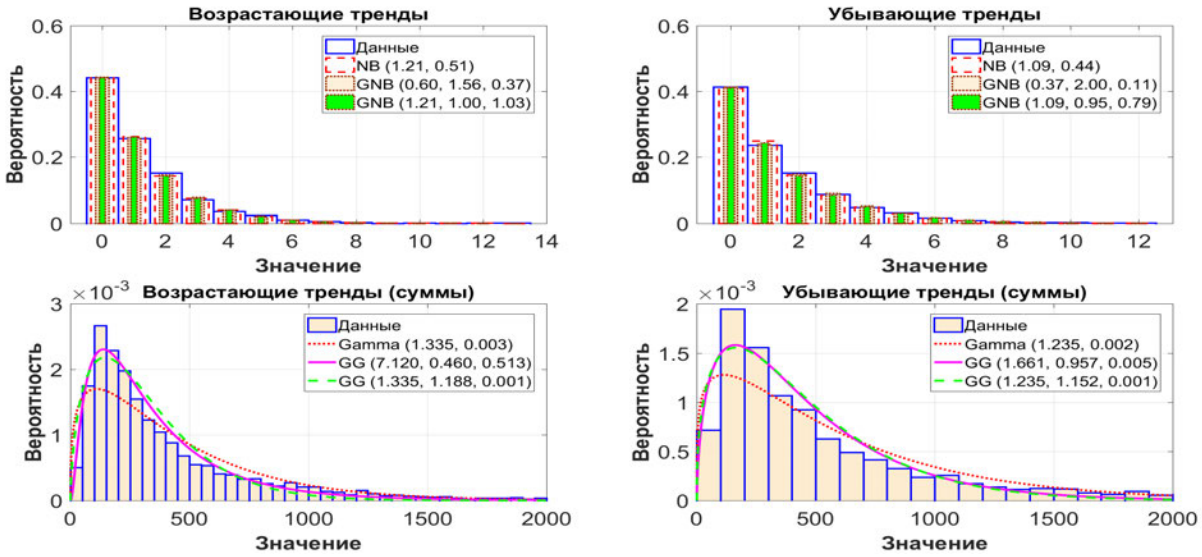

Рис. 6.69. Распределения характеристик (Гольфстрим, скрытые потоки)

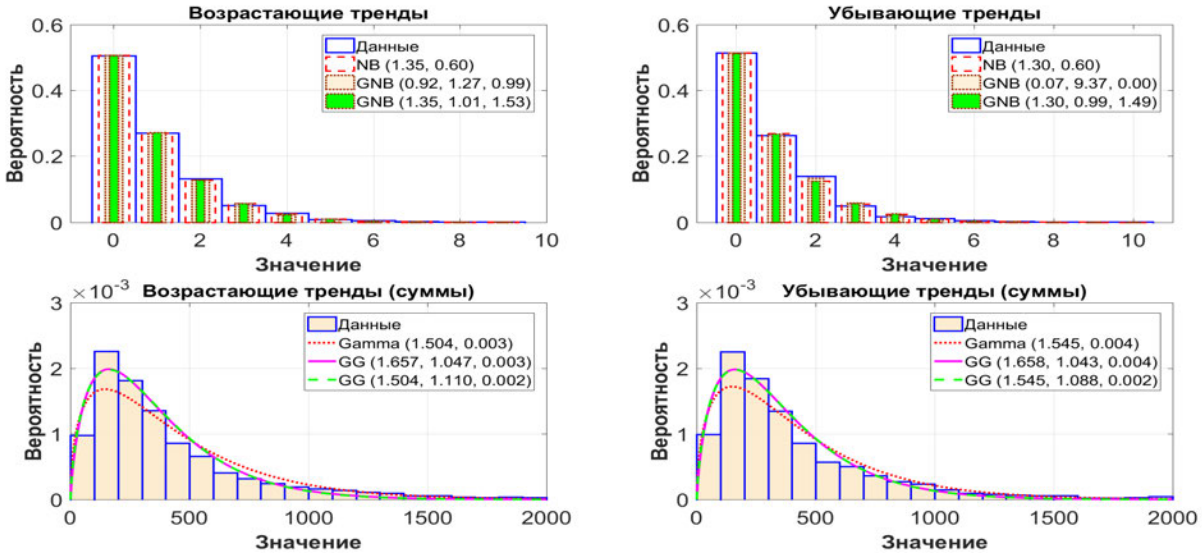

Рис. 6.70. Распределения характеристик (тропики, явные потоки)

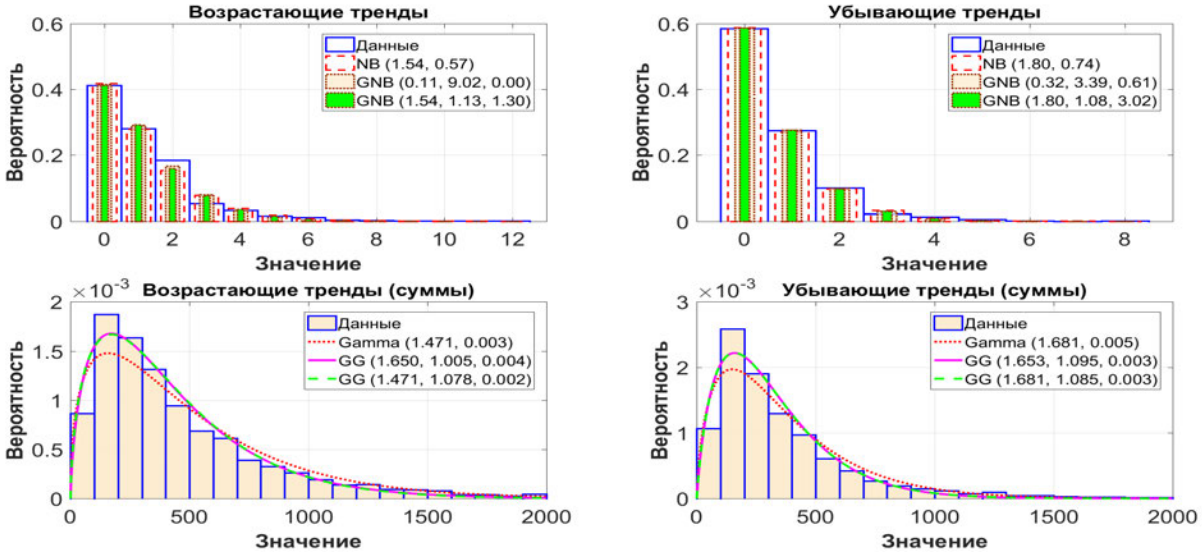

Рис. 6.71. Распределения характеристик (тропики, скрытые потоки)

В алгоритме 6.14 реализован поиск оценок и при фиксированном параметре формы r, а также производится сравнение с классическими отрицательным биномиальным и гамма-распределениями (см. параметры вызовов функций `GNBApprox GGApprox`). Во всех случаях критерий χ^2 свидетельствует против принятия гипотезы о гамма-распределении. Для отрицательного биномиального распределения были получены P-значения 0,55 и 0,08 (Гольфстрим, скрытые и явные потоки), 0,25 для скрытых потоков в тропиках. В остальных случаях гипотеза отвергалась. Необходимо отметить высокое визуальное соответствие гистограмм и приближающих кривых, которое наглядно демонстрирует преимущества использования обобщенных смешанных моделей.

Глава 7

Прикладные программные комплексы

В данной главе рассматриваются программные решения и комплексы, разработанные на их основе, которые использовались для анализа неоднородных данных и визуализации результатов в главах 3–6. Приводится описание разработанных интерфейсов, функциональных возможностей и архитектурных решений. Обсуждаются вопросы реализации разработанных методов вероятностного моделирования и алгоритмов анализа данных в виде онлайн-системы для поддержки междисциплинарных исследований и научных сервисов цифровой платформы.

7.1 Инструменты графического вывода результатов метода скользящего разделения смесей

При практической реализации CPC-метода возникает задача отображения изменяющихся во времени динамической и диффузионной компонент (см. раздел 2.1). Помимо числового значения соответствующих параметров на каждом шаге, оценки также имеют веса, отображение которых в большинстве случаев необходимо произвести на двумерном графике. Каждой точке по оси абсцисс соответствует текущее положение окна, а по оси ординат откладываются значения оценок, полученных на данном шаге. Для изображения весов на графиках используется цветовая шкала с плавной градацией от темно-синего до темно-красного, при этом каждому весу из сегмента $[0, 1]$ ставится в соответствие цвет в шка-

ле RGB. Вес определяет размер выводимой точки, который пределяется как $\lceil p_i^{(m)} \cdot Size_{max} \rceil$, где $p_i^{(m)}$ обозначает вес компоненты с номером i на m-м итерационном шаге, а $Size_{max}$ – некоторое заранее заданное максимальное значение размера. Ниже рассмотрим несколько вариантов программных инструментов, использовавшихся для визуализации результатов работы CPC-метода, созданных на языке программирования MATLAB.

7.1.1 Оконный пользовательский интерфейс

Для начала рассмотрим оконный пользовательский интерфейс для выполнения расчетов с помощью целого ряда модификаций EM-алгоритма (классический, медианный, стохастический, сглаженный, основанный на повторных вычислениях на окне, с матричными вычислениями на GPU) для различных типов смесей вероятностных распределений (нормальные, гамма). Реализация CPC-метода продемонстрирована в виде псевдокода в алгоритме 7.1.

Алгоритм 7.1. CPC-метод с оконным интерфейсом

```
1:  function MSM( )
2:      // Ввод данных через диалоговый интерфейс
3:      [Data, From, To]← InputDialog( );
4:      options← OptionsDialog( );          // Настройки CPC-метода
5:      // Запуск CPC-метода с заданными параметрами
6:      Params← EMs(Data, options);
7:      // Динамическая настройка диапазонов вывода компонент
8:      PlotOpt← PlotDialog( );
9:      // Визуализация динамической и диффузионной компонент
10:     FastPlotter(Params, PlotOpt);
11:     // Определение моментных характеристик, см. (3.9)–(3.12)
12:     [Exp, Var, Skew, Kurt]← Moments(Params);
13:     Quantiles(Params);      // Вычисление и отрисовка квантилей
14:     return ;
```

В начале работы пользователю предлагается ввести имя (путь) для файла с данными. Может быть выбран любой диапазон внутри ряда, при этом по умолчанию предлагаются все значения от первого до последнего (с автоматическим определением длины ряда при указании параметра End). В случае ввода пользователем некорректных значений (например,

отрицательных или превышающих объем выборки) программа возвращает настройки к предустановленным. Данные действия осуществляются функцией `OptionsDialog` (см. алгоритм 7.1). Вид начального окна представлен на рис. 7.1.

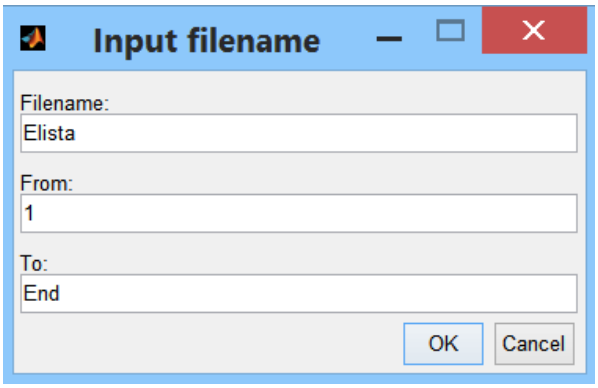

Рис. 7.1. Пример: начальное окно программы

Пользователь может выбрать отображение исходных данных или разностей, вывести гистограммы для этих рядов с автоматической аппроксимацией конечной смесью вероятностных распределений. После этого предлагается возможность запуска метода скользящего разделения смесей как для исходной выборки, так и для разностей.

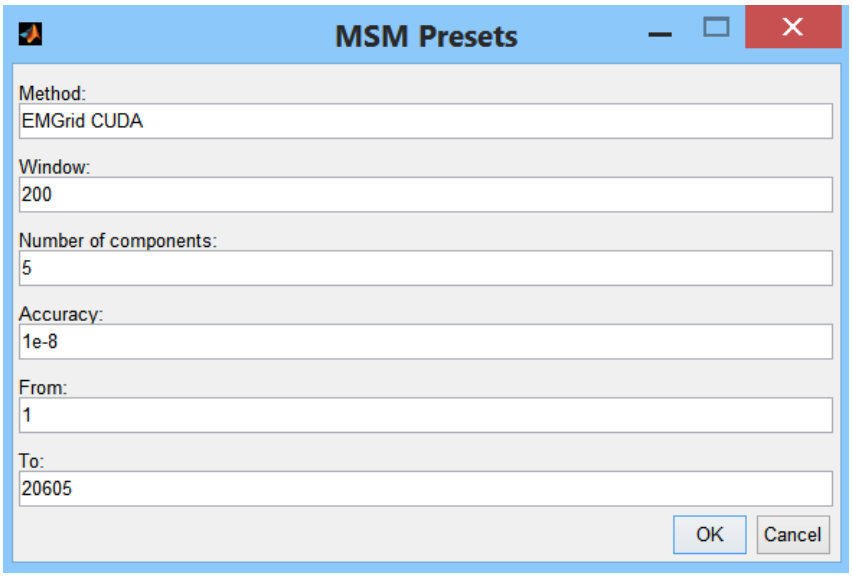

Рис. 7.2. Пример: настройки параметров CPC-метода

Настройка параметров CPC-метода осуществляется с помощью нового диалогового окна (см. рисунок 7.2), в котором задается название вычислительного алгоритма, размер подвыборки, максимальное число компонент в аппроксимирующей смеси, точность приближений и диапазон для сдвига (можно выбрать любую часть внутри выборки; если

заданные пользователем значения некорректны, расчет производится от первого элемента выборки до $N - width + 1$, где N – длина всей выборки для анализа, а $width$ – ширина окна). Для каждого из полей предусмотрены значения по умолчанию. Данные действия осуществляются функцией `InputDialog` (см. алгоритм 7.1).

После нажатия на кнопку «`OK`» и запуска шагов вычислительнго алгоритма (функция `EMs`, см. алгоритм 7.1) отображается окно с индикатором выполнения и названием выбранного вычислительного метода, позволяющего следить за текущим положением окна и состоянием работы программы. После завершения вычислений появляется график с динамической и диффузионной компонентами волатильности, а также диалоговое окно, с помощью которого можно настроить диапазон вывода параметров для сохранения (функция `PlotDialog`, см. алгоритм 7.1), включая и диапазон для окон (см. рисунок 7.3).

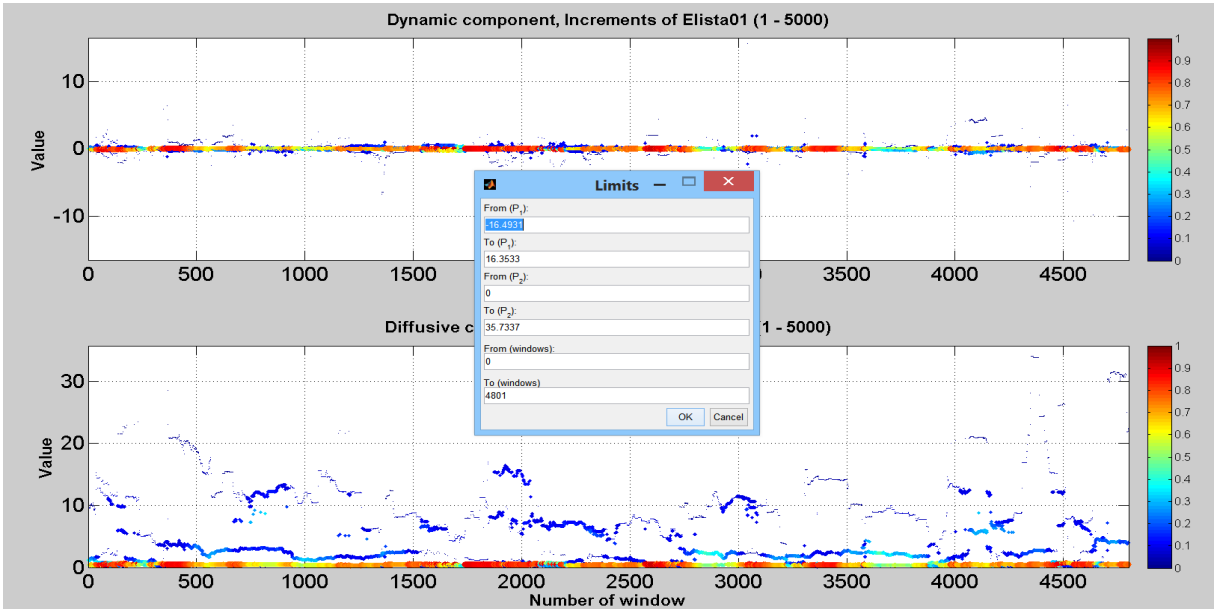

Рис. 7.3. Пример: графический вывод с окном изменения диапазонов по каждой из осей

Найденные оценки сохраняются в процессе расчетов, а также по окончании работы программы, что позволяет избежать потери результатов при возникновении прерывания. После закрытия диалогового окна график автоматически сохраняется в формате `PNG` на диск в папку с программой. Для графического вывода функция `FastPlotter` (см. алгоритм 7.1), которая реализует оптимизированную процедуру визуализации параметров, которая позволяет сохранить все заданные характеристики отображения параметров, но при этом не использовать поточеч-

ный вывод значений динамической и диффузионной компонент – эта процедура слишком требовательна к ресурсам компьютера, в то время как `FastPlotter` быстро и без существенной загрузки центрального процессора и оперативной памяти осуществляет отрисовку даже длинных рядов, в частности, размером более 100000 наблюдений, для трехчетырехкомпонентных смесей (свыше миллиона точек различных цветох и размеров на каждом из графиков).

Также реализован модуль визуализации моментных характеристик (см. раздел 3.1) и квантилей, который содержит функции для отыскания математического ожидания, дисперсии, коэффициента асимметрии и эксцесса, а также квантилей различного уровня для конечных смесей нормальных законов (см. функции `Moments` и `Quantiles`, алгоритм 7.1). Пример трехмерного вывода эволюции квантилей различного уровня приведен на рисунке 7.4.

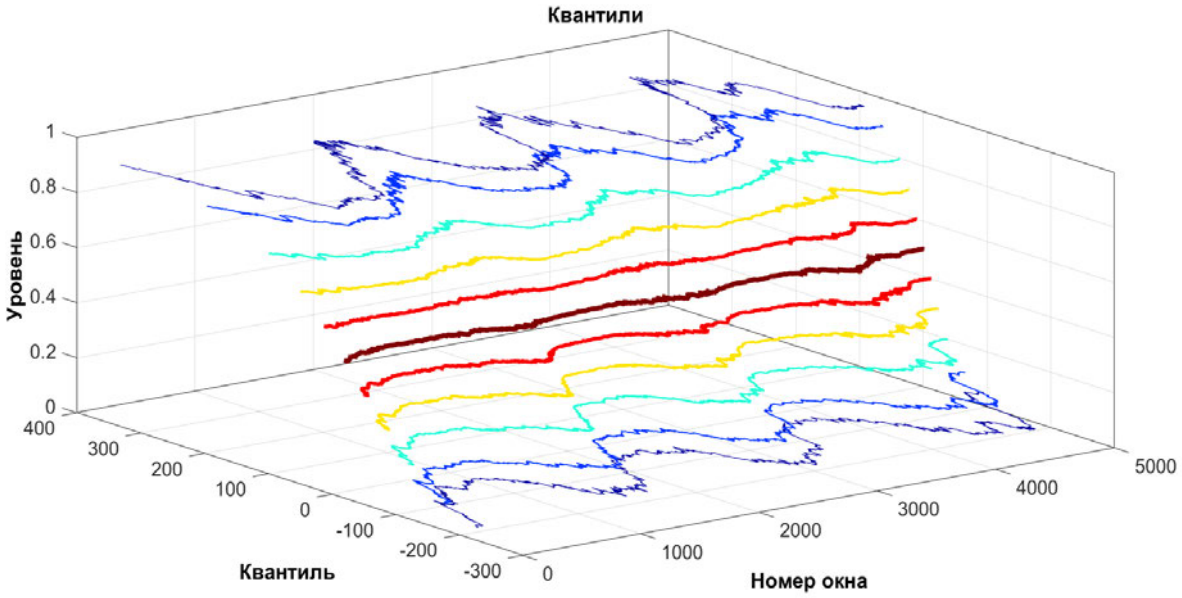

Рис. 7.4. Пример: эволюции квантилей уровней от $0,05$ до $0,95$

7.1.2 Графический пользовательский интерфейс

Описанный в предыдущем разделе оконный интерфейс обладает широкими функциональными возможностями, в рамках которого реализованы практически все возможные варианты СРС-анализа. В то же время, достаточно удобно использовать графические пользовательские интерфейсы, которые, безусловно, являются более подходящими для ши-

рокой пользовательской аудитории. Реализация подобного решения продемонстрирована в виде псевдокода в алгоритме 7.2.

Алгоритм 7.2. CPC-метод с графическим пользовательским интерфейсом

```
1: function MSMGUI( )
2:     Handles←GUI( ); // Порождение GUI системными средствами
3:     Objects←CreateObjects(Handles); // Создание объектов GUI
4:     // Создание надписей на полях формы
5:     SetProperties(Handles, Objects);
6:     // Настройка интерактивных действия полей формы GUI
7:     Callback(Objects);
8:     return ;
```

Перейдем к описанию возможностей средства визуализации результатов для CPC-метода. Начальный экран, формируемый в результате вызовов функций `GUI` и `CreateObjects` (см. алгоритм 7.2) при запуске приложения, представлен на рис. 7.5.

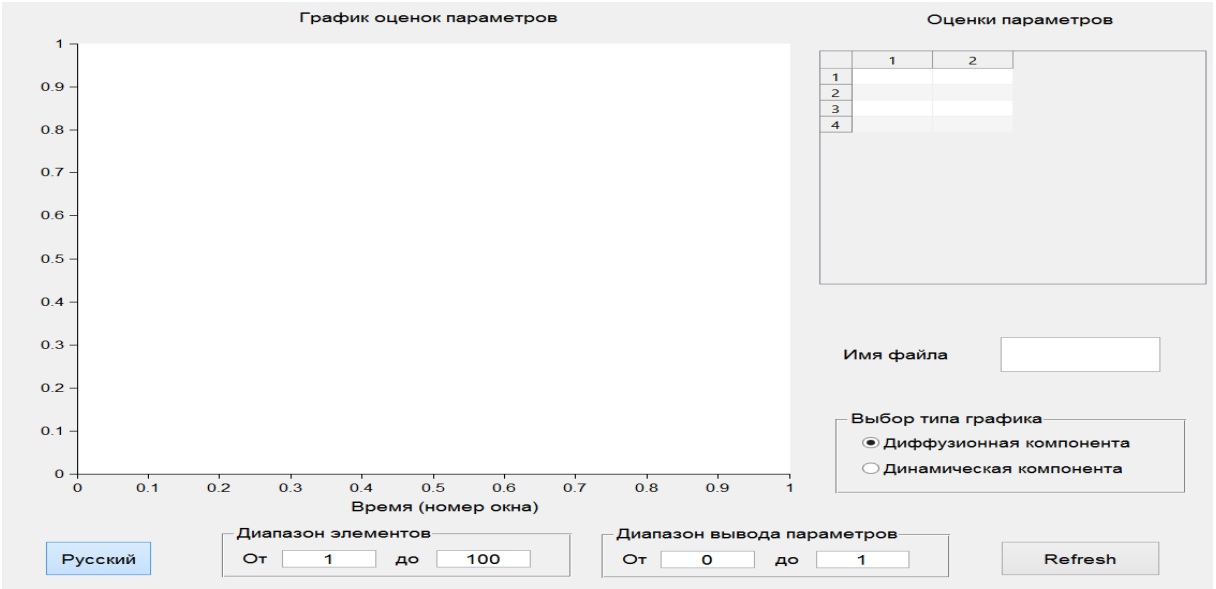

Рис. 7.5. Пример: начальный экран приложения

Область «`График оценок параметров`» (в англоязычном интерфейсе «`Figure`») предназначена для непосредственной визуализации оценок, полученных с помощью какого-либо алгоритма в CPC-методе. В начале работы обе оси размечены в диапазоне от 0 до 1 с шагом 0,1, однако при отрисовке актуального графика обозначения автоматически адаптируются к данным. Кнопка с надписью «`Русский`»/«`English`» позволяет

выбирать язык интерфейса (по умолчанию установлено отображение на русском языке, нажатие на кнопку изменяет все надписи на англоязычные варианты), содержимое остальных полей при переключении не изменяется, область вывода графика не перерисовывается, не происходит повторного вывода значений оценок.

Блок «`Диапазон элементов`»/«`Gap for windows`» задает область вывода оценок по временной оси, соответствующей количеству сдвигов окна в CPC-методе. Например, если есть необходимость рассмотреть крупнее отдельную область или анализируемый ряд слишком большой (оценки сливаются), то можно отобразить только часть данных. В качестве значений по умолчанию используется диапазон от первого до сотого элемента.

Блок «`Диапазон вывода параметров`»/«`Parameter gap`» задает область вывода значений оценок. Для разных данных получаются разные по порядку оценки, кроме того, динамическая компонента может принимать и отрицательные значения. Для удобства масштабирования и корректности вывода параметров и предназначен данный блок. В качестве значений по умолчанию используется диапазон от 0 до 1.

Таблица «`Оценки параметров`»/«`Estimations of parameters`» отображает полный набор оцененных параметров из блока «`Диапазон элементов`», который хранится в файле, адрес (имя) которого задается в поле «`Имя файла`»/«`Filename`».

По умолчанию предполагается, что отображается диффузионная компонента («`Diffusive component`»), однако в блоке «`Выбор типа графика`»/«`Type of figure`» это можно изменить, выбрав динамическую компоненту («`Dynamic component`»). Нажатие кнопки «`Refresh`» осуществит корректное обновление (либо первичное изображение) графика в соответствии с выбранными настройками с помощью специально разработанного для CPC-метода алгоритма рисования (фактически это функция `FastPlotter`, см. алгоритм 7.1).

На рисунках 7.6 и 7.7 изображены диффузионная и динамическая компоненты для некоторых рядов `Params` и `P1`, соответственно. В первом случае вывод осуществляется для положений окон от 4000 до 9000, при этом границы параметров установлены от 0 до $1,1 \cdot 10^{-4}$, а во втором – от первого положения окна до значения 4800, при этом границы параметров определяются диапазоном от $-0,08$ до $0,08$. Рассматриваемые случаи различаются и числом компонент, которое было использовано при аппроксимации (2 и 6, соответственно) – это проявляется, в частности,

в разных размерах таблиц оценок параметров справа от графика. Отметим, что на рисунке 7.7 приведен англоязычный вариант интерфейса.

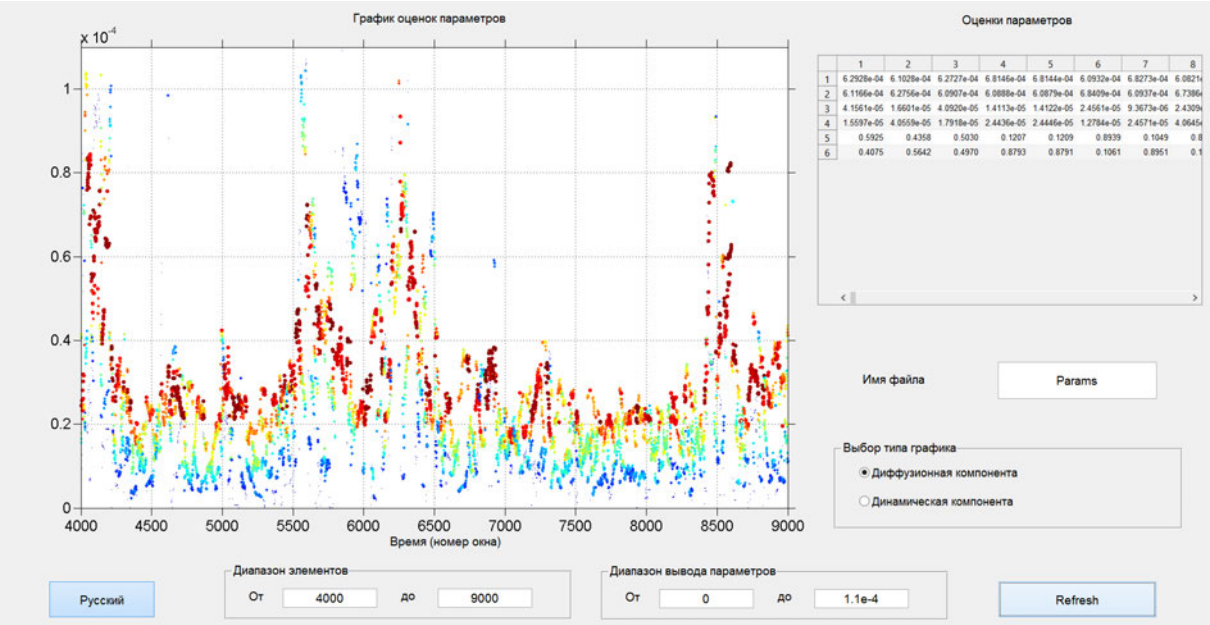

Рис. 7.6. Пример: диффузионная компонента

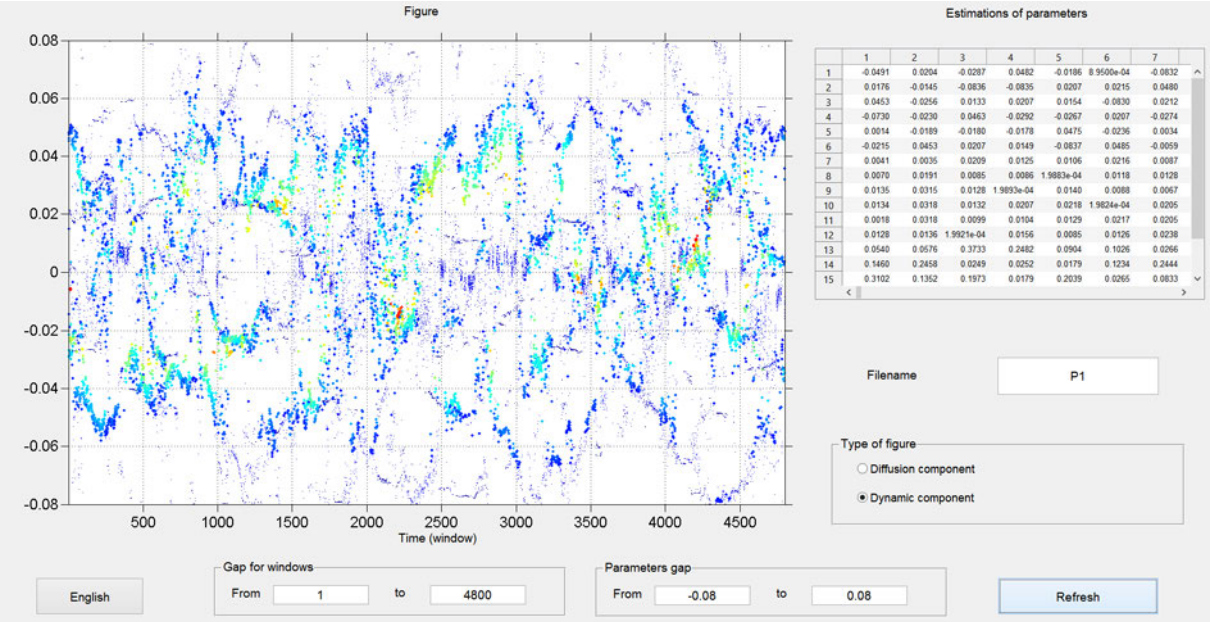

Рис. 7.7. Пример: динамическая компоненты

7.1.3 Динамическая визуализация

В завершении данного раздела рассмотрим еще один весьма наглядный способ визуализации эволюции параметров аппроксимирующей сме-

си в CPC-методе, а именно, видео (см. скриншот на рисунке 7.8).

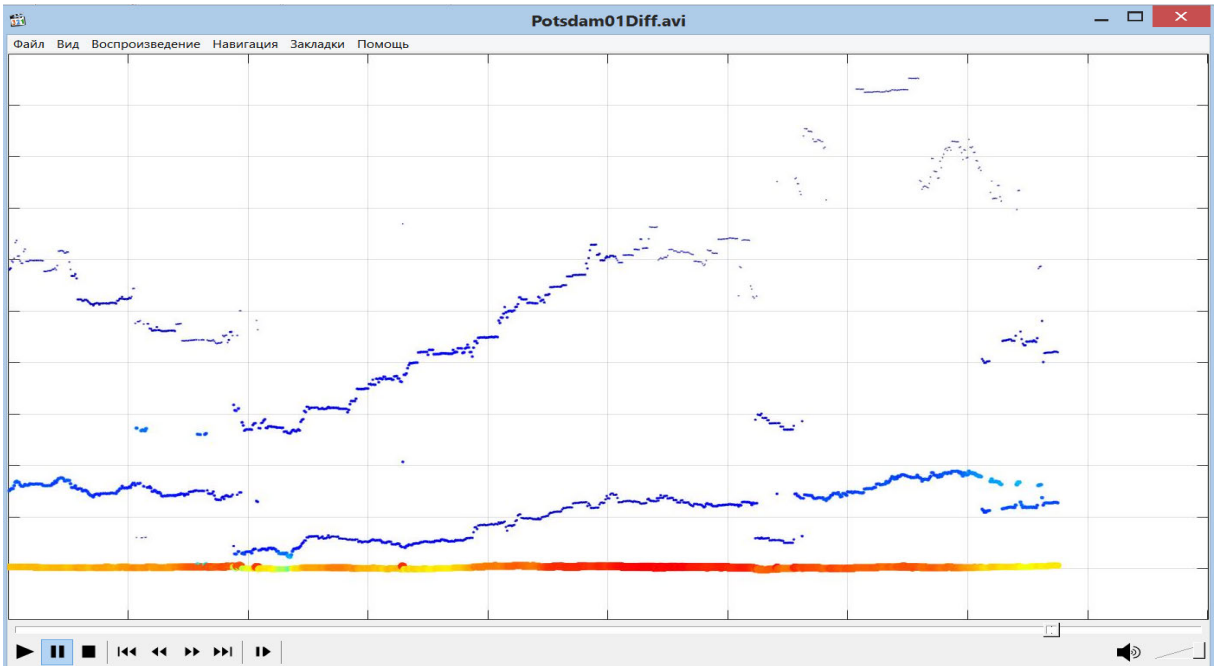

Рис. 7.8. Пример: динамическая визуализация эволюции компонент

Для создания подобных видео использована функция `VideoWriter` пакета `MATLAB`. С помощью аналога функции `FastPlotter` (см. алгоритм 7.1) производится рисование параметров для каждого шага CPC-метода, записываемых во фреймы, из которых формируется файл в формате `AVI`. Данный способ позволяет проследить, как компоненты волатильности появляются, исчезают и эволюционируют в течение всего процесса наблюдений за данными.

7.2 Приложение для анализа распределений длительностей и объемов осадков

В этом разделе будет рассмотрено программное решение, реализующих функциональные методы оценивания параметров обобщенных отрицательных биномиальных и гамма-распределений, описанные в разделе 6.3. В отличие от описанного в разделе 7.1.2 графического интерфейса, разработка которого частично базировалась на принципах визуального программирования (данный шаг в алгоритме 7.2 инкапсулирован в функциях `GUI` и `CreateObjects`), создание всех форм в данном случае производилось полностью с помощью программного кода `MATLAB` без использования инструментов `GUIDE` или `App Designer`.

7.2.1 Программная реализация

Как уже было упомянуто, все объекты и интерактивные компоненты интерфейса реализованы программно без привлечения средств `GUIDE` или `App Designer`. Отметим следующие ключевые преимущества данного подхода:

– в форме возможен вывод графиков любого типа, включая гистограммы и анимированные изображения – в `App Designer` вплоть до актуальной версии `MATLAB R2020a` для подобных объектов поддержка отсутствует;

– нет необходимости создавать специальный `FIG`-файл, в котором производится размещение элементов при визуальном программировании, как это реализуется при вызове окружения `GUIDE` – подобные файлы для нетривиальных приложений могут быть весьма большими, в то время как программный код занимает малую часть памяти и быстро исполняется.

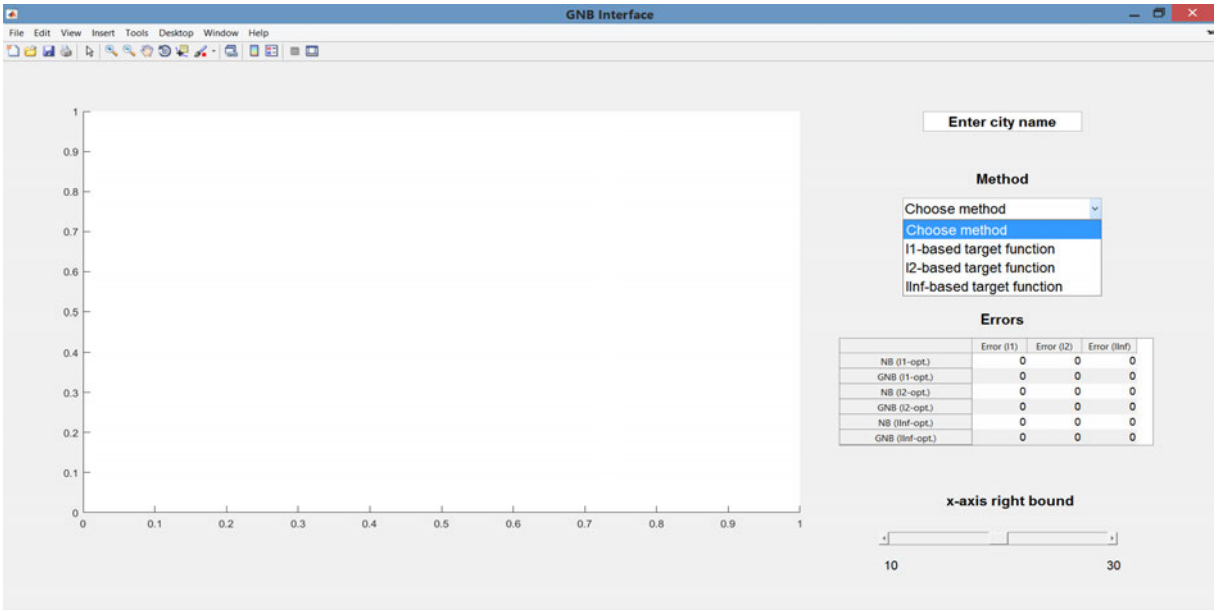

Рис. 7.9. Начальный экран приложения

Используемый подход требует только создания объекта `figure`, на котором размещаются интерактивные компоненты с помощью вызовов функции `uicontrol` пакета `MATLAB` с различными параметрами, ключевым из которых является `Style`. Перечислим основные настройки, используемые для создания графического интерфейса приложения (пример начального экрана привдеен на рисунке 7.9):

– `edit` для создания редактируемых текстовых полей (см., например,

рисунок 7.9 справа наверху), с помощью которых можно задавать название города, осадки в котором необходимо проанализировать: используется для открытия сохраненного на жестком диске файла с данными;

– `popupmenu` для создания дающих меню (см. рисунок 7.9 справа посередине формы) для выбора варианта метрик, в которых должна осуществляться оптимизация для поиска неизвестных параметров (ℓ^1, ℓ^2 и ℓ^∞ для обобщенного отрицательного биномиального и L^1, L^2 и L^∞ для обобщенного гамма-распределений): это влияет на параметры вызова соответствующих вычислительных функций, а также на настройки графиков при визуализации;

– `text` для указания имен объектов на форме (например, `Method` для выпадающего меню или `Errors` для таблицы).

– `slider` для создания ползунка для задания области вывода гистограмм (точнее – соответствующей правой границе графика, создаваемого с помощью функции `axes`, см. центральную область на рисунке 7.9): видимая область может быть изменена за счет сдвигания ползунка; в качестве допустимых границ используется диапазон $[10, 30]$ (шаг изменения – единица), значение по умолчанию – 20.

Под ползунком на рисунке 7.9) размещена таблица ошибок аппроксимации, создаваемая с помощью функции `uitable`. Столбцы таблицы можно изменять, например, перетаскивая с помощью мыши заголовки столбцов. Элементы таблицы инициализируются нулями, а после проведения расчетов автоматически заполняются полученными значениями. Реализация приложения представлена в виде псевдокода в алгоритме 7.3.

Алгоритм 7.3. Статистический анализ распределений длительностей дождливых периодов с графическим пользовательским интерфейсом

```
1:  function GNB_GG_Interface( )
2:      Fig←CreateFigure( );              // Создание объекта figure
3:      // Инициализация переменных, задание типа распределения
4:      options←Init(Fig, GNB, GG);
5:      // Создание объектов на форме
6:      Objects←CreateForm(Fig, uicontrol, Styles, uitable);
7:      // Настройка интерактивного взаимодействия
8:      Params_est←Callback(Fig, Objects, Approx, options);
9:      // Approx() – функция аппроксимации параметров
10:     return Params_est;
```

Данное приложение имеет две модификации – `GNBInterface` и

`GGInterface`, – в которых используемые инструменты разработки и получаемый интерфейс аналогичны (строки 2–6 в алгоритме 7.3), однако существенным образом отличаются модули вычисления параметров, поэтому для унификации данный момент инкапсулируется внутри функции `Approx`, а для коректного определения ситуации служат параметры `GNB` и `GG` (см. строку 8).

Оптимизационные процедуры основаны на симплекс-методе Нелдера–Мида [307] со значением точности для останова 10^{-12} при 3000 максимально допустимых итераций. Неизвестные параметры распределений оцениваются в соответствии с алгоритмами 6.5 и 6.6 (см. раздел 6.3). Отметим, что размер файла с интерфейсами и функциями обработки данных составляет всего 30 КБ.

7.2.2 Примеры использования

Рассмотрим некоторые конкретные примеры использования разработанного приложения для данных об осадках в Потсдаме, которые подробно изучались в разделе 6.3.

На рисунках 7.10–7.12 представлены примеры работы приложения `GNBInterface` для оценивания параметров распределений, аппроксимирующих длительности дождливых периодов с использованием метрик ℓ^1 (рисунок 7.10), ℓ^2 (рисунок 7.11) и ℓ^∞ (рисунок 7.12).

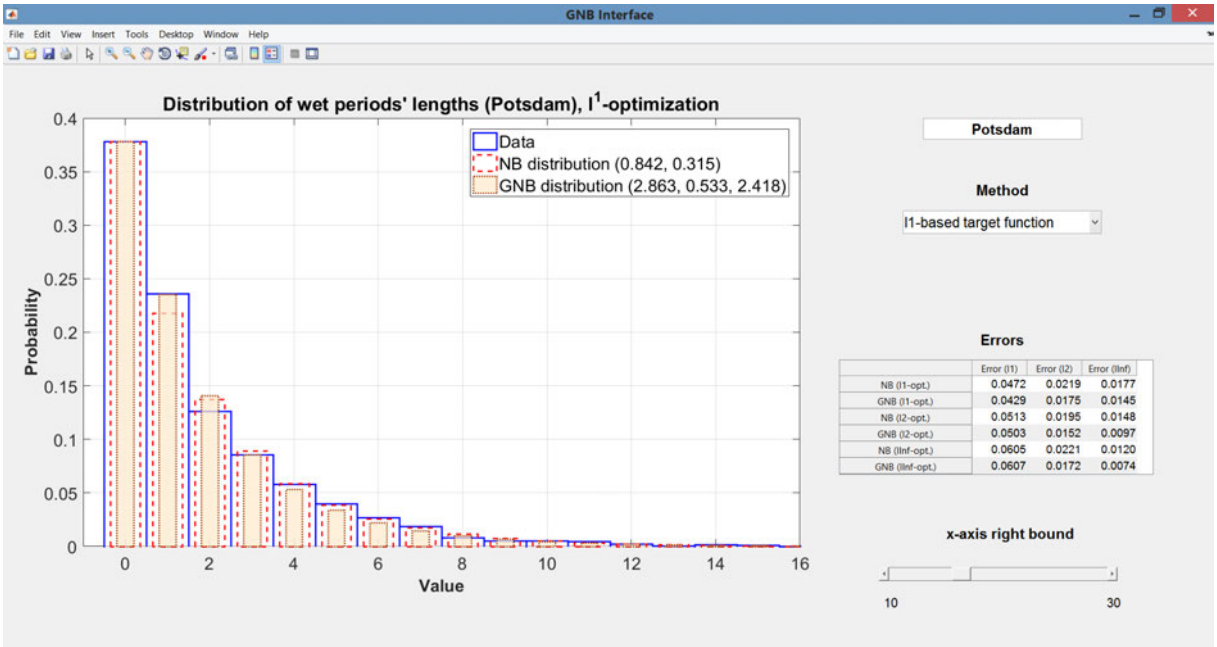

Рис. 7.10. Аппроксимация распределения длительностей дождливых периодов на основе метрики ℓ^1

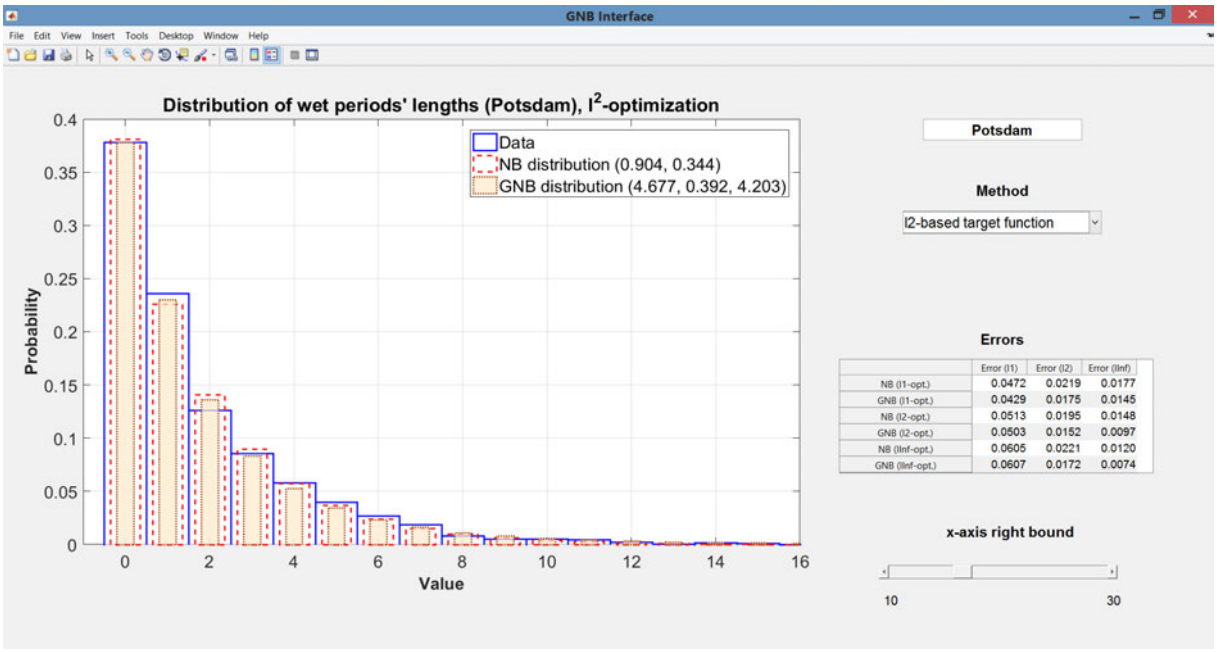

Рис. 7.11. Аппроксимация распределения длительностей дождливых периодов на основе метрики ℓ^2

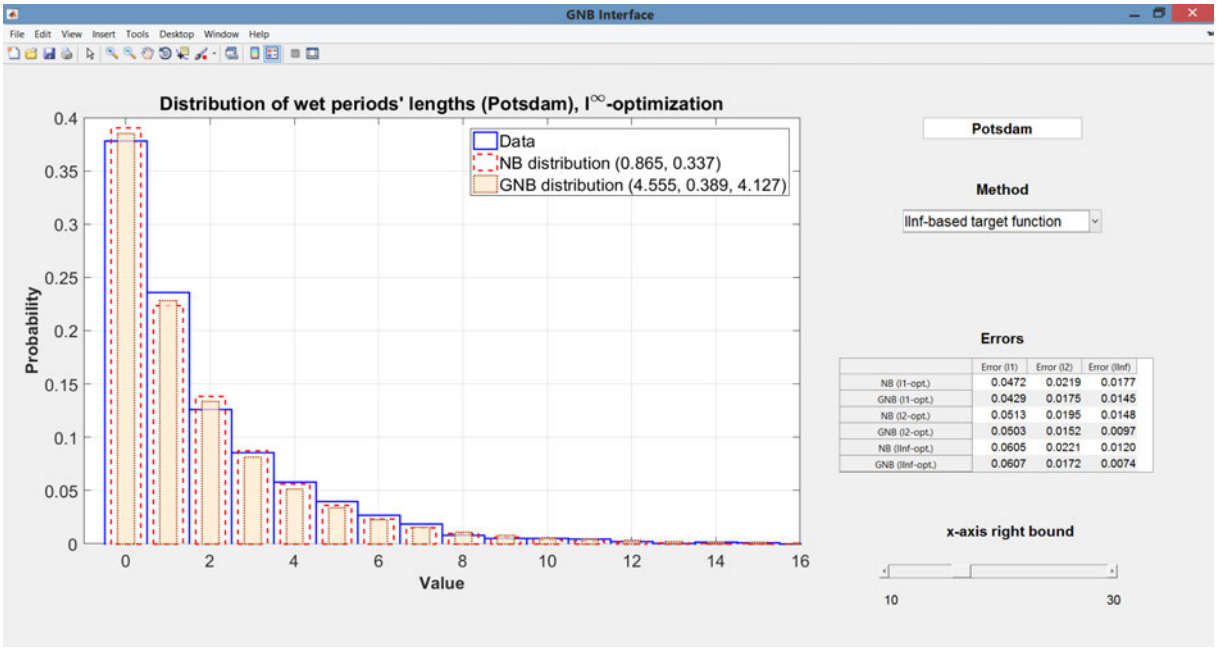

Рис. 7.12. Аппроксимация распределения длительностей дождливых периодов на основе метрики ℓ^∞

На приведенных рисунках изображены гистограммы для исходных данных, а также приближающих обобщенного и классического отрицательных биномиальных распределений. Область вывода по оси абсцисс средствами интерфейса задана как $[0, 16]$, соответствующий метод оптимизации отмечен в выпадающем меню `Method`. При этом, если параметры были оценены ранее, то повторные вычисление не производятся, при

этом они, а также величины ошибок в таблицах `Errors` загружаются из соответствующих `XLSX` файлов.

На рисунках 7.13–7.15 представлены примеры работы приложения `GNBInterface` для оценивания параметров распределений, аппроксимирующих объемs осадков за дождливые периодов с использованием метрик L^1 (рисунок 7.13), L^2 (рисунок 7.14) и L^∞ (рисунок 7.15).

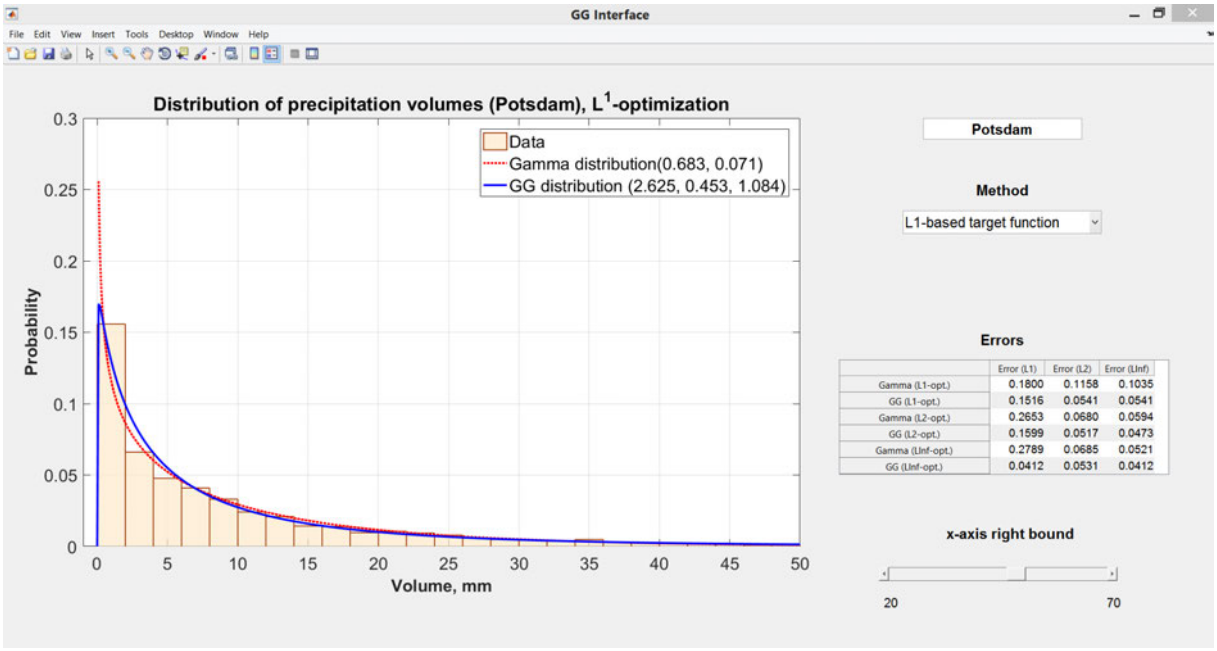

Рис. 7.13. Аппроксимация распределения объемов осадков за дождливые периоды на основе метрики L^1

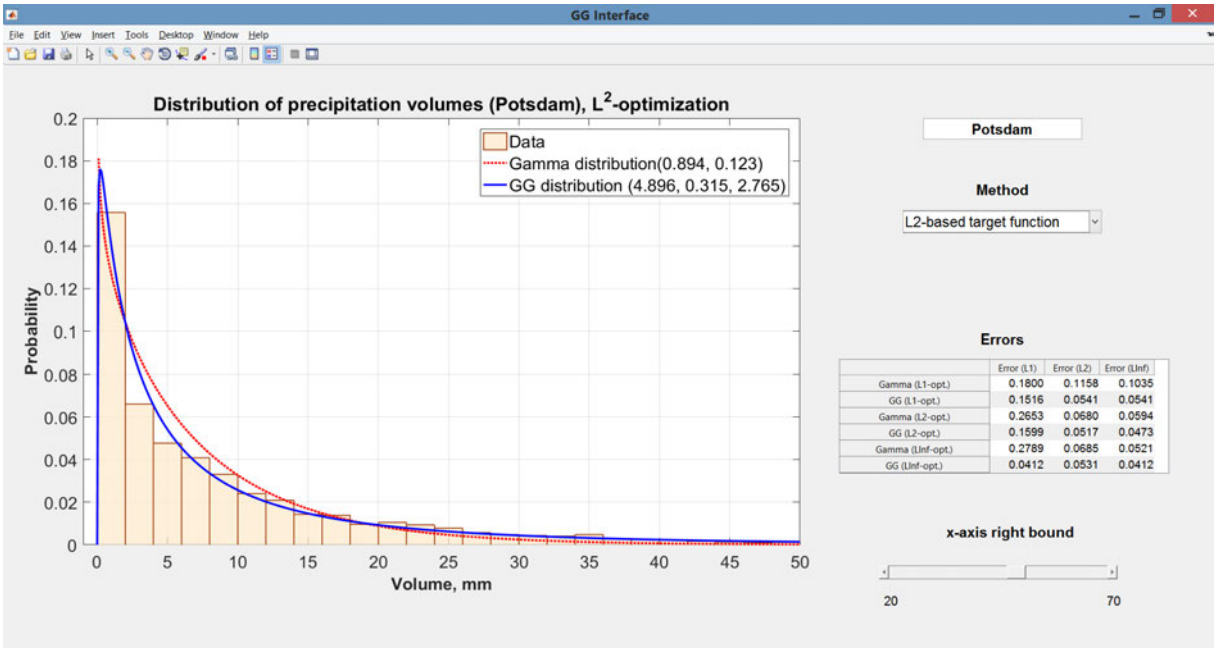

Рис. 7.14. Аппроксимация распределения объемов осадков за дождливые периоды на основе метрики L^2

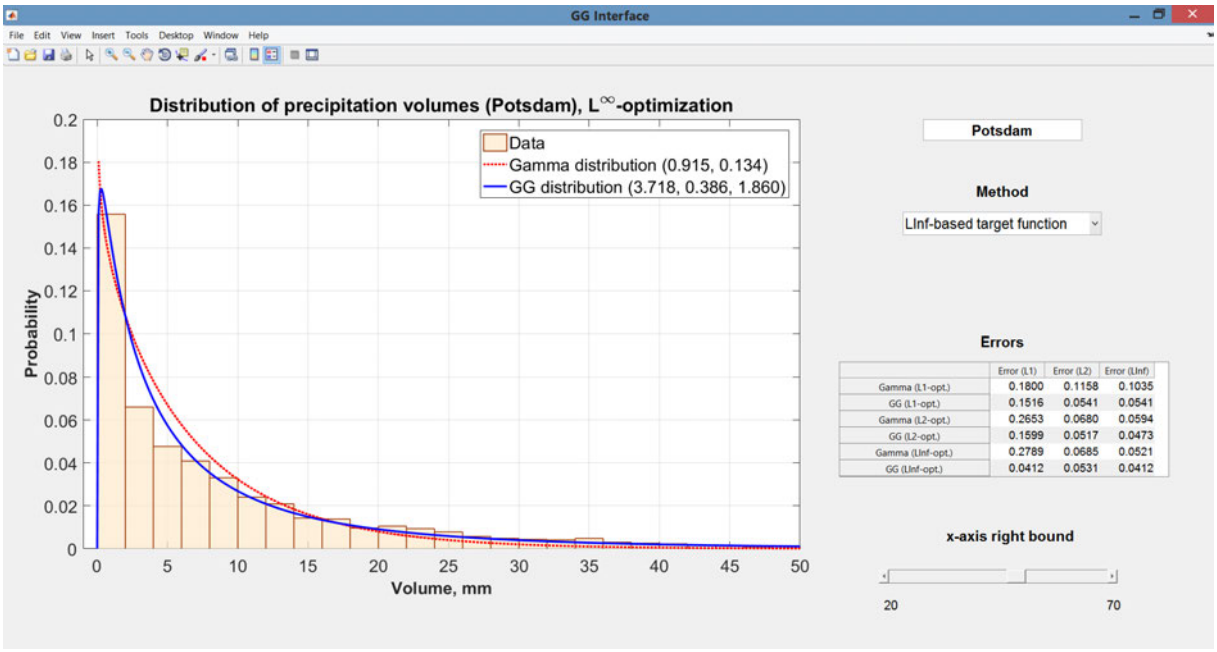

Рис. 7.15. Аппроксимация распределения объемов осадков за дождливые периоды на основе метрики L^∞

На приведенных рисунках изображены гистограммы для исходных данных, а также приближающих обобщенного и классического гамма-распределений. Область вывода по оси абсцисс средствами интерфейса установлена как $[0, 50]$, соответствующий метод оптимизации отмечен в выпадающем меню `Method`.

Разработанные программные решения позволяют получать представление результатов в наглядной форме для быстрого и удобного анализа данных. Данные инструменты использовались, в частности, при обработке осадков в Потсдаме и Элисте (см. раздел 6.3).

7.3 Информационная технология исследования стохастических процессов

В разделе 5.1 был представлен вероятностный бутстреп-подход для изучения с помощью спектрального анализа специфических структур, формирующих турбулентность в плазме, на основе конечных смесей различных вероятностных распределений. В этом разделе описана информационная технология, которая реализует данный метод, включая в себя инструменты первичной обработки и подготовки данных для анализа, различные реализации алгоритмов EM-типа, функции для бутстреп-анализа и визуализации результатов. Представленное программное ре-

шение является следующим шагом в направлении развития интерфейсов для разработанных в диссертации методов и алгоритмов анализа данных.

Используемые математические модели и полученные с помощью данного инструмента практические результаты в важном разделе современной физики (например, определение числа формирующих процессов и значений ряда величин), описаны в разделе 5.1. Здесь же рассмотрим вопросы функционирования и взаимосвязей программных модулей, составляющих вычислительную часть информационной технологии (см. рисунок 7.16).

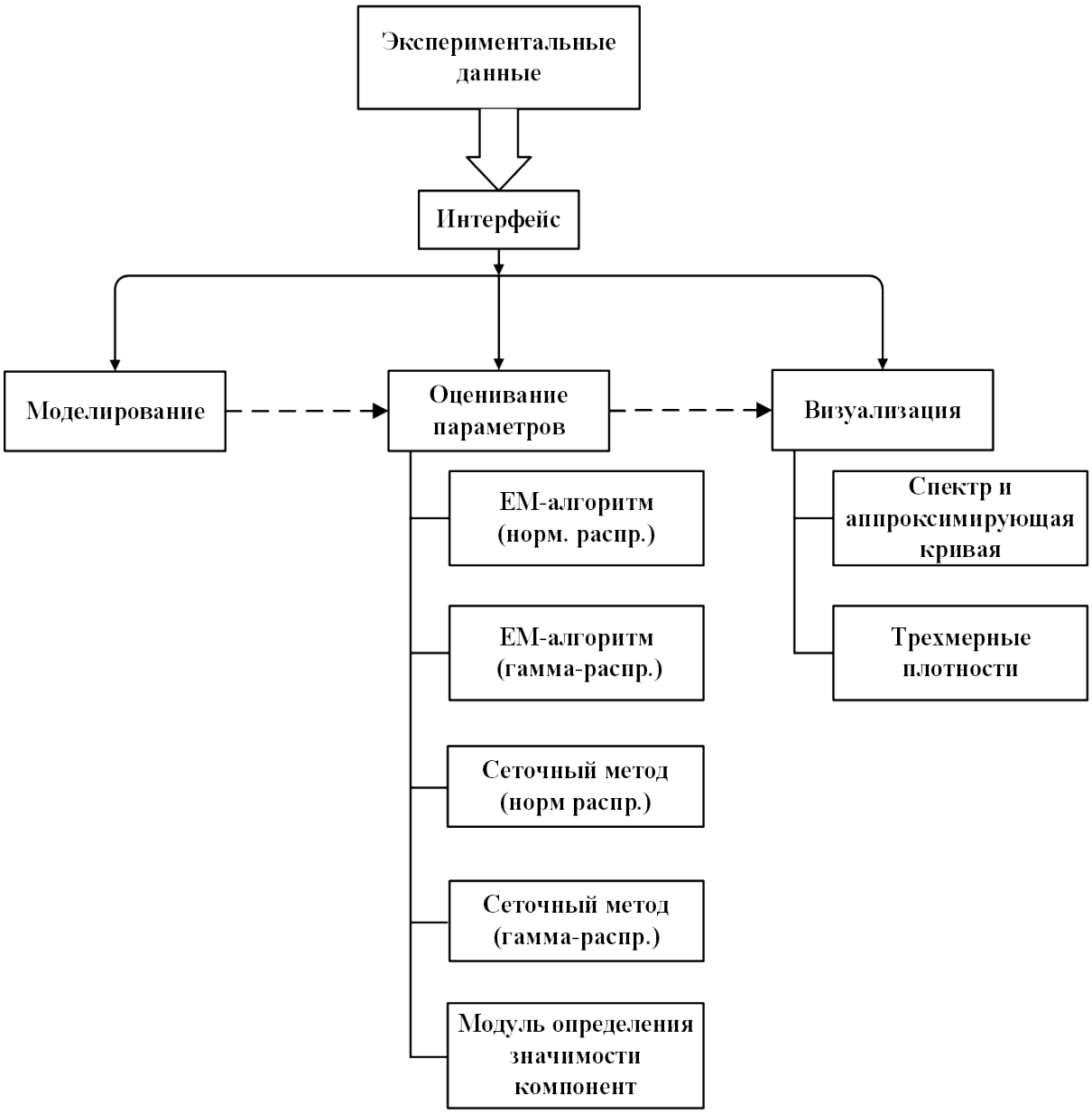

Рис. 7.16. Структура информационной технологии

Разработанная информационная технология включает в себя следующие основные компоненты:

– Блок «`Интерфейс`», в котором реализуются предварительная подготовка экспериментальных данных (загрузка, форматирование для вычислительных процедур, и т. д.), их передача в следующие блоки для обработки, получение результатов и сохранение их на диск.

– Блок «`Моделирование`», отвечающий за создание тестовой выборки для бутстреп-метода (см. раздел 5.1).

– Блок «`Оценивание параметров`», в котором производится аппроксимация спектра конечной смесью вероятностных распределенийс помощью полученной в предыдущем блоке выборке с помощью одного из EM-методов (аналогичная функция вызывается в алгоритме 7.1: см. строку 7 и функцию `EMs` – однако здесь не требуется осуществлять шаги в CPC-методе, ширина окна совпадает с числом всех наблюдений). Примеры некоторых методов приведены непосредственно на рисунке 7.16. Алгоритмы на основе нормального распределения можно использовать для одно- и двусторонних спектров, а на основе гамма-распределении – для односторонних спектров. В этот же логический блок включается `Модуль определения значимости компонент` (его теоретические основы были описаны в разделе 2.2, а тестирование возможностей проведено в статье [7]), который позволяет подобрать корректную смешанную модель при аппроксимации распределения бутстреп-выборки.

– Блок «`Визуализация`» осуществляет графическое отображение полученных результатов в виде спектров и трехмерных плотностей (примеры были рассмотрены ранее в разделе 5.1).

Очевидно, что на первом шаге каждый из этих этапов должен выполняться последовательно, однако возможен переход непосредственно к одному из блоков, если ранее предыдущий был выполнен: предусмотрена работа с данными, которые были ранее сохранены на диске, что позволяет избежать повторного моделирования или отыскания оценок. На рисунке 7.16 это продемонстрировано с помощью пунктирных и сплошных стрелок.

7.4 Онлайн-система вероятностного-статистического анализа данных

Одним из наиболее актуальных трендов современного анализа данных является использование исследователями из различных стран и

предметных областей высокопроизводительных вычислительных решений, в том числе удаленных. В предыдущих разделах были описаны комплексы, реализованные средствами программирования пакета MATLAB, который является проприетарным кросс-платформенным программное обеспечение, а значит, возможность предоставления разработок на его основе естественным образом ограничена. Кроме того, далеко не все алгоритмы могут эффективно выполняться на любой физической архитектуре или даже более того – быть запущены на ней. К таковым можно отнести, например, решения, предназначенные для выполнения вычислений с помощью CUDA. У конечного пользователя может оказаться графическая карта от другого производителя, либо его GPU не предназначена для осуществления вычислений на должном уровне. Поэтому в этом разделе обсуждаются вопросы реализации сходных функциональных возможностей в рамках онлайн-сервиса, поддерживающего гетерогенные вычисления GPGPU.

7.4.1 Архитектура онлайн-сервиса

Для систем онлайн-обработки данных в реальном времени большую важность представляют вопросы, связанные с масштабированием отдельных частей системы. Специфика решения исследовательских задач на основе смешанных вероятностных моделей такова, что вычислительная сложность используемых алгоритмов может быть значительна, хотя для хранения как анализируемых данных, так и результатов требуется умеренный объем памяти (например, типичный объем экспериментальных выборок составляет от нескольких десятков тысяч до миллиона наблюдений). Это соображение приводит к необходимости выделения сервисной и вычислительной частей в архитектуре онлайн-системы. Рассмотрим подробнее ключевые элементы с указанием потоков данных, представленные на рисунке 7.17.

Первый уровень представляет ПК пользователя, обеспечивающий доступ к интерфейсу онлайн-комплекса для загрузки данных, настройки параметров и получения результатов (в том числе, в визуальной форме). В пользовательских профилях, доступ к которым требует регистрации и авторизации для каждого сеанса работы, сохраняются ранее загруженные на сервер данные и полученные результаты обработки, которые доступны для дальнейшего использования и проведения дополнительного анализа. Из-за высокой вычислительной сложности в ряде ситуа-

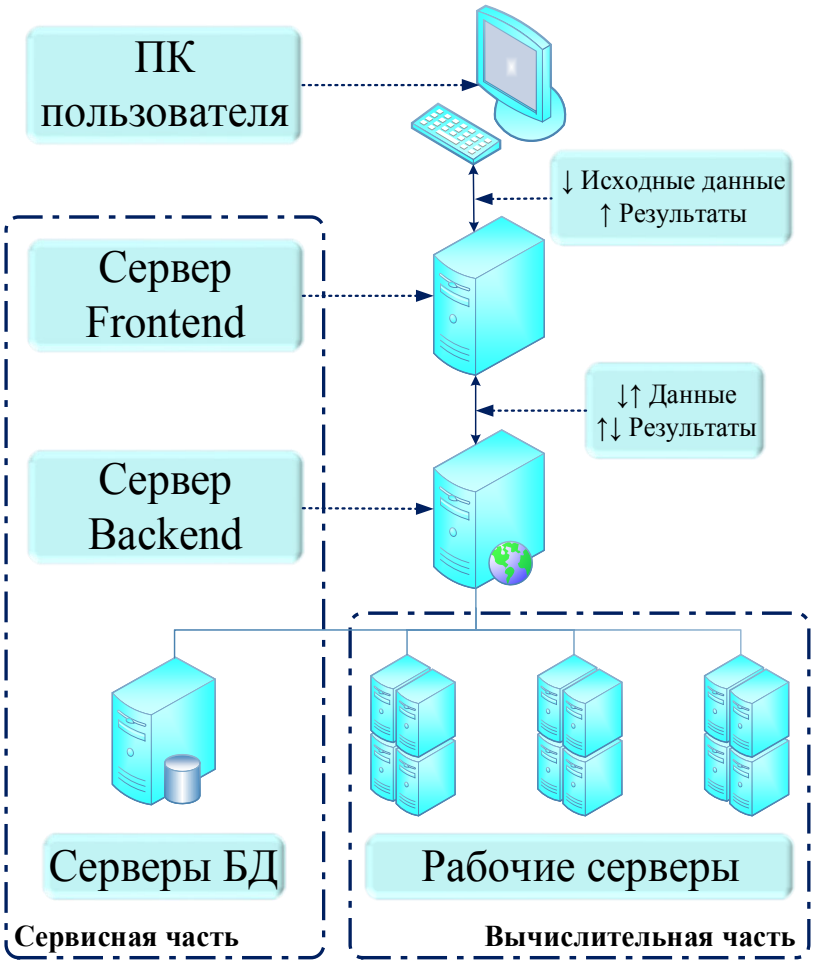

Рис. 7.17. Архитектура онлайн-сервиса и диаграмма потоков данных

ций может потребоваться значительное время для корректного определения параметров выбранной смешанной вероятностной модели. Поэтому предусмотрено оповещение пользователей о статусе процесса с помощью электронных каналов связи. В случае завершения расчетов, пользователь сможет в любое удобное время обратиться к результатам анализа как в числовом (с помощью экспорта оцененных параметров), так и в графическом виде.

Сервисная часть системы состоит из `Frontend`- и `Backend`-компонентов, а также серверов баз данных. С ее помощью реализуется основная бизнес-логика системы, поэтому вычислительные сложности, а также необходимость обработки больших данных отсутствуют. Решение задач масштабирования для сервисной части требуется только в случае серьезной пиковой активности пользователей, что является маловероятным сценарием при штатной работе такой системы, ориентированной на поддержку проведения научных исследований.

Frontend-сервер предназначен для реализации интерфейса взаимодействия между пользователем и вычислительными узлами системы, при этом непосредственная обработка данных на нем не производится. Backend-сервер осуществляет взаимодействие между Frontend-сервером, рабочими серверами и серверами баз данных, в частности, готовит данные для обработки и распределяет задачи по вычислительным компонентам системы. Взаимодействие Backend- и рабочих серверов должно осуществляться с помощью специализированных API для реализации возможностей, связанных с параллельной обработкой данных. После проведения вычислений для одного положения окна CPC-метода данные сохраняются на сервере БД и пересылаются с помощью Frontend-сервера пользователю для возможного контроля процесса выполнения анализа с его стороны. Очевидно, что наибольшая нагрузка по передаче данных ложится на взаимодействие Frontend- и Backend-серверов, поэтому необходима разработка механизмов ускорения их работы (в частности, за счет кэширования).

Вычислительная часть системы получает входные наборы данных и начальные параметры для методов, обеспечивает работу алгоритмов EM-типа, возвращает оценки неизвестных параметров смешанных вероятностных моделей. При этом детали внутренней реализации вычислительной части скрыты для сервисной: пользователи не могут работать с ней напрямую, всё взаимодействие производится исключительно за счет вызовов с Backend-сервера.

При увеличении нагрузки на систему или при создании приоритетной заявки на обработку данных в облачном сервисе может быть создан отдельный виртуальный сервер вычислительной части. После успешного запуска он добавляется в пул существующих обработчиков, хранящийся на Backend-сервере. При уменьшении нагрузки и отсутствии соответствующих задач данный обработчик удаляется из пула с целью экономии ресурсов.

Подобная архитектура позволит интегрировать в систему различные методы интеллектуального анализа данных, при этом их реализация может быть основана на специальных программных и аппаратных подходах, но для конечного пользователя особенности решений будут скрыты. Это достигается за счет реализации доступа к ним посредством соответствующих API. Кроме того, функции-обработчики могут быть размещены на иных серверных мощностях, нежели сервисная часть системы. Это, в частности, способствует увеличению стабильности работы систе-

мы с точки зрения конечного пользователя – сбой в работе одного из вычислительных серверов не сказывается на функционировании служебной части системы. Очевидна и возможность потенциального сокращения затрат на дорогостоящее оборудование без потери общего качества обслуживания.

7.4.2 Пользовательский интерфейс онлайн-сервиса

Основным элементом интерфейса является область графического вывода, предназначенная для отображения исходного временного ряда, его модификаций в результате предобработки, а также вывода результатов CPC-метода , включая вывод нескольких графиков одновременно (динамическая и диффузионная компоненты волатильности, моментные характеристики и др.). Поддерживается динамическое масштабирование для графиков в зависимости от настроек браузера.

Пользователю для каждого ряда (в том числе и модифицированного некоторым способом) предлагаются следующие инструменты:

– загрузка выборки для анализа в форматах `CSV` и `TXT`, а также экспорт результатов (в том числе сохранение изображений в формате `PNG`);

– отображение гистограммы для данных и экспорт в формате `PNG`;

– дублирование исходного ряда;

– переход к разностям (в том числе логарифмическим);

– отыскание выборочных моментных характеристик (математическое ожидание, дисперсия и т. п.);

– запуск классического EM-алгоритма (для всех параметров – ширина окна, точность итерационных приближений, величина сдвига и т. п. – предлагаются значения по умолчанию, которые основаны на предварительном анализе некоторого укороченного участка ряда с целью получения соответствующих начальных приближений);

– запуск сеточного EM-алгоритма (в том числе в `CUDA`-версии – в зависимости от доступных обработчиков);

– удаление любых данных (полное, выборочное) из области графического анализа.

На рисунке 7.18 продемонстрирован пример интерфейса сервиса, в котором для некоторого временного ряда выводятся математическое ожидание и дисперсии (верхний график), коэффициенты асимметрии и эксцесса (нижний график). При реализации были использованы формулы для вычисления моментных характеристик (см. раздел 3.1).

Рис. 7.18. Пример интерфейса онлайн-системы

Функциональное наполнение данного сервиса может существенным образом модифицироваться (безусловно, с внесением соответствующих изменений в интерфейс для предоставления новых инструментов), однако пользователю не придется устанавливать новое программное обеспечение, решать вопросы совместимости ранее написанного и обновленного программного кода и т. д. Безусловно, существенным образом повышается удобство использования сложными математическими моделями для не-специалистов. При этом необходимо упомянуть и ситуацию с менее гибкими настройками для профессионалов в области анализа данных, которые ориентированы на адаптацию методов для эффективной работы с конкретными наборами данных.

7.5 Сервисы научно-образовательных цифровых платформ

Данные в цифровом формате, инновационные принципы работы с которыми способствуют формированию информационного пространства с учетом потребностей граждан и общества, а также новой технологической индустрии составляют основу цифровой экономики. Одним из ее ключевых компонентов являются платформы и технологии, создающие

компетенции для развития рынков и отраслей экономики, а также новой среды для эффективного взаимодействия различных субъектов. Цифровые платформы предоставляют единую информационную среду для всех сторон за счет передовых ИТ-решений для сокращения транзакционных издержек. Они ориентированы на упрощение процедуры решение задач анализа, оптимизации и перестройки связей между участниками, а также позволяют создавать новые продукты и услуги, формировать экосистемы. Подобный подход в полной мере соответствует идеологии Индустрии 4.0 [118], в том числе и для научно-образовательной отрасли.

Как было отмечено во введении, одной из ключевых современных научных парадигм является интенсивный анализ огромных объемов накопленных в предметных областях данных [188]. Для их аналитической обработки требуется создание новых методов и подготовка специалистов с новым набором компетенций, которые не могут ограничиваться одной исследовательской группой или образовательным коллективом. Одним из наиболее эффективных инструментов решения данной задачи как раз и служат научно-образовательные цифровые платформы. Весьма перспективным представляется их создание как для поддержки индивидуальных исследовательских проектов, так для решения прорывных задач [202] на основе так называемых центров превосходства [78], осуществляющих прорывные фундаментальные и прикладные исследования в наиболее важных и инновационных областях знания, обладающих уникальными интеллектуальными и материально-техническими ресурсами. Прорывные достижения в рамках научных исследований стимулируют развитие технологических инновационных процессов, формирование и задействование новых компетенций,начинает проявляться эффект от трансфера технологий. Это позволяет сочетать вклад в фундаментальную науку с формированием новых инновационных отраслей [418].

Рассмотрим опыт центра превосходства в области компьютерных наук на базе ФИЦ ИУ РАН [80], который ориентирован на опережающее преодоление больших вызовов. На рисунке 7.19 продемонстрирована взаимосвязь исследовательских направлений ФИЦ ИУ РАН и задач цифровой экономики Российской Федерации. Очевидно, что результаты всех ключевых исследовательских направлений вносят вклад в различные области цифровой экономики в виде математических моделей, технологий, компетенций, развития кадрового потенциала и т. п. В том числе, безусловно, это касается и широкого круга вопросов по созданию и внедрения цифровых платформ и развитию соответствующих отрасле-

вых экосистем, а также инструментов, решений и необходимых ресурсов для ведения деятельности в цифровом пространстве.

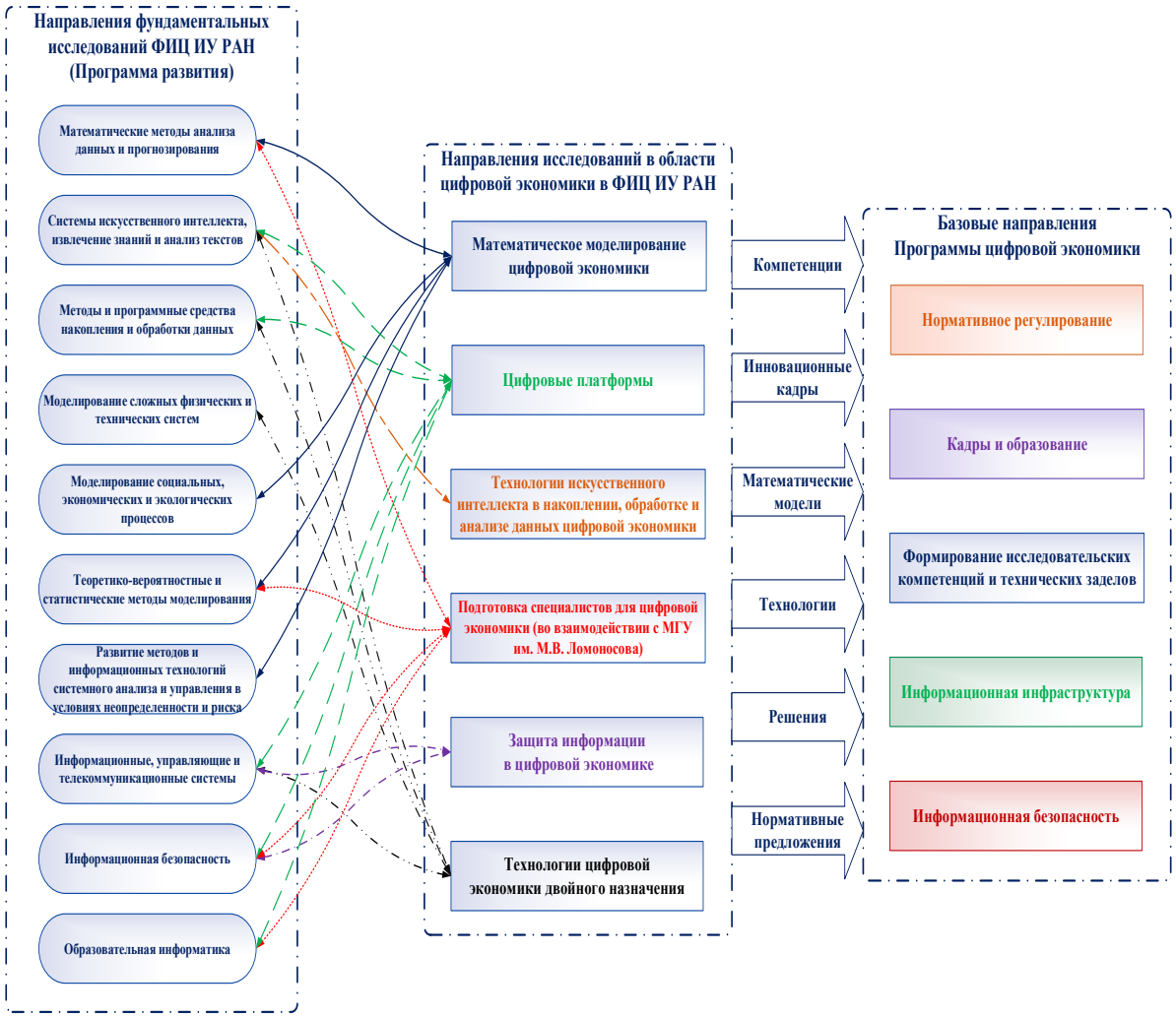

Рис. 7.19. Взаимосвязь исследовательских направлений ФИЦ ИУ РАН и задач цифровой экономики Российской Федерации

7.5.1 Цифровая система управления сервисами научной инфраструктуры

В создаваемой научно-образовательной платформе ФИЦ ИУ РАН значительное внимание уделяется цифровой системе управления сервисами (СУС) научной инфраструктуры, в том числе сложным и уникальным научным оборудованием, в целях поддержки и технологического обеспечения исследований, разработок и образовательного процесса. Подобный сервис ориентирован на обеспечение эффективного использования современной материально-технической исследовательской базы

в рамках центров обработки данных, центров коллективного пользования, уникальных научных установок, оцифрованных коллекций и банков данных, а также системы научной коммуникации в рамках больших вызовов Стратегии научно-технологического развития Российской Федерации. В рамках единой цифровой платформы предполагается развертывание множества других сервисов (научных, образовательных, аналитических, вычислительных, информационных, административных) на основе наиболее современных гибридных высокопроизводительных вычислительных решений, в том числе геораспределенных, для организации и проведения совместных исследований, в том числе с поддержкой виртуальных коллабораций. Очевидно, что ключевыми драйверами данной парадигмы должны стать облачные технологии и методы интеллектуального анализа больших данных. Для описываемой платформы характерны следующие черты, ориентированные на формирование опережающих ответов на большие вызовы:

– конвергентность характера научных исследований и разработок;

– осуществление эффективного трансфера научных технологий, знаний и компетенций для создания современных (в том числе, и коммерческих) технологий, продуктов и услуг;

– формирование и использование новых организационных и аппаратно-программных решений для эффективной обработки колоссально нарастающего объема информации в процессе проведения научных исследований и разработок;

– привлечение специалистов международного уровня, а также наукоемкие инвестиции в человеческий капитал.

7.5.2 Система управления обучением как ключевой сервис образовательной компоненты цифровой платформы

При этом, как было упомянуто выше, в рамках современной исследовательской парадигмы требуется развитие принципиально новых компетенций, поэтому крайне важной представляется образовательная составляющая данной цифровой платформы. Для нее ключевым сервисом должна стать система управления обучением с целью предоставления сетевых форм образовательной коммуникации и улучшения качества восприятия инновационных технологий, реализации инструментов поддержки развития талантливой молодежи и других аспектов трансфера

знаний для высших учебных заведений, различных частных и государственных компаний.

В настоящий момент системы управления обучением (без привязки к цифровым платформам) являются одним из наиболее современных и удобных средств поддержки общего администрирования учебных курсов, формирования отчетности по образовательным предметам и учебным программам, координации взаимодействия преподавателей и обучающихся, мониторинга показателей их деятельности. Известно [322], что пользователи подобных систем уделяют значительное внимание возможностям эффективной коммуникации, поэтому при разработке данный аспект является чрезвычайно важным, наравне с уровнем программного обеспечения, реализованных сервисов и внедренного контента [347]. Данный круг вопросов влияет на удовлетворенность от использования системы, повышая эффективность обучения [398]. Отметим популярность таких систем в мировом образовательном сообществе [351], а также запрос на возможность интеграции сторонних решений для поддержки обучения, например, для дистанционного выполнения лабораторных работ [364]. В качестве ИТ-основы выбираются облачные технологии [397] в сочетании с мобильными решениями [177] для повышения качества обучения и упрощение идентификации и контроля личности пользователей. Использование облачных решений позволяет значительным образом снизить инфраструктурные расходы, а также гибко реагировать на увеличение потребностей в вычислительных мощностях. Дополнительно в системы управления обучением могут быть внедрены современные методы интеллектуальной обработки данных [368] для персонализации образовательных курсов.

При этом становится возможным внедрение в образовательный процесс технологий искусственного интеллекта. Уже сейчас существуют решения для определения индивидуальных методов эффективного электронного обучения [415], планирования востребованности образовательных курсов [281] и самообучения [300]. В рамках цифровой платформы в системе управления обучением искусственный интеллект может использоваться для непосредственного взаимодействия со студентами (отслеживание удовлетворенности учебным процессом, решение различных вопросов) и с преподавателями (персональные ассистенты). Важной компонентой также является обеспечение максимальной безопасности пользовательских данных за счет использования биометрических решений. Реализация и развитие таких подходов ориентировано на формирование

передовой ИТ-экосистемы современного образования.

Архитектура сервиса управления обучением научно-образовательной цифровой платформы ФИЦ ИУ РАН представлена на рисунке 7.20. Взаимодействие между `Frontend` (тонкие клиенты) и `Backend`-компонентами (хранение пользовательских данных и реализация основного функционала) системы осуществляется посредством единого программного интерфейса `API` (`Application Programming Interface`) для унификации механизма взаимодействия клиентской и сервисной частей.

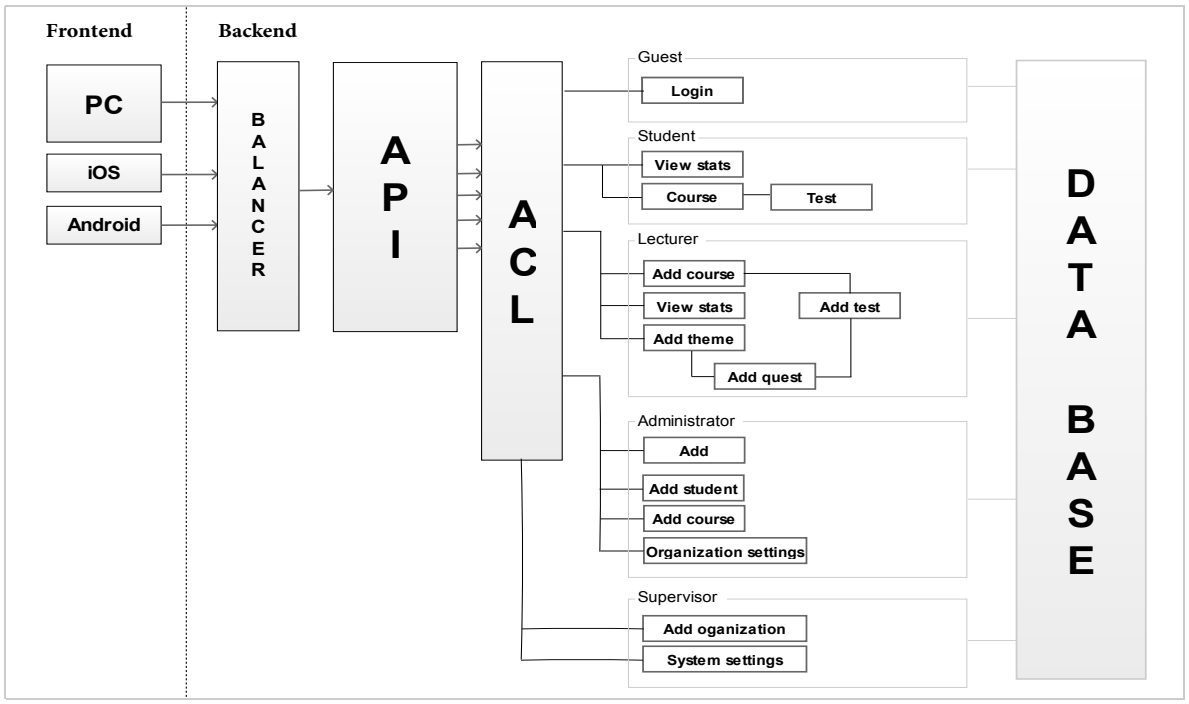

Рис. 7.20. Архитектура системы управления обучением научно-образовательной цифровой платформы ФИЦ ИУ РАН

В начале работы необходимо пройти авторизацию и получить идентификатор сессии, который будет использоваться во всех дальнейших запросах к `API`. Авторизация осуществляется посредством отправки пары значений логин/пароль, которая передается на сервер, где производится хеширование пароля, его проверка и поиск пользователя. В случае, если пользователь был найден, происходит генерация сессионного токена, который передается клиентскому приложению. Данный токен используется во всех последующих запросах для авторизации пользователя. Предусмотрена возможность ассоциации сессионного токена с `IP`-адресом пользователя для предотвращения его использования в случае перехвата третьими лицами. В случае обнаружения пользователя и

корректности введенного пароля производится авторизация и осуществляется загрузка списков контроля доступа `ACL` (`Access Control List`) для конкретного пользователя. Права каждого пользователя основаны на его персональной роли (`Супервизор`, `Администратор`, `Преподаватель`, `Студент`, `Гость`), а также возможностях, доступных для группы, в которой он состоит. Отметим, что проверка наличия разрешения у авторизированного пользователя на выполнение действия или доступа к запрошенным данным проводится при каждом обращении к `API`. В случае, если разрешение имеется, соответствующий запрос выполняется успешно. Все обращения к `API` осуществляются посредством балансировщика нагрузки, который направляет запросы к наименее загруженному узлу, что позволяет эффективно решать задачу масштабирования системы для изменяющегося числа пользователей.

Теперь можно привести схему соответствия сервисов научно-образовательной цифровой платформы ФИЦ ИУ РАН направлениям Стратегии научно-технологического развития Российской Федерации с учетом описанных выше сервисов (см. рисунок 7.21).

Рис. 7.21. Концептуальное соответствие сервисов цифровой платформы ФИЦ ИУ РАН направлениям Стратегии научно-технологического развития Российской Федерации

Представленная диаграмма наглядно демонстрирует тесную взаимосвязь между направлениями Стратегии научно-технологического развития Российской Федерации, поэтому предложенное разделение сервисов цифровой платформы «Наука и образование» по ним носит несколько

условный характер. Некоторые сервисы могут быть использованы для решения задач сразу из нескольких отраслей, при этом значительный эффект достигается именно при условии их комбинации. Подход на основе сервисов цифровой платформы представляется наиболее универсальным и современным.

Разработанные в диссертации методы и алгоритмы, как было отмечено при демонстрации результатов в главах 4–6, были успешно апробированы на высокопроизводительной вычислительной ресурсной базе, которая входит в состав создаваемой цифровой платформы. Они могут быть использованы в образовательном процессе при подготовке студентами дипломных работ, а также предоставляют новые возможности исследовательским коллективам для обработки широкого спектра данных с помощью вероятностно-статистических методов интеллектуального анализа в различных предметных областях.

Заключение

В диссертации предложены, развиты и теоретически обоснованы новые математические методы и вычислительные алгоритмы анализа неоднородных данных с неизвестной сложной стохастической структурой на основе конечных и непрерывных смешанных вероятностных моделей с применением машинного обучения и нейронных сетей. Продемонстрированы высокое соответствие созданных моделей и реальных данных, а также эффективность развитых методов и алгоритмов при анализе временных рядов в различных прикладных областях. А именно, были исследованы:

– пространственно-временные метеорологические (осадки и их интенсивности) и океанологические (турбулентные потоки тепла) данные;

– ансамбли экспериментальных рядов в физике турбулентной плазмы;

– категоризованные характеристики размеров частиц лунного реголита;

– процессы в информационных системах, включая времена выполнения программ, биржевую книгу заявок и интернет-трафик;

– данные медицинских экспериментов, ориентированных на выявление областей активности в головном мозге.

Созданные математические модели и аналитические методы не являются узкоспециализированными и подходят для анализа широкого спектра данных. В диссертации это продемонстрировано при их применении к анализу различных по своей природе временных рядов. Например, могут быть упомянуты:

– метод статистического оценивания распределений случайных параметров стохастических дифференциальных уравнений типа Ланжевена со случайной волатильностью и определения связности компонент для выявления числа структурных процессов в физических данных;

– метод скользящего разделения конечных нормальных и гамма-распределений для физических рядов, турбулентных потоков тепла между океаном и атмосферой, финансовых данных, интернет-трафика и ме-

дицинских исследований;

– бутстреп-процедуры для размеров частиц лунного реголита и физических спектров;

– модифицированный метод превышения порогового значения для осадков, интенсивностей и турбулентных потоков тепла между океаном и атмосферой;

– комбинация паттернов и нейронных сетей для физических и метеорологических выборок.

По результатам исследований, нашедших свое отражение в диссертации, получены следующие новые научные результаты. Доказан вариант центральной предельной теоремы для сумм со случайным числом независимых и необязательно одинаково распределенных слагаемых в схеме серий, в которой в качестве предельного закона возникают произвольные нормальные смеси. Данный результат использован для обоснования вида вероятностных аппроксимаций для размеров частиц лунного реголита.

Доказана теорема об асимптотическом распределении максимальной порядковой статистики в выборке, объем которой является обобщенной отрицательной биномиальной случайной величиной. Получены эквивалентные представления данного распределения в виде смесей известных классических распределений (Фреше, Снедекора-Фишера, строго устойчивого и других) и функциональный вид выражения для моментов произвольных порядков. Установлено, что при некоторых ограничениях на параметры данное распределение является безгранично делимым. В важном частном случае, когда элементы выборки имеют распределение Парето, получена оценка скорости сходимости к предельному закону.

Получен явный вид функциональной зависимости асимптотического распределения максимальной порядковой статистики при условии, что объем выборки является отрицательной случайной биномиальной величиной, а также выписана оценка скорости сходимости для элементов выборки с распределением Парето. Установлен вид асимптотического распределения для произвольных порядковых статистик, а также доказано, что выборочные квантили имеют асимптотическое распределение Стьюдента.

Доказан закон больших чисел для сумм с обобщенным отрицательным биномиальным распределением (обобщение теоремы Реньи), в котором для слагаемых не предполагается независимость и одинаковая распределенность.

Доказаны теоремы устойчивости сдвиговых конечных смесей нор-

мальных законов относительно возмущений параметров смешивающего распределения в терминах расстояния Леви для моделей добавления и расщепления компоненты. Данные результаты обосновывают корректность аппроксимации произвольных сдвиговых нормальных смесей, которые в общем случае не являются идентифицируемыми, конечными аналогами в задаче их статистического разделения.

Доказана теорема об устойчивости дисперсионно-сдвиговых смесей нормальных законов. Показано, что близость смешивающих распределений в смысле расстояния Леви необходимо влечет и близость соответствующих смесей. Данный результат важен, в частности, для обоснования корректности вычислительных процедур оценивания неизвестных параметров дисперсионно-сдвиговых смесей нормальных законов.

Разработан метод статистического оценивания распределений коэффициентов стохастического дифференциального уравнения Ланжевена на основе процедуры скользящего разделения смесей и предложенного в диссертации алгоритма последовательной идентификации (определения локальной связности) компонент смеси на базе специального жадного алгоритма для поиска их числа и одного из методов кластеризации, например, k-средних.

Исследованы теоретические основы для устранения ошибок в модели округления данных, в рамках которой предполагается, что наблюдения для анализа получены с аддитивной ошибкой с известными распределениями и дополнительно округляются до ближайшего целого. В рамках указанной модели получены оценки для неизвестного математического ожидания наблюдений в предположении, что исходные данные зашумлены с помощью случайных величин, имеющих распределения типа конечных смесей нормальных и гамма-законов; построены доверительные интервалы для неизвестного математического ожидания. Такой подход позволяет учесть большее число случайных факторов, влияющих на величину «дополнительной» ошибки, связанной с особенностями практической регистрации наблюдений.

Созданы новые методы интеллектуального анализа данных на основе алгоритма скользящего разделения смесей, а именно: предложен, теоретически обоснован и проверен на наборе тестовых смоделированных выборок алгоритм адаптивного выделения смешанного нормального сигнала на фоне смешанного гауссовского шума; развиты двухэтапный метод детектирования событий в скользящем режиме и процедура искусственного зашумления наблюдений для улучшения результатов СРС-анализа.

Разработана статистическая методология анализа данных с элементами машинного обучения для сгруппированных неизвестных наблюдений при заданных характерных точках их эмпирической функции распределения. Она успешно апробирована для аппроксимации распределений размеров частиц лунного реголита с использованием бутстреп-симуляции выборок и метода минимизации статистики χ^2. На основе центральной предельной теоремы для сумм со случайным числом независимых и необязательно одинаково распределенных слагаемых в специальной двумерной схеме обоснована корректность математической модели изменения размера части и аппроксимации с помощью логнормальных конечных смесей.

Решена задача разработки статистических методов и алгоритмов для обнаружения и идентификации экстремальных наблюдений в различных временных рядах. А именно, предложены методы определения пороговых уровней, развивающие подходы классической теории экстремальных значений на основе обобщения результатов теорем Реньи и Пикандса–Балкемы–Де Хаана. Создан метод классификации наблюдений как абсолютно, промежуточно и относительно экстремальных на основе проверки в скользящем режиме статистических гипотез об однородности выборки из объемов и интенсивностей. С использованием асимптотического распределения экстремальных наблюдений в случае, если их число является случайным с отрицательным биномиальным распределением, разработан подход к определению экстремальных суточных объемов осадков как превышающих квантили выбранных уровней данного распределения. Эти методы могут быть эффективно использованы и для других пространственно-временных метеорологических и иных данных, удовлетворяющих минимальным модельным предположениям относительно распределения элементов выборки и ее объема. Создание подобных инструментов необходимо для прогнозирования потенциально опасных явлений и процессов в глобальных климатических моделях.

Развит подход к анализу данных плазменной турбулентности на основе аппроксимации спектров с помощью конечных сдвиг-масштабных смесей вероятностных распределений. Для нескольких ансамблей спектров, полученных для разных режимов низкочастотной плазменной турбулентности, продемонстрирована эффективность использования предложенного метода, на основании которого удалось решить такие важные для прикладной области задачи, как идентификация амплитудного спектра, определение величины радиального электрического поля и фазовых

скоростей флуктуаций в турбулентной плазме.

Разработан вероятностно-статистический подход к анализу эволюции характеристик микротурбулентности в переходном процессе при электронно-циклотронном резонансном нагреве плазмы на основе конечных нормальных смешанных моделей. С помощью метода выявления локальной связности и CPC-алгоритма определено число структурных компонент (и их изменения во времени) для нескольких ансамблей экспериментальных данных. Продемонстрированы возможность получения содержательных физических результатов на основе анализа моментных характеристик (математическое ожидание, дисперсия, коэффициенты асимметрии и эксцесса) таких моделей для приращений наблюдений исходного процесса и повышение точности прогнозирования значений экспериментальных данных с помощью нейронных сетей за счет расширения признакового пространства этими моментами. Апробированы различные архитектуры для нейросетевого векторного прогнозирования значений указанных моментных характеристик.

Решена задача построения вероятностных и нейросетевых прогнозов на основе k-ичной дискретизации исходных непрерывных данных об объемах осадков. Полученная точность составляет до 97%. При этом для анализа использованы исключительно базовые статистические данные об объемах осадков и не привлекаются какие-либо дополнительные сведения о метеорологических условиях. Продемонстрирована эффективность использования метода случайного поиска для выбора оптимальной конфигурации гиперпараметров для метеорологических данных. Показано, что даже сравнительно небольшое число (порядка десяти) случайно выбранных комбинаций позволяет получить точность, сопоставимую с полным перебором, при этом затраченное время оказывается весьма умеренным. Полученные результаты означают возможность автоматизации выбора эффективной архитектуры для обработки конкретных наборов данных в рамках исследовательского сервиса научной цифровой платформы.

Исследована эффективность различных методов машинного обучения для заполнения пропущенных значений в пространственно-временных метеорологических данных на основе последовательного решения задач классификации и регрессии. Выбран ряд наиболее универсальных методов, которые продемонстрировали свою эффективность при анализе наблюдений метеостанций из регионов, существенно различающихся своим географическим местоположением, при одинаковых на-

стройках гиперпараметров. Это позволяет рассчитывать на сохранение эффективности разработанных подходов и при обработке временных рядов иной природы, например, данных экологического мониторинга окружающей среды.

Предложено и обосновано использование классических и обобщенных отрицательных биномиальных и гамма-моделей для распределений длительностей «дождливых» периодов и соответствующих им объемов осадков. Продемонстрировано высокое соответствие моделей с реальными данными. Разработан эффективный метод функционального оценивания параметров обобщенных распределений. Полученные результаты являются основой для разработки методов статистического определения экстремальных осадков и их интенсивностей.

Проведен анализ статистических закономерностей во временной эволюции тепловых потоков между океаном и атмосферой. Продемонстрировано применение развитого в диссертации метода статистического оценивания коэффициентов стохастического дифференциального уравнения Ланжевена для потоков тепла. Предложен метод определения доли экстремальных наблюдений в них на основе процедуры скользящего разделения конечных нормальных смесей. Продемонстрирована эффективность использования разработанного для осадков и их интенсивностей модифицированного метода превышения порогового значения для выявления аномальных данных и при анализе океанологических рядов. Предложен алгоритм анализа характеристик распределений локальных трендов в потоках тепла с помощью аппроксимации обобщенными отрицательным биномиальным и гамма-распределениями.

Созданы комплексы программных решений на языках программирования `MATLAB` и `Python`, реализующие статистические алгоритмы, методы машинного обучения и работу с нейронными сетями. Они предназначены для автоматизации моделирования, проведения анализа данных и возможности обработки значительных объемов массивов наблюдений. При этом для повышения скорости вычислений в большинстве случаев задействовались гибридные высокопроизводительные вычислительные ресурсы центра коллективного пользования «Информатика» Федерального исследовательского центра «Информатика и управление» Российской академии наук. Данные комплексы могут быть включены в виде сервисов в научно-образовательную цифровую платформу либо непосредственно, либо с учетом некоторых дополнительных архитектурных модификаций, необходимых для полноценного развертывания.

Все полученные результаты являются принципиально новыми, а проведенные исследования – комплексными и имеющими ярко выраженный междисциплинарный характер, поскольку созданные математические модели, вычислительные алгоритмы и программные инструменты анализа данных ориентированы на решение актуальных фундаментальных и прикладных задач в различных предметных областях.

Список литературы

1. *Батанов Г. М., Горшенин А. К., Королев В. Ю., Малахов Д. В., Скворцова Н. Н.* Эволюция вероятностных характеристик низкочастотной турбулентности плазмы в микроволновом поле // Математическое моделирование. – 2011. – Т. 23. Вып. 5. – С. 35–55.

2. *Бенинг В. Е., Горшенин А. К., Королев В. Ю.* Асимптотически оптимальный критерий проверки гипотез о числе компонент смеси вероятностных распределений // Информатика и ее применения. – 2011. – Т. 5. Вып. 3. – С. 4–16.

3. *Вапник В. Н., Червоненкис А. Я.* Теория распознавания образов. – М.: Наука, 1974. –416 с.

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

5. *Галамбош Я.* Асимптотическая теория экстремальных порядковых статистик. – М.: Наука, 1984. –304 с.

6. *Гнеденко Б. В., Колмогоров А. Н.* Предельные распределения для сумм независимых случайных величин. – М.-Л.: ГИТТЛ, 1949. – 264 с.

7. *Горшенин А. К.* О применении асимптотических критериев для определения числа компонент смеси вероятностных распределений // Компьютерные исследования и моделирование. – 2012. – Т. 4. Вып. 1. – С. 45–53.

8. *Горшенин А. К.* О сходимости последовательности SEM-оценок в задаче статистического разделения смесей // Вестник Тверского государственного университета. Серия: Прикладная математика. – 2011. Вып. 4(23). – С. 39–49.

9. *Горшенин А. К.* Проверка статистических гипотез в модели расщепления компоненты // Вестник Московского университета. Серия 15: Вычислительная математика и кибернетика. – 2011. Вып. 4. – С. 26—32.

10. *Горшенин А. К.* Устойчивость масштабных смесей нормальных

законов относительно смешивающего распределения // Системы и средства информатики, 2012. – Т. 22. Вып. 1. – С. 136–148.

11. *Горшенин А. К.* Об устойчивости сдвиговых смесей нормальных законов по отношению к изменениям смешивающего распределения // Информатика и ее применения. – 2012. – Т. 6. Вып. 2. – С. 22–28.

12. *Горшенин А. К.* Программа бутстреп-анализа спектров. Свидетельство о государственной регистрации программ для ЭВМ №2012617918 от 31.08.2012.

13. *Горшенин А. К.* Программа трехмерной визуализации плотностей и параметров распределений. Свидетельство о государственной регистрации программ для ЭВМ №2012660096 от 09.11.2012.

14. *Горшенин А. К.* Информационная технология исследования тонкой структуры хаотических процессов в плазме с помощью анализа спектров // Системы и средства информатики. – 2014. – Т. 24. Вып. 1. – С. 116–127.

15. *Горшенин А. К.* О принципах разработки электронных средств аттестации учащихся по курсам направления «Программирование» // Труды Международной научно-методической конференции «Информатизация инженерного образования» ИНФОРИНО-2014 (Москва, 15-16 апреля 2014 г.). – М.: Издательство МЭИ, 2014. – Р. 529–530.

16. *Горшенин А. К.* Визуализация результатов для метода скользящего разделения смесей // Информатика и ее применения. – 2014. – Т. 8. Вып. 4. – С. 78–84.

17. *Горшенин А. К.* Программный модуль анализа спектров с помощью смесей гамма-распределений. Свидетельство о государственной регистрации программ для ЭВМ №2014612083 от 18.02.2014.

18. *Горшенин А. К.* Информационная технология и программные средства исследования тонкой структуры хаотических процессов в плазме с помощью анализа спектров. Свидетельство о государственной регистрации программ для ЭВМ №2014612085 от 18.02.2014.

19. *Горшенин А. К.* Программный модуль вероятностного анализа спектров на основе логарифмических преобразований. Свидетельство о государственной регистрации программ для ЭВМ №2014661370 от 29.10.2014.

20. *Горшенин А. К.* Средство визуализации результатов для метода скользящего разделения смесей. Свидетельство о государственной регистрации программ для ЭВМ №2014661369 от 29.10.2014.

21. *Горшенин А. К.* Программный модуль «Ядро СРС-метода». Свидетельство о государственной регистрации программ для ЭВМ №2015618673 от 13.08.2015.

22. *Горшенин А. К.* Модуль визуализации моментных характеристик и квантилей для конечных смесей вероятностных распределений. Свидетельство о государственной регистрации программ для ЭВМ №2015618564 от 12.08.2015.

23. *Горшенин А. К.* Концепция онлайн-комплекса для стохастического моделирования реальных процессов // Информатика и ее применения. – 2016. – Т. 10. Вып. 1. – С. 72–81.

24. *Горшенин А. К.* Некоторые аспекты разработки мобильных приложений для аттестации учащихся // Труды Международной научно-методической конференции «Информатизация инженерного образования» – ИНФОРИНО-2016 (Москва, 12-13 апреля 2016 г.). – М.: Издательский дом МЭИ, 2016. – С. 92–95.

25. *Горшенин А. К.* Управляющий модуль для СРС-метода. Свидетельство о государственной регистрации программ для ЭВМ №2016613924 от 11.04.2016.

26. *Горшенин А. К.* Программный модуль динамической визуализации эволюции параметров СРС-метода. Свидетельство о государственной регистрации программ для ЭВМ №2016613925 от 11.04.2016.

27. *Горшенин А. К.* Оптимизированный модуль графического вывода для СРС-метода. Свидетельство о государственной регистрации программ для ЭВМ №2016618859 от 09.08.2016.

28. *Горшенин А. К.* Программный модуль анализа статистических характеристик осадков. Свидетельство о государственной регистрации программ для ЭВМ №2016618864 от 09.08.2016.

29. *Горшенин А. К.* О некоторых математических и программных методах построения структурных моделей информационных потоков // Информатика и ее применения. – 2017. – Т. 11. Вып. 1. – С. 58–68.

30. *Горшенин А. К.* Анализ вероятностно-статистических характеристик осадков на основе паттернов // Информатика и ее применения. – 2017. – Т. 11. Вып. 4. – С. 38–46.

31. *Горшенин А. К.* Программный модуль статистического анализа физических экспериментальных данных. Свидетельство о государственной регистрации программ для ЭВМ №2017617451 от 04.07.2017.

32. *Горшенин А. К.* Программный модуль поиска порогового значе-

ния для объемов и интенсивностей осадков. Свидетельство о государственной регистрации программ для ЭВМ № 2017662539 от 10.11.2017.

33. *Горшенин А. К.* Программный модуль анализа вероятностно-статистических характеристик объемов осадков на различных временных интервалах. Свидетельство о государственной регистрации программ для ЭВМ № 2017662540 от 10.11.2017.

34. *Горшенин А. К.* Зашумление данных конечными смесями нормальных и гамма-распределений с применением к задаче округления наблюдений // Информатика и ее применения. – 2018. – Т. 12. Вып. 3. – С. 28–34.

35. *Горшенин А. К.* Развитие сервисов цифровых платформ для преодоления нефинансовых барьеров // Информатика и ее применения. – 2018. – Т. 12. Вып. 4. – С. 109–115.

36. *Горшенин А. К.* Программа оценивания параметров обобщенного отрицательного биномиального распределения на основе функционального подхода. Свидетельство о государственной регистрации программ для ЭВМ № 2018619090 от 30.07.2018.

37. *Горшенин А. К.* Программа оценивания параметров обобщенного гамма-распределения на основе функционального подхода. Свидетельство о государственной регистрации программ для ЭВМ № 2018619794 от 10.08.2018.

38. *Горшенин А. К.* Программа скользящего разделения конечных смесей гамма-распределений с оптимизацией на основе векторных вычислений. Свидетельство о государственной регистрации программ для ЭВМ № 2018619795 от 10.08.2018.

39. *Горшенин А. К.* Программа классификации экстремальных объемов осадков. Свидетельство о государственной регистрации программ для ЭВМ № 2018619796 от 10.08.2018.

40. *Горшенин А. К.* Программный модуль статистического определения экстремальных пороговых уровней для максимумов дневных объемов осадков. Свидетельство о государственной регистрации программ для ЭВМ № 2018619922 от 14.08.2018.

41. *Горшенин А. К.* Программный модуль визуализации точности обучения нейронных сетей. Свидетельство о государственной регистрации программ для ЭВМ № 2018619923 от 14.08.2018.

42. *Горшенин А. К.* Программа статистического анализа распределений объемов осадков за дождливые периоды с графическим пользова-

тельским интерфейсом. Свидетельство о государственной регистрации программ для ЭВМ № 2018661221 от 04.09.2018.

43. *Горшенин А. К.* Программа статистического анализа распределений длительностей дождливых периодов с графическим пользовательским интерфейсом. Свидетельство о государственной регистрации программ для ЭВМ № 2018661222 от 04.09.2018.

44. *Горшенин А. К.* Программа двухэтапного определения аномальных интенсивностей осадков. Свидетельство о государственной регистрации программ для ЭВМ № 2018665545 от 06.12.2018.

45. *Горшенин А. К.* О выявлении смешанного нормального сигнала на фоне смешанного гауссовского шума // Обозрение прикладной и промышленной математики. – 2019. – Т. 26. Вып. 2. – С. 152–153.

46. *Горшенин А. К.* Программа анализа статистических свойств микротурбулентности в переходном процессе при электронно-циклотронном резонансном нагреве плазмы. Свидетельство о государственной регистрации программ для ЭВМ № 2019615238 от 22.04.2019.

47. *Горшенин А. К.* Программа анализа вероятностных характеристик данных метеорологических станций в пакетном режиме. Свидетельство о государственной регистрации программ для ЭВМ № 2019664376 от 06.11.2019.

48. *Горшенин А. К.* Программа кластеризации параметров вероятностной аппроксимации распределений размеров частиц лунного реголита. Свидетельство о государственной регистрации программ для ЭВМ № 2019664471 от 07.11.2019.

49. *Горшенин А. К.* Программа аппроксимации вероятностных распределений размеров частиц лунного реголита. Свидетельство о государственной регистрации программ для ЭВМ № 2019664472 от 07.11.2019.

50. *Горшенин А. К.* Программа аппроксимации вероятностных распределений характеристик локальных трендов в турбулентных потоках тепла между океаном и атмосферой. Свидетельство о государственной регистрации программ для ЭВМ № 2019664808 от 13.11.2019.

51. *Горшенин А. К., Данилович Е. С., Хромов Д. Р.* Система управления обучением ELIS. Архитектурные решения // Системы и средства информатики. – 2017. – Т. 27. Вып. 2. – С. 60–69.

52. *Горшенин А. К., Данилович Е. С., Хромов Д. Р.* Система управ-

ления обучением ELIS. Пользовательский интерфейс и функциональные возможности // Системы и средства информатики, 2017. – Т. 27. Вып. 2. – С. 70–84.

53. *Горшенин А. К., Зацаринный А. А.* Цифровизация науки: платформенный подход // Актуальные проблемы глобальных исследований: Россия в глобализирующемся мире. Сборник материалов VI Всероссийской научно-практической конференции, МГУ имени М. В. Ломоносова, 4–6 июня 2019 г. / под ред. И.В. Ильина. – М.: МООСИПНН Н. Д. Кондратьева, 2019. – 466 с. – С. 91–95.

54. *Горшенин А. К., Зейфман А. И., Королев В. Ю., Агафонов Е. С., Белоусов В. В., Дышкант Н. Ф.* О применении метода скользящего разделения смесей для стохастической верификации времени выполнения программ // Обозрение прикладной и промышленной математики. – 2015. – Т. 22. Вып. 5. – С. 350–351.

55. *Горшенин А. К., Королев В. Ю.* Применение смесей логнормальных распределений для аппроксимации неизвестных плотностей // Обозрение прикладной и промышленной математики. – 2014. – Т. 21. Вып. 4. – С. 350–351.

56. *Горшенин А. К., Королев В. Ю.* Программный модуль поиска моментов начала движения по миограмме с помощью анализа динамической компоненты. Свидетельство о государственной регистрации программ для ЭВМ №2015618672 от 13.08.2015.

57. *Горшенин А. К., Королев В. Ю.* Статистический подход для определения экстремальных пороговых значений // Информационно-коммуникационные технологии и математическое моделирование высокотехнологичных систем: материалы Всероссийской конференции с международным участием. – М.: РУДН, 2016. – С. 90–92.

58. *Горшенин А. К., Королев В. Ю.* Программный модуль предсказания осадков на основе исторических паттернов. Свидетельство о государственной регистрации программ для ЭВМ №2016618887 от 09.08.2016.

59. *Горшенин А. К., Королев В.Ю.* Определение экстремальности объемов осадков на основе модифицированного метода превышения порогового значения // Информатика и ее применения. – 2018. – Т. 12. Вып. 4. – С. 16–24.

60. *Горшенин А. К., Королев В.Ю.* Обобщенные вероятностные модели экстремальных осадков // Ломоносовские чтения: научная конференция. Тезисы докладов. – М: Издательский отдел факультета ВМК

МГУ, 2020. – С. 62–63.

61. *Горшенин А. К., Королев В.Ю.* Аппроксимация распределений размеров частиц лунного реголита на основе метода статистической симуляции выборок // Информатика и ее применения. – 2020. – Т. 14. Вып. 2. – С. 50–57.

62. *Горшенин А. К., Королев В. Ю., Казаков И. А.* Робастная версия EM-алгоритма для конечных смесей нормальных законов // Вестник Тверского государственного университета. Серия: Прикладная математика. – 2011. Вып. 3(22). – С. 63–71.

63. *Горшенин А. К., Королев В. Ю., Малахов Д. В., Скворцова Н. Н.* Анализ тонкой стохастической структуры хаотических процессов с помощью ядерных оценок // Математическое моделирование. – 2011. – Т. 23. Вып. 4. – С. 83–89.

64. *Горшенин А. К., Королев В. Ю., Малахов Д. В., Скворцова Н. Н.* Об исследовании плазменной турбулентности на основе анализа спектров // Компьютерные исследования и моделирование. – 2012. – Т. 4. Вып. 4. – С. 793–802.

65. *Горшенин А. К., Королев В. Ю., Турсунбаев А. М.* Медианные модификации EM- и SEM-алгоритмов для разделения смесей вероятностных распределений и их применение к декомпозиции волатильности финансовых временных рядов // Информатика и ее применения. – 2008. Т. 2. Вып. 4. – С. 12–47.

66. *Горшенин А. К., Королев В.Ю., Щербинина А. А.* Статистическое оценивание распределений случайных коэффициентов стохастического дифференциального уравнения Ланжевена // Информатика и ее применения. – 2020. – Т. 14. Вып. 3. – С. 3–12.

67. *Горшенин А. К., Кузьмин В. Ю.* Применение архитектуры CUDA при реализации сеточных алгоритмов для метода скользящего разделения смесей // Системы и средства информатики. – 2016. – Т. 26. Вып. 4. – С. 60–73.

68. *Горшенин А. К., Кузьмин В. Ю.* Портал MSM Tools как гетерогенный вычислительный сервис // Системы и средства информатики. – 2017. – Т. 27. Вып. 1. – С. 61–73.

69. *Горшенин А. К., Кузьмин В. Ю.* Программный модуль асинхронной конвейерной обработки данных на основе медианной модификации EM-алгоритма для системы поддержки научных исследований. Свидетельство о государственной регистрации программ для ЭВМ

№ 2017663370 от 30.11.2017.

70. *Горшенин А. К., Кузьмин В. Ю.* Программный модуль асинхронной конвейерной обработки данных на основе сеточных методов для системы поддержки научных исследований. Свидетельство о государственной регистрации программ для ЭВМ № 2017663371 от 30.11.2017.

71. *Горшенин А. К., Кузьмин В. Ю.* Прогнозирование моментов конечных нормальных смесей с использованием нейронных сетей прямого распространения // Системы и средства информатики. – 2018. – Т. 28. Вып. 3. – С. 61–70.

72. *Горшенин А. К., Кузьмин В. Ю.* Применение рекуррентных нейронных сетей для прогнозирования моментов конечных нормальных смесей // Информатика и ее применения. – 2019. – Т. 13. Вып. 3. – С. 114–121.

73. *Горшенин А. К., Кузьмин В. Ю.* Оптимизация гиперпараметров нейронных сетей с использованием высокопроизводительных вычислений для предсказания осадков // Информатика и ее применения. – 2019. – Т. 13. Вып. 1. – С. 75–81.

74. *Горшенин А. К., Кузьмин В. Ю.* Программа векторного прогнозирования временных рядов с использованием нейронных сетей. Свидетельство о государственной регистрации программ для ЭВМ № 2019665119 от 20.11.2019.

75. *Горшенин А. К., Кузьмин В. Ю.* Анализ конфигураций LSTM-сетей для построения среднесрочных векторных прогнозов // Информатика и ее применения. – 2020. – Т. 14. Вып. 1. – С. 10–16.

76. *Горшенин А. К., Лебедева М. А., Лукина С. С.* Программа заполнения пропусков в данных с использованием методов машинного обучения. Свидетельство о государственной регистрации программ для ЭВМ № 2019664807 от 13.11.2019.

77. *Горшенин А. К., Мартынов О. П.* Гибридные модели экстремального градиентного бустинга для восстановления пропущенных значений в данных об осадках // Информатика и ее применения. – 2019. – Т. 13. Вып. 3. – С. 34–40.

78. *Заиченко С. А.* Центры превосходства в системе современной научной политики // Форсайт. – 2008. – Т. 2. № 1. – С. 42–50.

79. *Захарова Т. В., Никифоров С. Ю., Гончаренко М. Б. и др.*. Методы обработки сигналов для локализации невосполнимых областей го-

ловного мозга // Системы и средства информатики. – 2012. – Т. 22. Вып. 2. – С. 157–175.

80. *Зацаринный А. А., Горшенин А. К., Волович К. И., Колин К. К., Кондрашев В. А., Степанов П. В.* Управление научными сервисами как основа национальной цифровой платформы «Наука и образование» // Стратегические приоритеты. – 2017. – Вып. 2 (14). – С. 103–113.

81. *Зацаринный А. А., Горшенин А. К., Волович К. И., Кондрашев В. А.* Основные направления развития информационных технологий в условиях вызовов цифровой экономики // Цифровая обработка сигналов. – 2018. Вып. 1. – С. 3–7.

82. *Ибрагимов И. А., Линник Ю. В.* Независимые и стационарно связанные величины. – М.: Наука, 1965. – 524 с.

83. *Колмогоров А. Н.* Избранные труды. Том 2: Теория вероятностей и математическая статистика. – М.: Наука, 2005 –581 с.

84. *Колмогоров А. Н.* О логарифмически нормальном законе распределения размеров частиц при дроблении // Доклады академии наук СССР. – 1941. – Т. 31. Вып. 2. – С. 99–101.

85. *Королев В. Ю.* Сходимость случайных последовательностей с независимыми случайными индексами. I // Теория вероятностей и ее применения. – 1994. – Т. 39. Вып. 2. – С. 313–333.

86. *Королев В. Ю.* Сходимость случайных последовательностей с независимыми случайными индексами. II // Теория вероятностей и ее применения. – 1995. – Т. 40. Вып. 4. – С. 907–910.

87. *Королев В. Ю.* О распределении размеров частиц при дроблении // Информатика и ее применения. – 2009. – Т. 3. Вып. 3. С. 60–68.

88. *Королев В. Ю.* Вероятностно-статистические методы декомпозиции волатильности хаотических процессов. – М.: Изд-во Моск. ун-та, 2011. – 512 с.

89. *Королев В. Ю.* Обобщенные гиперболические распределения как предельные для случайных сумм // Теория вероятностей и ее применения. – 2013. – Т. 58. Вып. 1. – С. 117–132.

90. *Королев В. Ю.* Предельные распределения для дважды стохастически прореженных процессов восстановления и их свойства // Теория вероятностей и ее применения. 2016. – Т. 61. Вып. 4. – С. 753–773.

91. *Королев В. Ю.* Аналоги теоремы Глезера для отрицательных биномиальных и обобщенных гамма-распределений и некоторые их приложения // Информатика и ее применения. – 2017. – Т. 11. Вып. 3. –

С. 2–17.

92. *Королев В. Ю., Арефьева Е. В., Нефедова Ю. С., Горшенин А. К., Лазовский Р. А.* Метод оценивания вероятностей катастроф в неоднородных потоках экстремальных событий и его применение к прогнозированию землетрясений в Арктике // Проблемы анализа риска. – 2016. – Т. 13. № 4. – С. 80–91.

93. *Королев В. Ю., Горшенин А. К.* О распределении вероятностей экстремальных осадков // Доклады Академии Наук. – 2017. – Т. 477. Вып. 5. – С. 604–609.

94. *Королев В. Ю., Горшенин А. К., Гулев С. К., Беляев К. П.* Вероятностно-статистическое моделирование турбулентных потоков тепла между океаном и атмосферой с помощью метода скользящего разделения смесей нормальных законов // Тихоновские чтения: Научная конференция, Москва, МГУ им. М. В. Ломоносова, 26 октября – 2 ноября 2015 г. Тезисы докладов. – М.: МАКС Пресс, 2015. – С. 72.

95. *Королев В. Ю., Горшенин А. К., Гулев С. К., Беляев К. П.* Статистическое моделирование турбулентных потоков тепла между океаном и атмосферой с помощью метода скользящего разделения конечных нормальных смесей // Информатика и ее применения. – 2015. – Т. 9. Вып. 4. – С. 3–13.

96. *Королев В. Ю., Корчагин А. Ю., Горшенин А. К.* Некоторые свойства дисперсионно-сдвиговых смесей нормальных законов // Статистические методы оценивания и проверки гипотез. – 2015. Вып. 26. – С. 134–153.

97. *Королев В. Ю., Соколов И. А.* Математические модели неоднородных потоков экстремальных событий. – М.: Торус-Пресс, 2008. – 200 с.

98. *Круглов В. М.* Дополнительные главы теории вероятностей. – М.: Высшая школа, 1984. – 264 с.

99. *Круглов В. М., Королев В. Ю.* Предельные теоремы для случайных сумм. – М.: Изд-во Моск. ун-та, 1990. – 269 с.

100. *Макс Ж.* Методы и техника обработки сигналов при физических измерениях. Т. 1,2. – М.: Мир. 1983. – 566 с.

101. *Малахов Д. В., Скворцова Н. Н., Васильков Д. Г., Смирнов В. А., Тедтоев Б. А., Горшенин А. К., Черноусов А. Д.* Программно-аппаратные методы сбора данных в плазменных экспериментах (на примере создания нового комплекса для стелларатора Л-2М) // Труды

IX Международной конференции «Современные средства диагностики плазмы и их применение», Москва, 5–7 ноября 2014 г. – М.: Изд-во НИЯУ МИФИ, 2014. – С. 60–61.

102. *Малахов Д. В., Скворцова Н. Н., Васильков Д. Г., Чирков А. Ю., Смирнов В. А., Тедтоев Б. А., Горшенин А. К., Черноусов А. Д.* Программно-аппаратный комплекс многопараметрической обработки данных на установке стелларатор Л-2М // XLII Международная Звенигородская конференция по физике плазмы и управляемому термоядерному синтезу, 9-13 февраля 2015 г., Звенигород. Сборник тезисов докладов – М.: ЗАО НТЦ «ПЛАЗМАИОФАН», 2015. – С. 79.

103. *Марпл-мл. С. П.* Цифровой спектральный анализ и его приложения. – М.: Мир, 1990. – 265 с.

104. *Перри А. Х., Уокер Дж. М.* Система океан-атмосфера. – Л.: Гидрометеоиздат, 1979. – 194 с.

105. *Петров В. В.* Суммы независимых случайных величин. – М.: Наука, 1972. – 416 с.

106. *Попель С. И., Голубь А. П., Захаров А. В. и др.* Формирование плазменно-пылевых облаков при ударе метеороида о поверхность Луны // Письма в Журнал экспериментальной и теоретической физики. – 2018. – Т. 108. Вып. 6. – С. 379–387.

107. *Прохоров Ю. В.* Избранные труды. – М.: Торус Пресс, 2012. – 775 с.

108. *Разумовский Н. К.* Характер распределения содержания металлов в рудных месторождениях // Доклады академии наук СССР. – 1940. – Т. 28. Вып. 9. – С. 815–817.

109. *Скворцова Н. Н., Горшенин А. К., Королев В. Ю., Малахов Д. В., Чернов Н. А.* Об исследовании низкочастотной структурной плазменной турбулентности на основе анализа Фурье-спектров // XL Международная Звенигородская конференция по физике плазмы и управляемому термоядерному синтезу, г. Звенигород, 11-15 февраля 2013 г. Тезисы докладов. М.: ЗАО НТЦ «ПЛАЗМАИОФАН», 2013. – С. 35.

110. *Скворцова Н. Н.,Малахов Д. В., Степахин В. Д., Майоров С. А.,Батанов Г. М. и др.* Инициация пылевых структур в цепных реакциях под воздействием излучения гиротрона на смесь порошков металла и диэлектрика с открытой границей // Письма в Журнал экспериментальной и теоретической физики. – 2017. – Т. 106. Вып. 3–4. –

С. 240–246.

111. *Скворцова Н. Н., Майоров С. А., Малахов Д. В., Степахин В. Д., Образцова Е. А., Кенжебекова А. И., Шишилов О. Н.* О пылевых структурах и цепных реакциях, возникающих над реголитом при воздействии излучения гиротрона // Письма в Журнал экспериментальной и теоретической физики. – 2019. – Т. 109. Вып. 7–8. – С. 452–459.

112. *Слюта Е. Н.* Физико-механические свойства лунного грунта (обзор) // Астрономический вестник. – 2014. – Т. 48. Вып. 5. – С. 358–382.

113. *Смирнов Н. В., Дунин-Барковский И. В.* Курс теории вероятностей и математической статистики для технических приложений. – М.: Наука, 1969. – 512 с.

114. *Ушаков В. Г., Ушаков Н. Г.* Об усреднении округленных данных // Информатика и ее применения. – 2015. Т. 9. Вып. 4. – С. 106–109.

115. *Ушаков В. Г., Ушаков Н. Г.* Границы точности восстановления информации, теряемой при округлении результатов наблюдений // Вестник Московского университета. Серия 15: Вычислительная математика и кибернетика. – 2017. Вып. 2. – С. 26–30.

116. *Флоренский К. П., Базилевский А. Т., Николаева О. В.* Лунный грунт: свойства и аналоги. – М.: Наука, 1975. – 50 с.

117. *П. Хьюбер.* Робастность в статистике. М.: Мир, 1984. – 304 с.

118. *Шваб К.* Четвертая промышленная революция. – М.: Эксмо, 2016. – 208 с.

119. *Ширяев А. Н.* Вероятность-1. – М.: МЦНМО, 2017. – 552 с.

120. *Ширяев А. Н.* Вероятность-2. – М.: МЦНМО, 2017. – 416 с.

121. *Ширяев А. Н.* Основы стохастической финансовой математики. Т. 1. Факты. Модели. – М.: МЦНМО, 2016. – 440 с.

122. *Ширяев А. Н.* Стохастические задачи о разладке. – М.: МЦНМО, 2016. – 392 с.

123. *Abanto-Vallea C. A., Bandyopadhyayb D., Lachosc V. H., Enriquezd I.* Robust Bayesian analysis of heavy-tailed stochastic volatility models using scale mixtures of normal distributions // Computational Statistics & Data Analysis. – 2010. – Vol. 54. Iss. 12. – P. 2883–2898.

124. *Abaurrea J.* Forecasting local daily precipitation patterns in a climate change scenario // Climate Research. – 2005. – Vol. 28. Iss. 3. – P. 183–197.

125. *Akaike H.* Information theory and an extension of the maximum likelihood principle // In: B. N. Petrov and F. Csake (eds.) Second International Symposium on Information Theory. – Budapest, 1973. – P. 267–281.

126. *Akamatsu Y., Yamamoto N.* Chiral Langevin theory for non-Abelian plasmas // Physical Review D. – 2014. – Vol. 90. Iss. 12. – Art. No. 125031.

127. *Albers W.* Asymptotic Expansions and the Deficiency Concept in Statistics. – Amsterdam: Mathematisch Centrum, 1974. – 144 p.

128. *Albers W.* Efficiency and deficiency considerations in the symmetry problem // Statistica Neerlandica. – 1975. – Vol. 29. – P. 81–92.

129. *Aler R., Galvan I. M., Ruiz-Arias J. A., Gueymard C. A.* Improving the separation of direct and diffuse solar radiation components using machine learning by gradient boosting // Solar Energy. – 2017. – Vol. 150. – P. 558–569.

130. *Alexander L. V., Zhang X., Peterson T.- C. et al.* Global observed changes in daily climate extremes of temperature and precipitation // Journal of Geophysical Research-Atmospheres. – 2006. – Vol. 111(D5).. – Art. No. D05109.

131. *Almgren H., Van de Steen F., Razi A., Friston K., Marinazzo D.* The effect of global signal regression on DCM estimates of noise and effective connectivity from resting state fMRI // Neuroimage. – 2020. – Vol. 208. – Art. No. 116435.

132. *Altman N.* An introduction to kernel and nearest-neighbor nonparametric regression // The American Statistician. – 1992. – Vol. 46. Iss. 3. – P. 175–185.

133. *D'Ambrosio D., Filippone G., Marocco D., Rongo R., Spataro W.* Efficient application of GPGPU for lava flow hazard mapping // Journal of supercomputing. – 2013. – Vol. 65. Iss. 2. – P. 630–644.

134. *Asadi H., Seyfe B.* Signal enumeration in Gaussian and non-Gaussian noise using entropy estimation of eigenvalues // Digital Signal Processing. – 2018. – Vol. 78. – P. 163–174.

135. *Athey S., Tibshirani J., Wager S.* Generalized Random Forests // Annals of Statistics. – 2019. – Vol. 47. Iss. 2. – P. 1148–1178.

136. *Audhkhasi K., Osoba O., Kosko B.* Noise-enhanced convolutional neural networks // Neural Networks. – 2016. – Vol. 78. – P. 15–23.

137. *Aymar R., Barabaschi P., Shimomura Y.* The ITER design //

Plasma Physics and Controlled Fusion. – 2002. – Vol. 44. Iss. 5. – P. 519–565.

138. *AyresF. J.* Schaum's Outline of Theory and Problems of Matrices. – New York: McGraw Hill Book Company, 1962. – 219 p.

139. *Bagnold R. A.* The Physics of Blown Sand and Desert Dunes. – London: Methuen, 1954. – 265 p.

140. *Bai Z., Zheng S., Zhang B., Hu G.* Statistical analysis for rounded data // Journal of Statistical Planning and Inference. – 2009. – Vol. 139. Iss. 8. – P. 2526–2542.

141. *Balkema A., de Haan L.* Residual life time at great age // Annals of Probability. – 1974. – Vol. 2. – P. 792–804.

142. *Ban N., Schmidli J., Schar C.* Heavy precipitation in a changing climate: Does short-term summer precipitation increase faster // Geophysical Research Letters. – 2015. – Vol. 42. Iss. 4. – P. 1165–1172.

143. *Barndorff-Nielsen O. E.* Exponentially decreasing distributions for the logarithm of particle size // Proc. R. Soc. London. – 1977. Vol. A 353. P. 401–419.

144. *Barrios A., Trincado G., Garreaud R.* Alternative approaches for estimating missing climate data: application to monthly precipitation records in South-Central Chile // Forest Ecosystems. – 2018. – Vol. 5 – Art. No. 28.

145. *Batanov G. M., Bening V. E., Korolev V. Yu. et al.* G. M. Low-Frequency Structural Plasma Turbulence in the L-2M Stellarator // JETP Letters. – 2003. Vol. 78 (8). – P. 502–510.

146. *Batanov G. M., Berezhetskii M. S., Borzosekov V. D. et al.* Reaction of turbulence at the edge and in the center of the plasma column to pulsed impurity injection caused by the sputtering of the wall coating in L-2M stellarator // Plasma Physics Reports. – 2017. – Vol. 43. Iss. 8. – P. 818–823.

147. *Batanov G. M., Borzosekov V. D., Gorshenin A. K., Kharchev N. K., Korolev V. Yu., Sarskyan K. A.* Evolution of statistical properties of microturbulence during transient process under electron cyclotron resonance heating of the L-2M stellarator plasma // Plasma Physics and Controlled Fusion. – 2019. – Vol. 61. Iss. 7. Art. No. 075006 (7 p.)

148. *Begueria S., Angulo-Martinez M., Vicente-Serrano S. M., Lopez-Moreno I. J., El-Kenawy A.* Assessing trends in extreme precipitation events intensity and magnitude using non-stationary peaks-over-threshold analysis:

a case study in northeast Spain from 1930 to 2006 // International Journal of Climatology. – 2011. – Vol. 31. Iss. 142. – P. 2102–2114.

149. *Belyaev K., Kuleshov A., Tuchkova N., Tanajura C. A. S.* An optimal data assimilation method and its application to the numerical simulation of the ocean dynamics // Mathematical and Computer Modelling of Dynamical Systems. – 2018. – Vol. 1. Iss. 24. – P. 12–25.

150. *Bening V. E.* Asymptotic Theory Of Testing Statistical Hypothesis: Efficient Statistics, Optimality, Power Loss and Deficiency. – Untrecht: VSP, 2000. – 277 p.

151. *Bening V.E., Korolev V.Yu.* Generalized Poisson Models and Their Applications in Insurance and Finance. – Berlin: De Gruyter , 2012. – 434 p.

152. *Bergstra J., Bengio Y.* Random Search for Hyper-Parameter Optimization // Journal of Machine Learning Research. – 2012. – Vol. 13. – P. 281–305.

153. *Berry D. I., Kent E. C.* A new air-sea interaction gridded dataset from ICOADS with uncertainty estimates // Bulletin of the American Meteorological Society. – 2009. – Vol. 90. Iss. 5. – P. 645–656.

154. *Berry D. I., Kent E. C.* A new air-sea interaction gridded dataset from ICOADS with uncertainty estimates // Bulletin of the American Meteorological Society. – 2009. – Vol. 90. Iss. 5. – P. 645–656.

155. *Bickel P. J., Ritov Y.* Non- and semiparametric statistics: compared and contrasted // Journal of Statistical Planning and Inference. – 2000. – Vol. 91. Iss. 2. – P. 209–228.

156. *Bloemer J., Brauer S., Bujna K., Kuntze D.* How well do SEM algorithms imitate EM algorithms? A non-asymptotic analysis for mixture models // Advances in Data Analysis and Classification. – 2020. – Vol. 14. Iss. 1. – P. 147–173 (2020).

157. *Bouchaud J. P., Cont R.* A Langevin approach to stock market fluctuations and crashes // European Physical Journal B. – 1998. – Vol. 6. Iss. 4. – P. 543–550.

158. *Bouras D.* Comparison of five satellite-derived latent heat flux products to moored buoy data // Journal of Climate. – 2006. – Vol. 19. – P. 6291–6313.

159. *Bouvet M., Schwartz Sc.* Underwater Noises–Statistical Modeling, Detection, and Normalization // Journal of the Acoustical Society of America. – 1988. – Vol. 83. Iss. 3. – P. 1023–1033.

160. Breiman L. Random forests // Machine Learning. – 2001. –

Vol. 45. – P. 5–32.

161. *Braga-Neto U. M., Dougherty E. R.* Machine Learning Requires Probability and Statistics // IEEE Signal Processing Magazine. – 2020. – Vol. 37. Iss. 4. – P. 118–122.

162. *Breitwieser C., Kaiser V., Neuper C., Muller-Putz G. R.* Stability and distribution of steady-state somatosensory evoked potentials elicited by vibro-tactile stimulation // Medical & Biological Engineering & Computing. – 2012. – Vol. 50. Iss. 4. – P. 347–357.

163. *Brodtkorb A. R., Dyken C., Hagen T. R., Hjelmervik J. M., Storaasli O. O.* State-of-the-art in heterogeneous computing // Scientific Programming. – 2010. – Vol. 185. Iss. 1. – P. 1–33.

164. *Broniatowski M., Celeux G., Diebolt J.* Reconnaissance de mélanges de densités par un algorithme d'apprentissage probabiliste // Data Analysis and Informatics. – 1983. – Vol. 3. – P. 359–373.

165. *Buduma N.* Fundamentals of Deep Learning: Designing Next-Generation Machine Intelligence Algorithms. – Sebastopol, CA: O'Reilly Media, 2017. – 298 p.

166. *Burger C. M., Kollet S., Schumacher J., Bosel D.* Introduction of a web service for cloud computing with the integrated hydrologic simulation platform ParFlow // Computers & Geosciences. – 2012. – Vol. 48. – P. 334–336.

167. *Burns R., Vogelstein J. T., Szalay A. S.* From Cosmos to Connectomes: The Evolution of Data-Intensive Science // Neuron. – 2014. – Vol. 83. Iss. 6. – P. 1249–1252.

168. *Byrd R. H., Hribar M. E., Nocedal J.* An Interior Point Algorithm for Large-Scale Nonlinear Programming // SIAM Journal on Optimization. – 1999. – Vol. 9. Iss. 4. – P. 877–900.

169. *Byrd R. H., Gilbert J. C., Nocedal J.* A Trust Region Method Based on Interior Point Techniques for Nonlinear Programming // Mathematical Programming. – 2000. – Vol. 89. Iss. 1. – P. 149–185.

170. *Bzdok D., Altman N., Krzywinski M.* Statistics versus machine learning // Nature Methods. – 2018. – Vol. 15. Iss. 4. – P. 232–233.

171. *Cai T. T., Ma J., Zhang L.* CHIME: Clustering of High-Dimensional Gaussian Mixtures with EM Algorithm and its Optimality // Annals of Statistics. – 2019. – Vol. 47. Iss. 3. – P. 1234–1267.

172. *Celeux G., Diebolt J.* Asymptotic properties of a stochastic EM algorithm for estimating mixing proportions // Communications in

statistics. Stochastic models. – 1993. – Vol. 9. – P. 599–613.

173. Celik T., Tjahjadi T. Automatic Image Equalization and Contrast Enhancement Using Gaussian Mixture Modeling // IEEE Transactions on Image Processing. – 2012. – Vol. 21. Iss. 1. – P. 145–156.

174. *Chandrashekar G., Sahin F.* A survey on feature selection methods // Computers & Electrical Engineering. – 2014. – Vol. 40. Iss. 1. – P. 16–28.

175. *Chatzis S. P., Siakoulis V., Petropoulos A., Stavroulakis E., Vlachogiannakis N.* Forecasting stock market crisis events using deep and statistical machine learning techniques // Expert Systems with Applications. – 2018. – Vol. 112. – P. 353–371.

176. *Che S., Boyer M., Meng J., Tarjan D., Sheaffer J. W., Skadron K.* A performance study of general-purpose applications on graphics processors using CUDA // Journal of Parallel and Distributed Computing. – 2008. – Vol. 68. Iss. 10. – P. 1370–1380.

177. *Chen H. R., Huang H. L.* User Acceptance of Mobile Knowledge Management Learning System: Design and Analysis // Educational Technology & Society. – 2010. – Vol. 13. Iss. 3. – P. 70–77.

178. *Chen Z., Chen N., Yang C., Di L.* Cloud computing enabled Web Processing Service for Earth Observation data processing // IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. – 2012. – Vol. 5. Iss. 6. – P. 1637–1649.

179. *Chen J., Gao J., Li D.* Estimation in semi-parametric regression with non-stationary regressors // Bernoulli. – 2012. – Vol. 18. Iss. 2. – P. 678–702.

180. *Chen T., Guestrin C.* XGBoost: A Scalable Tree Boosting System // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. – 2016. – P. 785–794.

181. *Chen L., Singh V. P., Xiong F.* An Entropy-Based Generalized Gamma Distribution for Flood Frequency Analysis // Entropy. – 2017. Vol. 19. Iss. 6. – Art. No. 239.

182. *Chibisov D. M.* Calculation of the deficiency of asymptotically efficient tests //Theory of Probability and Its Applications. – 1985. – Vol. 30. – P. 289–310.

183. *Christ M., Braun N., Neuffer J., Kempa-Liehr A. W.* Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package) // Neurocomputing. – 2018. – Vol. 307. – P. 72–77.

184. *Cortes C., Vapnik V. N.* Support-vector networks // Machine Learning. – 1995. – Vol. 20. Iss. 3. – P. 273–297.

185. *Costache R., Bui D. T.* Identification of areas prone to flash-flood phenomena using multiple-criteria decision-making, bivariate statistics, machine learning and their ensembles // Science of the Total Environment. – 2020. – Vol. 712. – Art. No. 136492.

186. *Christensen J. H., Boberg F., Christensen O. B., Lucas-Picher P.* On the need for bias correction of regional climate change projections of temperature and precipitation // Geophysical Research Letters. – 2008. – Vol. 35. Iss. 20. – Art. No. L20709.

187. *Ciuperca G., Mercadier C.* Semi-parametric estimation for heavy tailed distributions // Extremes. – 2010. – Vol. 13. Iss. 1. – P. 55–87.

188. *Critchlow T., Kleese van Dam K.* (Eds.) Data-Intensive Science. – London, UK: Chapman and Hall/CRC, 2013. – 446 p.

189. *David A., Vassilvitskii S.* K-means++: The Advantages of Careful Seeding // Proceedings of the 18^{th} Annual ACM-SIAM Symposium on Discrete Algorithms, 2007. – P. 1027–1035.

190. *Davis C.* The norm of the Schur product operation // Numerische Mathematik. – 1962. – Vol. 4. Iss. 1. – P. 343–344.

191. *Diebolt J., Ip E. H.* Stochastic EM: method and application // W. R. Gilks, S. Richardson, D. J. Spiegelhalter (Eds.) Markov Chain Monte Carlo in Practice. – London: Chapman and Hall, 1996.

192. *Dempster A., Laird N., Rubin D.* Maximum likelihood estimation from incompleted data // Journal of the Royal Statistical Society. Series B. – 1977. – Vol. 39. Iss. 1. – P. 1–38.

193. *Donat M., Angelil O., Ukkola A.* Intensification of precipitation extremes in the world's humid and water-limited regions // Environmental Research Letters. – 2019. – Vol. 14. Iss. 6. – Art. No. 065003.

194. *Donoghue J. F.* Phi Scale // Encyclopedia of Estuaries (part of Encyclopedia of Earth Sciences Series) / Ed. by M. J. Kennish. – Dordrecht: Springer, 2016.

195. *Dozat T.* Incorporating Nesterov Momentum into Adam // Proceedings of 4^{th} International Conference for Learning Representations ICLR 2016. 4 p.

196. *Dunn J. C.* A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters // Journal of Cybernetics. – 1973. – Vol. 3. Iss. 3. – P. 32–57.

197. *Embrechts P., Klüppelberg K., Mikosch T.* Modeling Extremal Events. – Berlin: Springer, 1998. – 648 p.

198. *Espinos D. O., Zhidkov A., Kodama R.* Langevin equation for coulomb collision in non-Maxwellian plasmas //Physics of Plasmas. – 2018. – Vol. 25. Iss. 7. – Art. No. 072307.

199. *Fabiani M., Gratton G., Federmeier K.* Event-Related Brain Potentials: Methods, Theory and Applications // Handbook of Psychophysiology. – Cambridge: Cambridge University Press, 2007. – P. 85–119.

200. *Fan J., Wang X., Wu L., Zhou H., Zhang F., Yu X., Lu X., Xiang Y.* Comparison of Support Vector Machine and Extreme Gradient Boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in China // Energy Conversion and Management. – 2018. – Vol. 164. – P. 102–111.

201. *Fernandez-Gonzalez P., Bielza C., Larranaga P.* Random forests for regression as a weighted sum of k-potential Nearest Neighbors
IEEE Access. – 2019. – Vol. 7. – P. 25660–25672.

202. *Fortin J.-M., Currie D. J.* Big Science vs. Little Science: How Scientific Impact Scales with Funding // PLoS ONE. – 2013. Vol. 8. Iss. 6. – Art. No. e65263.

203. *Freedman D., Diaconis P. Z.* On the histogram as a density estimator: L_2 theory // Zeitschrift Fur Wahrscheinlichkeitstheorie und Verwandte Gebiete. – 1981. – Vol. 57. Iss. 4. – P. 453–476.

204. *De Freitas J., Niranjan M., Gee A.* Dynamic Learning with the EM Algorithm for Neural Networks // The Journal of VLSI Signal Processing-Systems for Signal, Image, and Video Technology. – 2000. – Vol. 26. Iss. 1–2. – P. 119–131.

205. *Frenkel S., Gorshenin A., Korolev V.* Adaptive model of data predictability in designing of information systems //Proceedings of the 7^{th} International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT). – Piscataway, NJ, USA: IEEE, 2015. – P. 206–209.

206. *Friedman J. H.* Greedy function approximation: A gradient boosting machine // Annals of Statistics. – 2001. – Vol. 29. Iss. 5. – P. 1189–1232.

207. *Friman O., Volosyak I.,Graser A.* Multiple channel detection of steady-state visual evoked potentials for brain-computer interfaces // IEEE Transactions on Biomedical Engineering. – 2007. – Vol. 54. Iss. 4. – P. 742–

750.

208. *Fritsch F. N., Carlson R. E.* Monotone Piecewise Cubic Interpolation // SIAM Journal on Numerical Analysis. – 1980. – Vol. 17. – P. 238–246.

209. *Galeano P., Pena D.* Data science, big data and statistics // Test. – 2019. – Vol. 28. Iss. 2. – P. 289–329.

210. *Gammaitoni L., Hänggi P., Jung P., Marchesoni F.* Stochastic resonance // Reviews of Modern Physics. – 1988. – Vol. 70. – P. 223–287.

211. *Gao G., Ouyang K., Luo Y., Liang S., Zhou S.* Scheme of Parameter Estimation for Generalized Gamma Distribution and Its Application to Ship Detection in SAR Images // IEEE Transactions on Geoscience and Remote Sensing. – 2017. – Vol. 55. Iss. 3. – P. 1812–1832.

212. *Garland M., Le Grand S., Nickolls J., Anderson J., Hardwick J., Morton S., Phillips E., Zhang Y., Volkov V.* Parallel computing experiences with CUDA // IEEE Micro. – 2008. – Vol. 28. Iss. 4. – P. 13–27.

213. *Gelman A., Carlin J. B., Stern H. S., Dunson D. B., Vehtari A., Rubin D. B.* Bayesian Data Analysis. Third Edition. – Boca Raton, Florida, USA: CRC Press, 2013. – 675 p.

214. *Giabbiconi C M., Dancer C., Zopf R., Gruber T., Muller M. M.* Selective spatial attention to left or right hand flutter sensation modulates the steady-state somatosensory evoked potential // Cognitive Brain Research. – 2004. – Vol. 20. Iss. 1. – P. 58–66.

215. *Gleser L. J.* The gamma distribution as a mixture of exponential distributions // American Statistician. 1989. Vol. 43. P. 115–117.

216. *Glorot X., Bordes A., Bengio Y.* Deep sparse rectifier neural networks // Journal of Machine Learning Research. – 2011. – Vol. 15. – P. 315–323.

217. *Gnedenko B. V., Korolev V. Yu.* Random Summation: Limit Theorems and Applications. – Boca Raton, USA: CRC Press, 1996. 288 p.

218. *Gneiting T., Balabdaoui F., Raftery A. E.* Probabilistic forecasts, calibration and sharpness // Journal of the Royal Statistical Society. Series B-Statistical Methodology. – 2007. – Vol. 69. – P. 243–268.

219. *Goldie C.* A class of infinitely divisible distributions // Mathematical Proceedings of the Cambridge Philosophical Society. – 1967. – Vol. 63. – P. 1141–1143.

220. *O'Gorman P. A., Schneider T.* The physical basis for increases in precipitation extremes in simulations of 21st-century climate change //

Proceedings of the National Academy of Sciences of the United States of America. – 2009. – Vol. 106. Iss. 35. – P. 14773–14777.

221. *Gorshenin A. K.* On information technology for the plasma turbulence research // XXXI International Seminar on Stability Problems for Stochastic Models. Book of Abstracts. – M.: Institute of Informatics Problems, RAS, 2013. – P. 26–28.

222. *Gorshenin A. K.* On Implementation of EM-type Algorithms in the Stochastic Models for a Matrix Computing on GPU // AIP Conference Proceedings. – 2015. – Vol. 1648. – Art. No. 250008 (4 p.)

223. *Gorshenin A. K.* Investigation of Parameters of Meteorological Models Based on Patterns // CEUR Workshop Proceedings. – 2018. – Vol. 2177. – P. 4–10.

224. *Gorshenin A. K.* Software tools for statistical analysis of some precipitation characteristics // Pattern Recognition and Image Analysis. – 2018. – Vol. 28. No. 4. – P. 783–791.

225. *Gorshenin A.* Toward modern educational IT-ecosystems: from learning management systems to digital platforms // Proceedings of the 10^{th} International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT 2018). – Piscataway, NJ, USA: IEEE. – 2018. – P. 329–333. 978-1-5386-9360-5

226. *Gorshenin A. K.* Adaptive detection of normal mixture signals with pre-estimated Gaussian mixture noise // Pattern Recognition and Image Analysis. – 2019. – Vol. 29. No. 3. – P. 377–383.

227. *Gorshenin A., Doynikov A., Korolev V. and Kuzmin V.* Statistical Properties of the Dynamics of Order Books: Empirical Results // XXX International Seminar on Stability Problems for Stochastic Models. Book of Abstracts. – M.: Institute of Informatics Problems, RAS, 2012. – P. 31–51.

228. *Gorshenin A., Frenkel S., Korolev V.* On a stochastic approach to a code performance estimation // AIP Conference Proceedings, 2016. – Vol. 1738. – Art. No. 220010 (4 p.)

229. *Gorshenin A., Korolev V.* Modelling of statistical fluctuations of information flows by mixtures of gamma distributions // Proceedings of 27^{th} European Conference on Modelling and Simulation (May 27-30, 2013, Alesund, Norway). – Dudweiler, Germany: Digitaldruck Pirrot GmbHP. – P. 569–572.

230. *Gorshenin A. K., Korolev V. Yu.* A methodology for the identification of extremal loading in data flows in information systems //

Communications in Computer and Information Science. – 2016. – Vol. 638. – P. 94–103.

231. *Gorshenin A. K., Korolev V. Yu.* A noising method for the identification of the stochastic structure of information flows // Communications in Computer and Information Science. – 2016. – Vol. 678. – P. 279–289.

232. *Gorshenin A. K., Korolev V. Yu.* A functional approach to estimation of the parameters of generalized negative binomial and gamma distributions // Communications in Computer and Information Science. – 2018. – Vol. 919. – P. 353–364.

233. *Gorshenin A. K., Korolev V. Yu.* Scale mixtures of Frechet distributions as asymptotic approximations of extreme precipitation // Journal of Mathematical Sciences. – 2018. – Vol. 234. Iss. 6. – P. 886–903.

234. *Gorshenin A., Korolev V., Kuzmin V., Zeifman A.* Coordinate-wise versions of the grid method for the analysis of intensities of non-stationary information flows by moving separation of mixtures of gamma-distribution // Proceedings of 27^{th} European Conference on Modelling and Simulation (May 27-30, 2013, Alesund, Norway). – Dudweiler, Germany: Digitaldruck Pirrot GmbH. – P. 565–568.

235. *Gorshenin A. K., Korolev V. Yu., Batanov G. M., Skvortsova N. N., Malakhov D. V.* On investigation of the fine structure of processes in low-frequency plasma turbulence // AIP Conference Proceedings. – 2013. – Vol. 1558. – P. 2381–2384.

236. *Gorshenin A. K., Korolev V. Yu., Korchagin A. Yu., Zakharova T. V., Zeifman A. I.* Statistical detection of movement activities in a human brain by separation of mixture distributions // Journal of Mathematical Sciences. – 2016. – Vol. 218. Вып. 3. – P. 278–286.

237. *Gorshenin A., Korolev V., Malakhov. D., Skvortsova N., Shorgin S., Kuzmin V.* On the development of an information technology for plasma turbulence research // Proceedings of 28^{th} European Conference on Modelling and Simulation (May 27-30, 2014, Brescia, Italy). – Dudweiler, Germany: Digitaldruck Pirrot GmbHP. – P. 570–576.

238. *Gorshenin A. K., Korolev V. Yu., Skvortsova N. N., Malakhov D. V.* On non-parametric methodology of the plasma turbulence research // AIP Conference Proceedings. – 2013. – Vol. 1558. – P. 2377–2380.

239. *Gorshenin A. K., Korolev V. Yu., Tursunbaev A. M.* Median modifications of the EM-algorithm for decomposing mixtures of probability

distributions and their applications to the decomposition of volatility of financial indexes // Journal of Mathematical Sciences. – 2018. – Vol. 227. Iss 2. – P. 176–195.

240. *Gorshenin A., Korolev V., Zakharova T., Goncharenko M., Nikiforov S., Khaziakhmetov M., Zeifman A.* On the statistical methods to locate the areas of a human brain activity by the MEG signals and myograms // Proceedings of 29^{th} European Conference on Modelling and Simulation (May 26-29, 2015, Albena (Varna), Bulgaria). – Dudweiler, Germany: Digitaldruck Pirrot GmbH. – P. 631–636.

241. *Gorshenin A. K., Korolev V. Yu., Zeifman A. I.* Modeling particle size distribution in lunar regolith via a central limit theorem for random sums // Mathematics. – 2020. – Vol. 8. Iss. 9. – Art. No. 1409 (24 p.)

242. *Gorshenin A., Kuzmin V.* Online system for the construction of structural models of information flows // Proceedings of the 7^{th} International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT). – Piscataway, NJ, USA: IEEE, 2015. – P. 216–219.

243. *Gorshenin A., Kuzmin V.* On an interface of the online system for a stochastic analysis of the varied information flows // AIP Conference Proceedings. – 2016. – Vol. 1738. – Art. No. 220009 (4 p.)

244. *Gorshenin A. K., Kuzmin V. Yu.* Research support system for stochastic data processing // Pattern Recognition and Image Analysis, 2017. – Vol. 27. No. 3. – P. 518–524.

245. *Gorshenin A. K., Kuzmin V. Yu.* Neural network forecasting of precipitation volumes using patterns // Pattern Recognition and Image Analysis. – 2018. – Vol. 28. No. 3. – P. 450–461.

246. *Gorshenin A. K., Kuzmin V. Yu.* Improved architecture and configurations of feedforward neural networks to increase accuracy of predictions for moments of finite normal mixtures // Pattern Recognition and Image Analysis. – 2019. – Vol. 29. No. 1. – P. 79–88.

247. *Gorshenin A., Kuzmin V.* A machine learning approach to the vector prediction of moments of finite normal mixtures // Advances in Intelligent Systems and Computing. – 2020. – Vol. 1127. – P. 307–314.

248. *Gorshenin A., Lebedeva M., Lukina S., Yakovleva A.* Application of machine learning algorithms to handle missing values in precipitation data // Lecture Notes in Computer Science. – 2019. – Vol. 11965. – P. 563–577.

249. *Gorshenin A. K., Malakhov D. V.* Evolution of histograms and Fourier spectra in structural plasma turbulence in L-2M stellarator // XXX International Seminar on Stability Problems for Stochastic Models. Book of Abstracts. – M.: Institute of Informatics Problems, RAS, 2012. – P. 26–28.

250. *Gorshenin A. K., Shcherbinina A. A.* Efficiency of the method for detecting normal mixture signals with pre-estimated Gaussian mixture noise // Pattern Recognition and Image Analysis. – 2020. – Vol. 30. No. 3. – P. 470–479.

251. *Gould P. G., Koehler A. B., Ord J. K., Snyder R. D., Hyndman R. J., Vahid-Araghi F.* Forecasting time series with multiple seasonal patterns // European Journal of Operational Research. – 2008. – Vol. 191. Iss. 1. – P. 207–222.

252. *Graf J.C.* Lunar Soils Grain Size Catalog // NASA Reference Publication, 1265. – NASA, 1993. – 484 p.

253. *Grandell J.* Mixed Poisson Processes. – London: Chapman and Hall, 1997. – 260 p.

254. *Greenwood M., Yule G. U.* An inquiry into the nature of frequency-distributions of multiple happenings, etc. // J. Roy. Statist. Soc. 1920. Vol. 83. P. 255–279.

255. *Greff K., Srivastava R. K., Koutnik J., Steunebrink B. R., Schmidhuber J.* LSTM: A Search Space Odyssey // IEEE Transactions on Neural Networks and Learning Systems. – 2017. – Vol. 28. Iss. 10. – P. 2222–2232.

256. *Groisman P., Legates D.* Documenting and detecting long-term precipitation trends: Where we are and what should be done // Climate Change. – 1995. – Vol. 31. – P. 601–622

257. *Groisman P. Y., Karl T. R., Easterling D. R. et al.* Changes in the probability of heavy precipitation: important indicators of climatic change // Journal of Climate. – 1999. – Vol. 42. – P. 243–285.

258. *Groisman P., Knight R., Karl T.* Changes in Intense Precipitation over the Central United States // Journal of Hydrometeorology. – 2012. – Vol. 13. Iss. 1. – P. 47–66,

259. *Gulev S. K., Jung T., Ruprecht E.* Estimation of the impact of sampling errors in the VOS observations on air-sea fluxes. Part I. Uncertainties in climate means // Journal of Climate. – 2007. – Vol. 20. – P. 279–301.

260. *Gulev S.K., Jung T., Ruprecht E.* Estimation of the impact of

sampling errors in the VOS observations on air-sea fluxes. Part II. Impact on trends and interannual variability // Journal of Climate. – 2007. – Vol. 20. – P. 302–315.

261. *Gulev S. K., Belyaev K. P.* Probability distribution characteristics for surface air-sea turbulent heat fluxes over the global ocean // Journal of Climate. – 2012. – Vol. 25. Iss. 1. – P. 184–206.

262. *Gulev S. K., Latif M., Keenlyside N., Park W., Koltermann K. P.* North Atlantic Ocean control on surface heat flux on multidecadal timescales // Nature. – 2013. – Vol. 499. – P. 464–467.

263. Guo J., Zhang H., Zhen D., Shi Z., Gu F., Ball A. D. An enhanced modulation signal bispectrum analysis for bearing fault detection based on non-Gaussian noise suppression // Measurement. –2020. – Vol. 151. – Art. No. 107240.

264. *Halmos P. R.* Finite-Dimensional Vector Spaces. – Princeton: Princeton University Press, 1948. – 202 p.

265. *Han Z.-C., Lin J.-G., Zhao Y.-Y.* Adaptive semiparametric estimation for single index models with jumps // Computational Statistics & Data Analysis. – 2020. – Vol. 151. – Art. No.. 107013.

266. *Hartley H.* Maximum likelihood estimation from incomplete data // Biometrics. – 1958. – Vol. 14. – P. 174–194.

267. *Hartmann D.* Cross-shore selective sorting process and grain size distributional shape // Aeolian Grain Transport. Acta Mechanica Supplementum – 1991. – Vol. 2. – P. 49–63.

268. *Hey T., Gannon D., Pinkelman J.* The Future of Data-Intensive Science // Computer. – 2012. – Vol. 45. Iss. 5. – P. 81–82.

269. *Hu X., Song L.* Hydrodynamic modeling of flash flood in mountain watersheds based on high-performance GPU computing // Natural Hazards. – 2018. – Vol. 91. Iss. 2. – P. 567–586.

270. *Huang J., Ling C. X.* Using AUC and accuracy in evaluating learning algorithms // IEEE Transactions on Knowledge and Data Engineering. – 2005. – Vol. 17. Iss. 3. – P. 299–310.

271. *Huang P.-H., Hwang T. Y.* New moment estimation of parameters of the generalized gamma distribution using it's characterization // Taiwanese Journal of Mathematics. – 2006. – Vol. 10. Iss. 4. – P. 1083–1093.

272. *Ilter M. C., Sokun H. U., Yanikomeroglu H., Wichman R., Hamalainen J.* The Joint Impact of Fading Severity, Irregular Constellation, and Non-Gaussian Noise on Signal Space Diversity-Based Relaying

Networks // IEEE Access. – 2019. – Vol. 7. – P. 116162–116171.

273. *Iosup A., Ostermann S., Yigitbasi M. N., Prodan R., Fahringer T., Epema D. H. J.* Performance analysis of cloud computing services for many-tasks scientific computing // IEEE Transactions on Parallel and Distributed Systems. – 2011. – Vol. 22. Iss. 6. – P. 931–945.

274. *Ivanov M. V., Levitsky L. I., Bubis J. A., Gorshkov M. V.* Scavager: A Versatile Postsearch Validation Algorithm for Shotgun Proteomics Based on Gradient Boosting // Proteomics. – 2019. – Vol. 19. Iss. 3. – Art. No. 1800280.

275. *Johnson N., Kot S., Balakrishnan N.*. Continuous Univariate Distributions, Vol. 2, 2nd Edition. – New York: Wiley, 1995. – 752 p.

276. *Jordan M. I., Mitchell T. M.* Machine learning: Trends, perspectives, and prospects // Science. – 2015. – Vol. 349. Iss. 6245. – P. 255–260.

277. *Josey S. A.* A comparison of ECMWF, NCEP-NCAR and SOC surface heat fluxes with moored buoy measurements in the subduction region of the Northeast Atlantic // Journal of Climate. – 2001. – Vol. 14. – P. 1780–1789.

278. *Kadir S. N., Goodman D. F. M., Harris K. D.* High-Dimensional Cluster Analysis with the Masked EM Algorithm // Neural Computation. – 2014. – Vol. 26. Iss. 11. – P. 2379–2394.

279. *Kalashnikov V.* Geometric Sums: Bounds for Rare Events with Applications. – Dordrecht: Kluwer Academic Publishers, 1997. – 270 p.

280. *Kalteh A., Hjorth P.* Imputation of missing values in a precipitation-runoff process database // Hydrology Research. – 2009. Vol. 40. Iss. 4. – P. 420–432.

281. *Kardan A. A., Sadeghi H., Ghidary S. S., Sani M. R. F.* Prediction of student course selection in online higher education institutes using neural network // Computers & Education. – 2013. – Vol. 65. – P. 1–11.

282. *Kates-Harbeck J., Svyatkovskiy A., Tang W.* Predicting disruptive instabilities in controlled fusion plasmas through deep learning // Nature. – 2019. – Vol. 568. Iss. 7753. – P. 526–531.

283. *Kelling S., Hochachka W. M., Fink D., Riedewald M., Caruana R., Ballard G., Hooker G.* Data-intensive Science: A New Paradigm for Biodiversity Studies //Bioscience. – 2009. – Vol. 59. Iss. 7. – P. 613–620.

284. *Kharin V. V., Zwiers F. W., Zhang X., Hegerl G. C.* Changes in temperature and precipitation extremes in the IPCC ensemble of global

coupled model simulations // Journal of Climate. – 2007. – Vol. 20. Iss. 8. – P. 1419–1444.

285. *Kingma D., Ba J.* Adam: A Method for Stochastic Optimization // Conference Paper at the 3^{rd} International Conference for Learning Representations (ICLR 2015) // arXiv:1412.6980. – 2015. – 13 p.

286. *Kingman J. F. C.* Poisson processes. – Oxford: Clarendon Press, 1993.

287. *Korolev V. Yu.* A general theorem on the limit behavior of superpositions of independent random processes with applications to Cox processes // Journal of Mathematical Sciences. – 1996. – Vol. 81. Iss. 5. – P. 2951–2956.

288. *Korolev V.* On convergence of distributions of compound Cox processes to stable laws // Theory of Probability and its Applications. – 1999. – Vol. 43. Iss. 4. – P. 644–650.

289. *Korolev V. Yu., Skvortsova N. N.* (Eds) Stochastic Models of Structural Plasma Turbulence. Untrecht: VSP, 2006.

290. *Korolev V.* Product representations for random variables with the Weibull distributions and their applications // Journal of Mathematical Sciences. – 2016. – Vol. 218. Iss. 3. – P. 298–313.

291. *Korolev V. Yu., Gorshenin A. K.* Probability models of statistical regularities in rainfall data // XXXV International Seminar on Stability Problems for Stochastic Models. Book of Abstracts. – Perm: Perm State University, 2018. – P. 52–54.

292. *Korolev V. Yu., Gorshenin A. K.* Probability models and statistical tests for extreme precipitation based on generalized negative binomial distributions // Mathematics. – 2020. – Vol. 8. Iss. 4. – Art. No. 604 (30 p.)

293. *Korolev V. Yu., Gorshenin A. K., Belyaev K. P.* Statistical tests for extreme precipitation volumes // Mathematics. – 2019. – Vol. 7. Iss. 7. – Art. No. 648 (20 p.)

294. *Korolev V. Yu., Gorshenin A. K., Gulev S. K., Belyaev K. P.* Statistical modeling of air-sea turbulent heat fluxes by finite mixtures of Gaussian distributions // Communications in Computer and Information Science. – 2015. – Vol. 564. – P. 152–162.

295. *Korolev V. Yu., Gorshenin A. K., Gulev S. K., Belyaev K. P., Grusho A. A.* Statistical Analysis of Precipitation Events // AIP Conference Proceedings. – 2017. – Vol. 1863. – Art. No. 090011 (4 p.).

296. *Korolev V., Gorshenin A., Korchagin A., Zeifman A.* Generalized

gamma distributions as mixed exponential laws and related limit theorems // Proceedings of 31^{st} European Conference on Modelling and Simulation (May 23-26, 2017, Budapest, Hungary). – Dudweiler, Germany: Digitaldruck Pirrot GmbHP. – P. 642–648.

297. *Korolev V. Yu., Sokolov I. A., Gorshenin A. K.* Max-compound Cox processes. I // Journal of Mathematical Sciences, 2019. – Vol. 237. Iss. 6. – P. 789–803.

298. *Korolev V. Yu., Zeifman A. I.* On convergence of the distributions of random sequences with independent random indexes to variance–mean mixtures // Stochastic Models. – 2016. – Vol. 32. Iss. 3. – P. 414–432.

299. *Korolev V. Yu., Zeifman A. I.* Generalized negative binomial distributions as mixed geometric laws and related limit theorems // Lithuanian Mathematical Journal. – 2019. –. Vol. 59. – P. 1461–1466.

300. *Kose U., Arslan A.* Optimization of self-learning in computer engineering courses: an intelligent software system supported by artificial neural network and vortex optimization algorithm // Computer Applications in Engineering Education. – 2017. – Vol. 25. Iss. 1. – P. 142–156.

301. *Kosko B., Mitaim S.* Stochastic resonance in noisy threshold neurons // Neural Networks. – 2003. – Vol. 16. Iss. 5. – P. 755–761.

302. *Krzysztofowicz R.* The case for probabilistic forecasting in hydrology // Journal of Hydrology. – 2001. – Vol. 249. Iss. 1–4. – P. 2–9.

303. *Kullback S., Leibler R. A.* On Information and Sufficiency // Annals of Mathematical Statistics. – 1951. – Vol. 22. – P. 79–86.

304. *Kunkel K. E., Karl T. R., Easterling D. R. et al.* Probable maximum precipitation and climate change // Geophysical Research Letters. – 2013. – Vol. 40. Iss. 7. – P. 1402–1408.

305. *Kurbanmuradov O., Sabelfeld K.* Lagrangian stochastic models for turbulent dispersion in the atmospheric boundary layer // Boundary-Layer Meteorology. – 2000. – Vol. 97. Iss. 2. – P. 191–218.

306. *Kysely J., Picek J., Beranova R.* Estimating extremes in climate change simulations using the peaks-over-threshold method with a non-stationary threshold // Global and Planetary Change. – 2010. – Vol. 72. Iss. 1-2. – P. 55–68.

307. *Lagarias J.C., Reeds J.A., Wright M.H., Wright P.E.* Convergence Properties of the Nelder-Mead Simplex Method in Low Dimensions // SIAM Journal of Optimization. – 1998. – Vol. 9. Iss. 1. – P. 112–147.

308. *Leadbetter M. R.* On a basis for «Peaks over Threshold» modeling // Statistics & Probability Letters. – 1991. – Vol. 12. Iss. 4. – P. 357–362.

309. *LeCam L.* Asymptotic Methods in Statistical Decision Theory. – New York: Springer, 1986. – 742 p.

310. *LeCun Y., Bengio Y.,Hinton G.* Deep learning // Nature. – 2015. – Vol. 521. Iss. 7553. – P. 436–444.

311. *Lee C. A., Gasster S. D., Plaza A., Chang C.-I., Huang B.* Recent Developments in High Performance Computing for Remote Sensing: A Review // IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. – 2011. – Vol. 4. Iss. 3. – P. 508–527.

312. *Lee G., Scott C.* EM algorithms for multivariate Gaussian mixture models with truncated and censored data // Computational Statistics & Data Analysis. – 2012. – Vol. 56. Iss. 9. – P. 2816–2829.

313. *Lee S. X., Leemaqz K. L., McLachlan G. J.* A Block EM Algorithm for Multivariate Skew Normal and Skew t-Mixture Models // IEEE Transactions on Neural Networks and Learning Systems. – 2018. – Vol. 29. Iss. 11. – P. 5581–5591.

314. *Hodges J. L., Lehmann E. L.* The efficiency of some nonparametric competitors of the t-test // Annals of Mathematical Statistics. – 1956. – Vol. 27. – P. 324–335.

315. *Hodges J. L., Lehmann E. L.* Comparison of the normal scores and Wilcoxon tests // Proceedings of 4th Berkeley Symposium. – 1960. – Vol. 1. – P. 307–317.

316. *Hodges J. L., Lehmann E. L.* Deficiency // Annals of Mathematical Statistics. – 1970. – Vol. 41. – P. 783–801.

317. *Li Y., Li Z., Wei K., Xiong W., Yu J., Qi B.* Noise Estimation for Image Sensor Based on Local Entropy and Median Absolute Deviation // Sensors. – 2019. – Vol. 19. Iss. 2. – Art. No. 339.

318. *Liu C., Li H.-C., Fu K., Zhang F., Datcu M., Emery W. J.* A robust EM clustering algorithm for Gaussian mixture models // Pattern Recognition. – 2019. – Vol. 87. – P. 269–284.

319. *Lo Y., Mendell N. R., Rubin D. B.* Testing the number of components in a normal mixture // Biometrika. – 2001. – Vol. 88. Iss. 3. – P. 767–778.

320. *Lo Y.* Likelihood ratio tests of the number of components in a normal mixture with unequal variances // Statistics and Probability Letters. – 2005. – Vol. 71. – P. 225–235.

321. *Loéve M.* Probability Theory. – New York: Springer, 1977. – 704 p.

322. *Lonn S., Teasley S. D.* Saving time or innovating practice: Investigating perceptions and uses of Learning Management Systems // Computers & Education. – 2009. – Vol. 53. Iss. 3. – P. 686–694.

323. *Lu F., Song J., Cao X., Zhu X.* CPU/GPU computing for long-wave radiation physics on large GPU clusters // Computers & Geosciences. – 2012. – Vol. 41. – P. 47–55.

324. *Marquez-Figueroa S., Shmaliy Y. S., Ibarra-Manzano O.* Optimal extraction of EMG signal envelope and artifacts removal assuming colored measurement noise // Biomedical Signal Processing and Control. – 2020. – Vol. 57. – Art. No. 101679.

325. *McArthur D. S.* Distinctions between grain-size distribution of accretion and encroachment deposits in an inland dune // Sedimentary Geology. – 1987. – Vol. 54. – P. 147–163.

326. *MacQueen J.* Some methods for classification and analysis of multivariate observations // Proceedings of the 5^{th} Berkeley Symposium on Mathematical Statistics and Probability. – 1967. – P. 281–297.

327. *Maia J. D. C., Carvalho G. A. U., Mangueira C. P., Santana S. R., Cabral L. A. F., Rocha G. B.* GPU linear algebra libraries and GPGPU programming for accelerating MOPAC semiempirical quantum chemistry calculations // Journal of Chemical Theory and Computation. – 2012. – Vol. 8. Iss. 9. – P. 3072–3081.

328. *Malakhov D., Skvortsova N., Gorshenin A., Korolev V., Chirkov A., Tedtoev B.* Spectral analysis and modeling of non-Gaussian processes of structural plasma turbulence // XXXII International Seminar on Stability Problems for Stochastic Models. Book of Abstracts. – M.: Institute of Informatics Problems, RAS, 2014. – P. 68–72.

329. *Malakhov D. V., Skvortsova N. N., Gorshenin A. K., Korolev V. Yu., Chirkov A. Yu., Konchekov E. M., Kharchevsky A. A.* On a spectral analysis and modeling of non-Gaussian processes in the structural plasma turbulence // Journal of Mathematical Sciences. – 2016. – Vol. 218. Вып. 2. – P. 208–215.

330. *Mann M. E., Lees J. M.* Robust estimation of background noise and signal detection in climatic time series // Climatic Change. –1996. – Vol. 33. Iss. 3. – P. 409–445.

331. *Martinez-Villalobos C., Neelin J.* Why Do Precipitation Intensities Tend to Follow Gamma Distributions? // Journal of the Atmospheric

Sciences. – 2019. – Vol. 76. – P. 3611–3631,

332. *der Maur A.N.F.* Statistical tools for drop size distributions: Moments and generalized gamma // Journal of the Atmospheric Sciences. – 2001. – Vol. 58. Iss. 4. – P. 407–418.

333. *McGillem C. D., Aunon J. I.* Analysis of Event-Related Potentials. – Methods of Analysis of Brain Electrical and Magnetic Signals: EEG Handbook. A.S. Gevins, A. Remond (Eds.). Amsterdam: Elsevier Science Publishers, 1987. – P. 131–169.

334. *McLachlan G. J., Lee S. X., Rathnayake S. I.* Finite Mixture Models // Annual Review of Statistics and Its Application. – 2019. – Vol. 6. – P. 355–378.

335. *Meneghini O., Luna C. J., Smith S. P., Lao L. L.* Modeling of transport phenomena in tokamak plasmas with neural networks // Physics of Plasmas. – 2014. – Vol. 21. Iss. 6. – Art. No. 060702.

336. *Mesbah A., Graves D. B.* Machine learning for modeling, diagnostics, and control of non-equilibrium plasmas // Journal of Physics D-Applied Physics. – 2019. – Vol. 52. Iss. 30. – Art. No. 30LT02.

337. *Michener W. K., Jones M. B.* Ecoinformatics: supporting ecology as a data-intensive science // Trends in Ecology & Evolution. – 2012. – Vol. 27. Iss. 2. – P. 85–93.

338. *Mo C., Ruan Y., He J., Jin J., Liu P., Sun G.* Frequency analysis of precipitation extremes under climate change // International Journal of Climatology. – 2019. – Vol. 39. – P. 1373–1387.

339. *Mustapha I. B., Saeed F.* Bioactive Molecule Prediction Using Extreme Gradient Boosting // Molecules. – 2016. – Vol. 21. Iss. 8. – Art. No. 983.

340. *Narita E., Honda M., Nakata M., Yoshida M., Hayashi N., Takenaga H.* Neural-network-based semi-empirical turbulent particle transport modelling founded on gyrokinetic analyses of JT-60U plasmas // Nuclear Fusion. – 2019. – Vol. 59. Iss. 10. – Art. No. 106018.

341. *Newman H. B., Ellisman M. H., Orcutt J. A.* Data-intensive e-science - Frontier research // Communications of the ACM. – 2003. – Vol. 46. Iss. 11. – P. 67–75.

342. *Ng S. K., McLachlan G. J.* Using the EM algorithm to train neural networks: Misconceptions and a new algorithm for multiclass classification // IEEE Transactions on Neural Networks. – 2004. – Vol. 15. Iss. 3. – P. 738–749.

343. *Nielsen S. F.* Stochastic EM algorithm: Estimation and asymptotic results // Bernoulli. – 2000. – Vol. 6. – P. 457–489.

344. *Noether G. E.* On a theorem of Pitman // Annals of Mathematical Statistics. – 1955. – Vol. 26. – P. 64–68.

345. *Oreizy P., Medvidovic N., Taylor R. N.* Runtime software adaptation: framework, approaches, and styles // Proceedings of ICSE-08. – 2008. – P. 899-910.

346. *Osoba O., Mitaim S., Kosko B.* The noisy Expectation-Maximization algorithm // Fluctuation Noise Letters. – 2013. – Vol. 12. Iss. 3. – Art. No. 1350012.

347. *Ozkan S., Koseler R.* Multi-dimensional students' evaluation of e-learning systems in the higher education context: An empirical investigation // Computers & Education. – 2009. – Vol. 53. Iss. 4. – P. 1285–1296.

348. *Page E. S.* On problems in which a change in a parameter occurs at an unknown point // Biometrika. – 1957. – Vol. 44. Iss. 1–2. – P. 248–252.

349. *Papadrakakis M., Stavroulakis G., Karatarakis A.* A new era in scientific computing: Domain decomposition methods in hybrid CPU-GPU architectures // Computer Methods in Applied Mechanics and Engineering. – 2011. – Vol. 200. Iss. 13–16. – P. 1490–1508.

350. *Parsons M. S.* Interpretation of machine-learning-based disruption models for plasma control // Plasma Physics and Controlled Fusion. – 2017. – Vol. 59. Iss. 8. – Art. No. 085001.

351. *Paulsen M. F.* Experiences with Learning Management Systems in 113 European institutions // Educational Technology & Society. – 2003. Vol. 6. Iss. 4. P. 134–148.

352. *Picard D.* Testing and estimating change-points in time series // Advances in Applied Probability. – 1985. – Vol. 6. – P. 841–867.

353. *Pickands J.* Statistical inference using extreme order statistics // Annals of Statistics. – 1975. – Vol. 3. – P. 119–131.

354. *Pinson P., Madsen H., Nielsen H. A., Papaefthymiou G., Klockl B.* From Probabilistic Forecasts to Statistical Scenarios of Short-term Wind Power Production // Wind Energy. – 2009. – Vol. 12. Iss. 1. – P. 51–62.

355. *Pitman E. J. G..* Lecture notes on nonparametric statistical inference: lectures given for the University of North Carolina. – North Carolina: Institute of Statistics, 1948. – 77 p.

356. *Portmann R. W., Solomon S., Hegerl G. C.* Spatial and seasonal

patterns in climate change, temperatures, and precipitation across the United States // Proceedings of the National Academy of Sciences of the United States of America. – 2009. – Vol. 106. Iss. 18. – P. 7324–7329.

357. *Prokhorenkova L., Gusev G., Vorobev A., Dorogush A. V., Gulin A.* CatBoost: unbiased boosting with categorical features // Advances in Neural Information Processing Systems. – 2018. – Vol. 31. – P. 6638–6648.

358. *Pshenichnikov A. A., Kolik L. V., Malykh N. I., Petrov A. E. et al.* The use of Doppler reflectometry in the L-2M stellarator // Plasma Phys. Rep. – 2005. – Vol. 31. No. 7. – P. 554–561.

359. *Punmiya R., Choe S.* Energy Theft Detection Using Gradient Boosting Theft Detector With Feature Boost Engineering-Based Preprocessing // IEEE Transactions on Smart Grid. – 2019. – Vol. 10. Iss. 2. – P. 2326–2329.

360. *Qin X., Zou H., Zhou S., Ji K.* Region-Based Classification of SAR Images Using Kullback-Leibler Distance Between Generalized Gamma Distributions // IEEE Geoscience and Remote Sensing Letters. 2015. – Vol. 12. Iss. 8. – P. 1655–1659.

361. *Quarteroni A.* The role of statistics in the era of big data: A computational scientist' perspective // Statistics & Probability Letters. – 2018. – Vol. 136. – P. 63–67.

362. *Quinn R. A., Goree J.* Single-particle Langevin model of particle temperature in dusty plasmas // Physical Review E. – 2000. – Vol. 61. Iss. 3. – P. 3033–3041.

363. *Raja M. A. Z., Shah F. H., Tariq M., Ahmad I., Ahmad S. U.* Design of artificial neural network models optimized with sequential quadratic programming to study the dynamics of nonlinear Troesch's problem arising in plasma physics // Neural Computing & Applications. – 2018. – Vol. 29. Iss. 6. – P. 83–109.

364. *Rapuano S., Zoino F.* A learning management system including laboratory experiments on measurement instrumentation // IEEE Transactions on Instrumentation and Measurement. – 2006. – Vol. 55. Iss. 5. – P. 1757–1766.

365. *Reed W. J., Jorgensen M.* The double Pareto-Lognormal distribution – a new parametric model for size distribution // Communications in Statistics – Theory and Methods. – 2004. – Vol. 33. Iss. 8. – P. 1733–1753.

366. *Rényi A.* On an extremal property of the Poisson process // Annals

of the Institute of Statistical Mathematics. – 1964. – Vol. 16. – P. 129–133.

367. *Rojas-Dominguez A., Padierna L. C., Valadez J. M. C., Puga-Soberanes H. J., Fraire H. J.* Optimal Hyper-Parameter Tuning of SVM Classifiers With Application to Medical Diagnosis // IEEE Access. – 2018. – Vol. 6. – P. 7164–7176.

368. *Romero C., Ventura S., Garcia E.* Data mining in course management systems: Moodle case study and tutorial // Computers & Education. – 2008. – Vol. 51. Iss. 1. – P. 368–384.

369. *Roth M., Buishand T. A., Jongbloed G., Tank A. M. G., van Zanten J. H.* A regional peaks-over-threshold model in a nonstationary climate // Water Resources Research. – 2012. – Vol. 48. – Art. No. W11533.

370. *Sabelfeld K. K.* A stochastic model and Monte Carlo algorithm for fluctuation-induced H-2 formation on the surface of interstellar dust grains // Journal of Cosmology and Astroparticle Physics. – 2015. – Vol. 9. – Art. No. 061.

371. *Sabelfeld K. K.* Stochastic simulation algorithms for solving narrow escape diffusion problems by introducing a drift to the target // Journal of Computational Physics. – 2020. – Vol. 410. – Art. No. 109406.

372. Sattari M., Rezazadeh-Joudi A., Kusiak A. Assessment of different methods for estimation of missing data in precipitation studies // Hydrology Research. – 2017. – Vol. 48. Iss. 4. – P. 1032–1044.

373. *Schwartz G.* Estimating the dimension of a model // The Annals of Statistics. – 1978. – Vol. 6. – P. 461–464.

374. *Schmidhuber J., Bengio Y.,Hinton G.* Deep learning in neural networks: An overview // Neural Networks. – 2015. – Vol. 61. – P. 85–117.

375. *Service R. F.* Materials Scientists Look To a Data-Intensive Future // Science. – 2012. – Vol. 335. Iss. 6075. – P. 1434–1435.

376. *Sexty D.* Calculating the equation of state of dense quark-gluon plasma using the complex Langevin equation // Physical Review D. – 2019. – Vol. 100. Iss. 7. – Art. No. 074503.

377. *Sichel H,S.* On a family of discrete distributions particularly suited to represent long tailed frequency data // Proceedings of the 3rd Symposium on Mathematical Statistics. – Pretoria: CSIR, 1971. – P. 51–97.

378. *Simolo C., Brunetti M., Maugeri M., Nanni T.* Improving estimation of missing values in daily precipitation series by a probability density function-preserving approach // International Journal of Climatology. –2010. – Vol. 30. Iss. 10. – P. 1564–1576.

379. *Sipin A. S.* On stochastic algorithms for solving boundary-value problems for the Laplace operator // Journal of Mathematical Sciences. – 2017. – Vol. 225. Iss. 5. – P. 812–817.

380. *Sipin A.* Monte Carlo Algorithms for the Parabolic Cauchy Problem // Mathematics. – 2019. – Vol. 7. Iss. 2. – Art. No. 177.

381. *Sipin A. S., Zeifman A. I.* Numerical experiments for some Markov models for solving boundary value problems // Lecture Notes in Computer Science. – 2019. – Vol. 11386. – P. 493–500.

382. *Skvortsova N. N., Akulina D. K., Batanov G. M. et al.* Effect of ECRH Regime on Characteristics of Short-Wave Turbulence in Plasma of the L-2 // Plasma Phys. Control. Fusion. – 2010. – Vol. 52. 055008.

383. *Skvortsova N. N., Chirkov A. Yu., Kharchevsky A. A., Malakhov D. V., Gorshenin A. K., Korolev V. Yu.* Doppler reflectometry studies of plasma gradient instabilities in L-2M stellarator // Journal of Physics: Conference Series. – 2016. – Vol. 666. – Art. No. 012007 (7 p.)

384. *Skvortsova N. N, Korolev V. Y., Batanov G M.,et al.* Statistical analysis and modelling of turbulent fluxes in the plasma of the L-2M stellarator and the FT-2 tokamak // Plasma Physics and Controlled Fusion. – 2006. – Vol. 48. – Art. No. A393–9 .

385. *Small R. J., Bryan F. O., Bishop S. P., Tomas R. A.*: Air-Sea Turbulent Heat Fluxes in Climate Models and Observational Analyses: What Drives Their Variability? // Journal of Climate. – 2019. – Vol. 32. Iss. 8. – P. 2397–2421.

386. *Song K.-S.* Globally convergent algorithms for estimating generalized gamma distributions in fast signal and image processing // IEEE Transactions on Image Processing. – 2008. – Vol. 17. Iss. 8. – P. 1233–1250.

387. *Sørensen M.* On the Size Distribution of Sand: Working paper. – Copenhagen: Department of Applied Mathematics and Statistics, University of Copenhagen. – 2006. – P. 1–11.

388. *Sportouche H., Nicolas J.-M., Tupin F.* Mimic Capacity Of Fisher And Generalized Gamma Distributions For High-Resolution SAR Image Statistical Modeling // IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. – 2017. – Vol. 10. Iss. 12. – P. 5695–5711.

389. *Srivastava N., Hinton G., Krizhevsky A., Sutskever I., Salakhutdinov R.* Dropout: A Simple Way to Prevent Neural Networks from Overfitting // Journal of Machine Learning Research. – 2014. – Vol. 15. –

P. 1929–1958.

390. *Stacy E. W.* A generalization of the gamma distribution. Annals of Mathematical Statistics. – 1962. – Vol. 33. – P. 1187–1192.

391. *Stein D. W. J.* Detection of random signals in Gaussian mixture noise // IEEE Transactions on Information Theory. – 1995. – Vol. 41. Iss. 6. – P. 1788–1801.

392. *Steinhaus H.* Sur la division des corps materiels en parties // Bulletin de L'Academie Polonaise Des Sciences. – 1956. – Vol. 4. Iss. 12. – P. 801–804.

393. *Stineman R. W.* A Consistently Well Behaved Method of Interpolation // Creative Computing. – 1980. – Vol. 6. Iss. 7. – P. 54–57.

394. *Stopa J. E., Cheung K. F., Tolman H. L., Chawla A.* Patterns and cycles in the Climate Forecast System Reanalysis wind and wave data // Ocean Modelling. – 2013. – Vol. 70. – P. 207–220.

395. *Strauch M., Bernhofer C., Koide S., Volk M., Lorz C., Makeschin F.* Using precipitation data ensemble for uncertainty analysis in SWAT streamflow simulation // Journal of Hydrology. – 2012. – Vol. 414. – P. 413–424.

396. *Styan G. P. H.* Hadamard Products and Multivariate Statistical Analysis // Linear Algebra and its Applications. – 1973. – Vol. 6. – P. 217–240.

397. *Sultan N.* Cloud computing for education: A new dawn? // International Journal of Information Management. – 2010. – Vol. 30. Iss. 2. – P. 109–116.

398. *Sun P.-Ch., Tsai R. J., Finger G., Chen Y.-Y., Yeh D.* What drives a successful e-Learning? An empirical investigation of the critical factors influencing learner satisfaction // Computers & Education. – 2008. – Vol. 50. Iss. 4. – P. 1183–1202.

399. *Szabo C., Sheng Q. Z., Kroeger T., Zhang Y., Yu J.* Science in the Cloud: Allocation and Execution of Data-Intensive Scientific Workflows // Journal of Grid Computing. – 2014. – Vol. 12. Iss. 2. – P. 245–264.

400. *Tang F., Ishwaran H.* Random forest missing data algorithms // Statistical Analysis and Data Mining. – 2017. – Vol. 10. Iss. 6. – P. 363–377.

401. *Tang Y.* Beyond EM: A faster Bayesian linear regression algorithm without matrix inversions // Neurocomputing. – 2020. – Vol. 378. – P. 435–440.

402. *Teegavarapu R., Aly A., Pathak C., Ahlquist J., Fuelberg H., Hood J.* Infilling missing precipitation records using variants of spatial interpolation and data-driven methods: use of optimal weighting parameters and nearest neighbour-based corrections // International Journal of Climatology. – 2018. – Vol. 38. Iss. 12. – P. 776–793.

403. *Teicher H.* Identifiability of mixtures // Annals of Mathematical Statistics. – 1961. – Vol. 32. – P. 244–248.

404. *Teicher H.* Identifiability of Finite Mixtures // The Annals of Mathematical Statistics. – 1963. – Vol. 34. Вып. 4. – P. 1265–1269.

405. *Tian G., Xia Y., Zhang Y., Feng D.* Hybrid Genetic and Variational Expectation-Maximization Algorithm for Gaussian-Mixture-Model-Based Brain MR Image Segmentation // IEEE Transactions on Information Technology in Biomedicine. – 2011. – Vol. 15. Iss. 3. – P. 373–380.

406. *Tikhonov A. N., Leonov A. S., Yagola A. G.* Nonlinear Ill-Posed Problems. – Heidelberg: Springer, 1998. – 386 p.

407. *Torres-Barran A., Alonso A., Dorronsoro J. R.* Regression tree ensembles for wind energy and solar radiation prediction // Neurocomputing. – 2019. – Vol. 326. – P. 151–160.

408. *Trenberth K. E.* Changes in precipitation with climate change // Climate Research. – 2011. – Vol. 47. Iss. 1–2. – P. 123–138.

409. *Tuac Y., Guney Y., Arslan O.* Parameter estimation of regression model with AR(p) error terms based on skew distributions with EM algorithm // Soft Computing. – 2020. – Vol. 24. Iss. 5. – P. 3309–3330.

410. *Uppu S., Krishna A.* A deep hybrid model to detect multi-locus interacting SNPs in the presence of noise // International Journal of Medical Informatics. – 2018. – Vol. 119. – P. 134–151.

411. *Ushakov N. G., Ushakov V. G.* Statistical analysis of rounded data: recovering of information lost due to rounding // Journal of the Korean Statistical Society. – 2017. – Vol. 46. No. 3. – P. 426–437.

412. *Vapnik V. N.* An overview of statistical learning theory // IEEE Transactions on Neural Networks. – 1999. – Vol. 10. Iss. 5. – P. 988–999.

413. *Vasilieva M., Gorshenin A., Korolev V.* Statistical analysis of probability characteristics of precipitation in different geographical regions // Advances in Intelligent Systems and Computing. – 2020. – Vol. 902. – P. 629–639.

414. *Verbeek J. J., Vlassis N., Krose B.* Efficient greedy learning of Gaussian mixture models // Neural Computation. – 2003. – Vol. 15. Iss. 2. –

P. 469–485.

415. *Villaverde J. E., Godoy D., Amandi A.* Learning styles' recognition in e-learning environments with feed-forward neural networks // Journal of Computer Assisted Learning. – 2006. – Vol. 22. Iss. 3. – P. 197–206.

416. *Vincent P.* Differentiation of modern beach and coastal dune sands – a logistic regression approach using the parameters of the hyperbolic function // Sedimentary Geology. – 1986. – Vol. 49. – P. 167–176.

417. *Viroli C., McLachlan G. J.* Deep Gaussian mixture models // Statistics and Computing. – 2019. – Vol. 29. Iss. 1. – P. 43–51.

418. *Vuola O., Hameri A.-P.* Mutually benefiting joint innovation process between industry and big-science // Technovation. – 2006. – Vol. 26. Iss. 1. – P. 3–12.

419. *Waltz R. A., Morales J. L., Nocedal J., Orban D.* An interior algorithm for nonlinear optimization that combines line search and trust region steps // Mathematical Programming. – 2006. – Vol. 107. Iss. 3. – P. 391–408.

420. *Wang J. J. J., Chan J. S. K., Choy S. T. B.* Stochastic volatility models with leverage and heavytailed distributions: A Bayesian approach using scale mixtures // Computational Statistics & Data Analysis. – 2011. – Vol. 55. No. 1. – P. 852–862.

421. *Wang Y., Chee C.-S.* Density estimation using non-parametric and semi-parametric mixtures // Statistical Modelling. – 2012. – Vol. 12. No. 1. – P. 67–92.

422. *Wright D. E., Bray I.* A mixture model for rounded data // Journal of the Royal Statistical Society Series D – The Statistician. – 2003. – Vol. 52. – P. 3–13.

423. *Wu X., Kumar V., Quinlan J., et al.* Top 10 algorithms in data mining // Knowledge and Information Systems. – 2008. – Vol. 14. Iss. 1. – P. 1–37.

424. *Wu D., Ma J.* An effective EM algorithm for mixtures of Gaussian processes via the MCMC sampling and approximation // Neurocomputing. – 2019. – Vol. 331. – P. 366–374.

425. *Xia Y., Liu C., Li Y., Liu N.* A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring // Expert Systems with Applications. – 2017. – Vol. 78. – P. 225–241.

426. *Xie J., Yang C., Zhou B., Huang Q.* High-performance computing for the simulation of dust storms // Computers Environment and Urban

Systems. – 2010. – Vol. 34. – P. 278–290.

427. *Xu C., Qiao Y., Jian M.* Interdecadal Change in the Intensity of Interannual Variation of Spring Precipitation over Southern China and Possible Reasons // Journal of Climate. – 2013. – Vol. 32. – P. 5865–5881.

428. *Yang X., Deka S., Righetti R.* A hybrid CPU-GPGPU approach for real-time elastography // IEEE Transactions on Ultrasonics Ferroelectrics and Frequency Control. – 2011. – Vol. 58. Iss. 12. – P. 2631–2645.

429. *Yang M.-Sh., Lai Ch.-Yo, Lin C.-Y.* A robust EM clustering algorithm for Gaussian mixture models // Pattern Recognition. – 2012. – Vol. 45. Iss. 11. – P. 3950–3961.

430. *Yu L.* Global variations in oceanic evaporation (1958-2005): The role of the changing wind speed // Journal of Climate. – 2007. – Vol. 20. – P. 5376–5390.

431. *Yu L., Weller R. A.* Objectively analyzed air-sea heat fluxes for the global ice-free oceans (1981–2005) // Bulletin of the American Meteorological Society. – 2007. – Vol. 88. – P. 527–539.

432. *Yu L., Yang T., Chan A. B.* Density-Preserving Hierarchical EM Algorithm: Simplifying Gaussian Mixture Models for Approximate Inference // IEEE Transactions On Pattern Analysis and Machine Intelligence. – 2019. – Vol. 41. Iss. 6. – P. 1323–1337.

433. *Zatsarinny A., Gorshenin A., Kondrashev V., Volovich K., Denisov S.* Toward high performance solutions as services of research digital platform // Procedia Computer Science. – 2019. – Vol. 150. – P. 622–627.

434. *Zeiler M. D.* ADADELTA: An Adaptive Learning Rate Method // arXiv:1212.5701. – 2012. – 6 p.

435. *Zeller C. B., Cabral C. R. B., Lachos V. H., Benites L.* Finite mixture of regression models for censored data based on scale mixtures of normal distributions // Advances in Data Analysis and Classification. – 2019. – Vol. 13. Iss. 1. – P. 89–116.

436. *Zeng D., Lin D. Y.* Maximum likelihood estimation in semiparametric regression models with censored data // Journal of the Royal Statistical Society. Series B-Statistical Methodology. – 2007. – Vol. 69. – P. 507–536.

437. *Zhang B., Liu T., Bai Z. D.* Analysis of rounded data from dependent sequences // Annals of the Institute of Statistical Mathematics. – 2010. – Vol. 62. Iss. 6. – P. 1143–1173.

438. *Zhao N., Bai Z.* Analysis of rounded data in mixture normal

model // Statistical Papers. – 2012. – Vol. 53. – P. 895–914.

439. *Zhao W., Li J., Yang X., Peng Q., Wang J.* Innovative CFAR detector with effective parameter estimation method for generalised gamma distribution and iterative sliding window strategy // IET Image Processing. – 2018. – Vol. 12. Iss. 1. – P. 60–69.

440. *Zheng M., Tang W., Zhao X.* Hyperparameter optimization of neural network-driven spatial models accelerated using cyber-enabled high-performance computing // International Journal of Geographical Information Science. – 2019. – Vol. 33. Iss. 2. – P. 314–345.

441. Zhou Y., Zhu H. Segmentation Using a Trimmed Likelihood Estimator in the Asymmetric Mixture Model Based on Generalized Gamma and Gaussian Distributions // Mathematical Problems in Engineering. – 2018. – Art. No. 3468967.

442. *Zolina O., Simmer C., Kapala A., Bachner S., Gulev S., Maechel H.* Seasonally dependent changes of precipitation extremes over Germany since 1950 from a very dense observational network // Journal of Geophysical Research. – 2008. – Vol. 113. – Art. No. D06110.

443. *Zolina O., Simmer C., Belyaev K., Kapala A., Gulev S. K.* Improving estimates of heavy and extreme precipitation using daily records from European rain gauges // Journal of Hydrometeorology. – 2009. – Vol. 10. – P. 701–716.

444. *Zolina O., Simmer C., Belyaev K., Gulev S., Koltermann P.* Changes in the duration of European wet and dry spells during the last 60 years // Journal of Climate. – 2013. – Vol. 26. – P. 2022–2047.

445. *Zolina O., Simmer C., Belyaev K., Kapala A., Gulev S. K., Koltermann P.* Multidecadal trends in the duration of wet spells and associated intensity of precipitation as revealed by a very dense observational German network // Environmental Research Letters. – 2014. – Vol. 9. Iss. 2. – Art. No. 025003.

446. *Zolina O., Simmer C., Kapala A., Gulev S. K.* On the robustness of the estimates of centennial-scale variability in heavy precipitation from station data over Europe // Geophysical Research Letters. – 2005. – Vol. 32. – Art. No. L14707.

447. *Zolotarev V.* Modern Theory of Summation of Random Variables. – Utrecht: VSP, 1997. – 412 p.