Министерство науки и высшего образования
Российской Федерации

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

ГОРШЕНИН Андрей Константинович

Автореферат диссертации
на соискание ученой степени
доктора физико–математических наук

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

Специальность 05.13.18 — математическое моделирование,
численные методы и комплексы программ

Работа выполнена в отделении 6
«Стохастические и интеллектуальные методы и средства
моделирования и построения систем с интенсивным
использованием данных» ФИЦ ИУ РАН

Научный консультант: **Королев Виктор Юрьевич**,
доктор физико-математических наук, профессор, заведующий
кафедрой математической статистики факультета ВМК МГУ
имени М. В. Ломоносова

Москва – 2020

Общая характеристика работы

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`)[1]. 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[2,3,4].

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 задач, 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.

[1] *Critchlow T., Kleese van Dam K.* (Eds.) Data-Intensive Science. – London, UK: Chapman and Hall/CRC, 2013. – 446 p.

[2] *Bzdok D., Altman N., Krzywinski M.* Statistics versus machine learning // Nature Methods, 2018. Vol. 15. Iss. 4. P. 232–233.

[3] *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

[4] *Galeano P., Pena D.* Data science, big data and statistics // Test, 2019. Vol. 28. Iss. 2. P. 289–329.

Процесс накопления данных зачастую протекает в условиях неопределенности, обусловленной: а) стохастическим характером интенсивностей потоков информативных событий и взаимодействием большого числа не поддающихся исчерпывающему прогнозированию факторов, которые можно считать случайными; б) неоднородностью или нестационарностью изучаемых закономерностей; в) 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[5]. 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[6], Б. В. Гнеденко[7], И. А. Ибрагимова, Ю. В. Линника[8], Ю. В. Прохорова[9], А. Н. Ширяева[10], В. М. Золотарева[11], В. В. Калашникова[12], В. В. Петрова[13], В. М. Круглова[14],

[5] *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.

[6] *Колмогоров А. Н.* Избранные труды. Том 2: Теория вероятностей и математическая статистика. – М.: Наука, 2005 –581 с.

[7] *Гнеденко Б. В., Колмогоров А. Н.* Предельные распределения для сумм независимых случайных величин. – М.-Л.: ГИТТЛ, 1949. –264 с.

[8] *Ибрагимов И. А., Линник Ю. В.* Независимые и стационарно связанные величины. – М.: Наука, 1965. – 524 с.

[9] *Прохоров Ю. В.* Избранные труды. – М.: Торус Пресс, 2012. – 775 с.

[10] *Ширяев А. Н.* Вероятность-1. – М.: МЦНМО, 2017. – 552 с.

[11] *Zolotarev V.* Modern Theory of Summation of Random Variables. – Utrecht: VSP, 1997. – 412 p.

[12] *Kalashnikov V.* Geometric Sums: Bounds for Rare Events with Applications. – Dordrecht: Kluwer Academic Publishers, 1997, 270 p.

[13] *Петров В. В.* Суммы независимых случайных величин. – М.: Наука, 1972. – 416 с.

[14] *Круглов В. М., Королев В. Ю.* Предельные теоремы для случайных сумм. – М.: Изд-во Моск. ун-та, 1990. – 269 с.

В. Ю. Королева[15]. 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[15,16]. Известны многочисленные применения смешанных вероятностных моделей в различных прикладных задачах, например, для описания процессов в турбулентной плазме, при анализе финансовых данных, в процессе обработки изображений в медицине, в ряде социологических исследований.

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[17,18] 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[19] в 1977 году. Данный метод получения оценок максимального правдоподобия применялся еще в 1958 году Х. Хартли 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 интеллектуаль-

[15] *Королев В. Ю.* Вероятностно-статистические методы декомпозиции волатильности хаотических процессов. – М.: Изд-во Моск. ун-та, 2011. – 512 с.

[16] *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.

[17] *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.

[18] *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. **107013**.

[19] *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.

ного анализа данных[20].

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, регрессии, обработки цензурированных и усеченных данных, оценивания параметров различных распределений и процессов, в том числе с организацией параллельных вычислительных алгоритмов и обучением нейронных сетей. Однако в процессе модификации обычно сохраняется общий принцип наличия E- (от `expectation`) и M-шагов (от `maximization`). Например, в стохастическом (`SEM`) варианте алгоритма[21,22] вводится дополнительный 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[23] 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[24].

Ряд модификаций направлен на повышения скорости сходимости. Так, в статье[25] предложено введение дополнительного «зашумляющего» этапа, улучающего эффективность метода примерно на 10–15%. Идея введения подобной модификации основана на явлении стохастического резонанса, 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 ав-

[20] *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.

[21] *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.

[22] *Nielsen S. F.* Stochastic EM algorithm: Estimation and asymptotic results // Bernoulli, 2000. Vol. 6. P. 457–489.

[23] *Viroli C., McLachlan G. J.* Deep Gaussian mixture models // Statistics and Computing, 2019. Vol. 29. Iss. 1. P. 43–51.

[24] *Audhkhasi K., Osoba O., Kosko B.* Noise-enhanced convolutional neural networks // Neural Networks, 2016. Vol. 78. P. 15–23.

[25] *Osoba O., Mitaim S., Kosko B.* The noisy Expectation-Maximization algorithm // Fluctuation Noise Letters, 2013. Vol. 12. Iss. 3. **1350012**.

томатизации 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 обнаружения и отслеживания эволюции структуры формирующих стохастических процессов в рамках процедуры, называемой методом скользящего разделения смесей (СРС)[15]. 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 СРС-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 процессов (в среднем от 3 до 5), формирующих ионно-звуковую турбулентность в плазме. Также получены значимые результаты в области анализа волатильности финансовых индексов.

Возможны и другие подходы к построению стохастических моделей различных явлений, в том числе на основе стохастических дифференциальных уравнений (СДУ)[26,27,28]. 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[29], ассимиляции данных при анализе разномасштабной изменчивости геофизических

[26] *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.

[27] *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. **061**.

[28] *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.

[29] *Ширяев А. Н.* Основы стохастической финансовой математики. Т. 1. Факты. Модели. – М.: МЦНМО, 2016. – 440 с.

переменных[30], взаимодействия частиц в плазме[31]. 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[32]. 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 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 обучения и нейронные сети, особенно глубокие, стали одним из наиболее востребованных и эффективных инструментов всестороннего анализа данных[33]. Существенный вклад в их развитие внесли, в частности, В. Н. Вапник и А. Я. Червоненкис[34], Я. Лекун, И. Бенджио и Дж. Хинтон[35]. Подобные процедуры успешно применяются для обработки наблюдений в самом широком спектре областей, включая метеорологию, финансы, медицину и многие другие. При этом получение прорывных результатов обеспечивается не только выбором

[30] *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.

[31] *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. **074503.**

[32] *Espinos D. O., Zhidkov A., Kodama R.* Langevin equation for coulomb collision in non-Maxwellian plasmas // Physics of Plasmas, 2018. Vol. 25. Iss. 7. **072307**.

[33] *Jordan M. I., Mitchell T. M.* Machine learning: Trends, perspectives, and prospects // Science, 2015. Vol. 349. Iss. 6245. P. 255–260.

[34] *Вапник В. Н., Червоненкис А. Я.* Теория распознавания образов. – М.: Наука, 1974. –416 с.

[35] *LeCun Y., Bengio Y.,Hinton G.* Deep learning // Nature, 2015. Vol. 521. Iss. 7553. P. 436–444.

подходящих типов архитектур и настройкой гиперпараметров[36], 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 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[37]. 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[38]. Применение гетерогенных вычислений[39] 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 и гидродинамическое моделирование, геопространственный анализ данных, квантовохимические вычисления.

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

[36] *Bergstra J., Bengio Y.* Random Search for Hyper-Parameter Optimization // Journal of Machine Learning Research, 2012. Vol. 13. P. 281–305.

[37] *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.

[38] *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.

[39] *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.

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.

Соответствие паспорту специальности

Исследования и полученные результаты соответствуют следующим пунктам паспорта специальности 05.13.18 – математическое моделирование, численные методы и комплексы программ:

– разработка новых математических методов моделирования объектов и явлений (п. 1);

– развитие качественных и приближенных аналитических методов исследования математических моделей (п. 2);

– разработка, обоснование и тестирование эффективных вычислительных методов с применением современных компьютерных технологий (п. 3);

– реализация эффективных численных методов и алгоритмов в виде комплексов проблемно-ориентированных программ для проведения вычислительного эксперимента (п. 4);

– комплексные исследования научных и технических проблем с применением современной технологии математического моделирования и вычислительного эксперимента (п. 5);

– разработка новых математических методов и алгоритмов интерпретации натурного эксперимента на основе его математической модели (п. 7).

Научная новизна и основные результаты диссертации

В диссертации разработаны эффективные 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 по-

мощью 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 явлений 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 and International Workshop «Applied Problems in Theory of Probabilities and Mathematical Statistics related to modeling of information systems» (ISSPSM): 2012–2014, 2018, 2020 гг. [77, 83, 104, 106, 113];

– European Conference on Modelling and Simulation (ECMS): 2013–2015, 2017 гг. [85, 90, 93, 95];

– International Conference of Numerical Analysis and Applied Mathematics (ICNAAM): 2013–2016 гг. [78, 84, 91, 94, 98, 110];

– International Conference on Modern Techniques of Plasma Diagnostics and their Application: 2014 г. [72, 115];

– International Congress on Ultra Modern Telecommunications and Control Systems (ICUMT): 2015, 2018 гг. [76, 81, 97];

– International Scientific Conference on Information Technologies and Mathematical Modelling (ITMM): 2015, 2016 гг. [86, 109];

– International Conference on Distributed Computer and Communication Networks: Control, Computation, Communications (DCCN): 2016, 2018, 2019 гг. [87, 88, 103];

– International Conference of Artificial Intelligence, Medical Engineering, Education (AIMEE): 2018, 2020 гг. [116];

– International Symposium «Intelligent Systems» (INTELS): 2018 г. [117];

– International Symposium on Computer Science, Digital Economy and Intelligent Systems (CSDEIS): 2019, 2020 гг. [102];
– Международная Звенигородская конференция по физике плазмы и управляемому термоядерному синтезу: 2013, 2015 гг. [73, 74];
– Международная научно-методическая конференция «Информатизация инженерного образования» (ИНФОРИНО): 2014, 2016 гг. [5, 14];
– Всероссийская конференция (с международным участием) «Информационно-телекоммуникационные технологии и математическое моделирование высокотехнологичных систем»: 2016, 2018 гг. [47, 79];
– Всероссийская научная конференция «Ломоносовские чтения»: 2018–2020 гг. [50];
– Всероссийский Симпозиум по прикладной и промышленной математике: 2014, 2015, 2019 гг. [35, 44, 45];
– Всероссийская научно-практическая конференция с международным участием «Актуальные проблемы глобальных исследований: Россия в глобализирующемся мире»: 2019 г. [43];
– научная конференция «Тихоновские чтения»: 2015 г. [69];
– 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 стелларатоpе 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 [1, 4–6, 13, 14, 19, 20, 24, 25, 35, 41–45, 47, 49–55, 58–60, 62, 64–117], из них:
– **31** статья в журналах, включенных в перечень ВАК [1, 4, 6, 13, 19, 20, 24, 25, 41, 42, 49, 51–55, 58–60, 62, 64, 66–68, 70, 80, 82, 99–101, 105];
– **51** статья в изданиях, индексируемых базами Web of Science Core Collection и/или Scopus [13, 19, 20, 24, 25, 49, 51, 53, 59, 60, 62, 64, 68, 75, 76, 78–82, 84–103, 105, 107–112, 114–117], включая журналы первого и второго квартилей [68, 75, 96, 103, 107, 108].

Получены **35** свидетельств о государственной регистрации про-

грамм для ЭВМ [2,3,7–12,15–18,21–23,26–34,36–40,46,48,56,57,61,63], зарегистрированных в Федеральной службе по интеллектуальной собственности (Роспатент).

Личный вклад автора

Основные результаты диссертации получены лично автором. В работах [41–45, 54, 55, 58–60, 62, 64–66, 84, 85, 93, 97–105, 116, 117] А. К. Горшениным выполнены постановка исследовательских задач, определение ключевых концепций и методов решения, а также проведен всесторонний анализ полученных результатов. В работах [47, 49–53, 67–76, 83, 86–92, 94–96, 106–115] 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 [46, 48, 56, 57, 61, 63] А. К. Горшениным реализованы алгоритмы анализа данных в виде значимых компонентов зарегистрированных инструментов.

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.

Содержание работы

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

В **первой главе** 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 (GNB) как смешанного пуассоновского со смешивающим обобщенным гамма-распределением (GG).

Определение 1.1. Случайная величина (с. в.) $N_{r,\gamma,\mu}$, $r > 0$, $\gamma \in \mathbb{R}$, $\mu > 0$, имеет дискретное распределение, называемое *обобщенным отрицательным биномиальным*, если оно для всех целых значений k определяется вероятностями

$$\mathbb{P}(N_{r,\gamma,\mu} = k) = \frac{1}{k!} \int_0^\infty e^{-z} z^k f^{GG}_{r,\gamma,\mu}(x)\, dz,$$

то есть является смешанным пуассоновским со смешивающим GG-распределением $f^{GG}_{r,\gamma,\mu}(x) = \frac{|\gamma|\mu^r}{\Gamma(r)} x^{\gamma r - 1} e^{-\mu x^\gamma}$, $x \geqslant 0$.

Получены рекуррентные представления для данного распределения и формулы для математического ожидания и дисперсии (утверждения 1.1 и 1.2).

В §1.3 доказана теорема об асимптотическом распределении максимальной порядковой статистики в выборке, объем которой является обобщенной отрицательной биномиальной с. в.. Здесь и далее символ $\stackrel{d}{=}$ обозначает равенство по распределению, W_λ, $\lambda > 0$ – с. в. с распределением Вейбулла, $Q_{r,1}$ – с. в. с распределением Снедекора-Фишера, $\overline{G}_{r,\gamma,\mu}$, $r > 0$, $\gamma \in \mathbb{R}$, $\mu > 0$, и $G_{r,\mu}$– с. в. с обобщенным и классическим гамма-распределениями, соотвественно. Всюду далее подразумевается, что рассматриваемые с. в. независимы.

Теорема 1.1. *Пусть $n \in \mathbb{N}$, $r > 0$, $\gamma > 0$, $\mu > 0$ и $N_{r,\gamma,\mu/n^\gamma}$ – с. в., имеющая обобщенное отрицательное биномиальное распределение. Пусть $X_1, X_2, \ldots$ – независимые одинаково распределенные с. в. с общей функцией распределения (ф. р.) $F(x)$. Предположим, что $\operatorname{rext}(F) = \sup\{x : F(x) < 1\} = \infty$ и существует такое число $\lambda > 0$, что при любом $x > 0$ справедливо соотношение $\lim\limits_{y\to\infty} \frac{1-F(xy)}{1-F(y)} = x^{-\alpha}$. 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 $F_{\lambda,\gamma,\mu,r}(x) = \int_0^\infty e^{-zx^{-\lambda}} f^{GG}_{r,\gamma,\mu}(z)\, dz \equiv \mathbb{P}(M_{\lambda,\gamma,\mu,r} < x), \quad x \geqslant 0.$

При этом предельная с. в. $M_{\lambda,\gamma,\mu,r}$ допускает представления:

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

Для важного частного случая GNB-распределения – классического отрицательного биномиального – предельная ф. р. имеет простой функциональный вид.

Теорема 1.2. *Пусть выполнены условия теоремы 1.1, однако объем выборки является отрицательным биномиальным, то есть рассматривается с. в. 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*

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

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

При этом предельная с. в. $M_{\lambda,\mu,r}$ допускает представления:

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

Рассмотрим более подробно свойства предельного распределения в теореме 1.1. Определим с. в. $Z_{r,1} \stackrel{d}{=} \left(1 + \frac{1-r}{r} Q_{1-r,r}\right)$ и обозначим через $S_{\gamma,1}$ с. в. со строго устойчивым распределением, а через Π_λ – с. в. с распределением Парето.

Теорема 1.3. *Распределение с. в. $M_{\lambda,\gamma,\mu,r}$ представимо в виде:*

(i) Если $r \in (0,1]$, то $M_{\lambda,\gamma,\mu,r} \stackrel{d}{=} (\mu Z_{r,1})^{-1/\lambda\gamma} \cdot W_{\lambda\gamma}/W_\gamma$.

(ii) Если $\gamma \in (0,1]$, то $M_{\lambda,\gamma,\mu,r} \stackrel{d}{=} \left((\mu r)^{-1} S_{\gamma,1} \cdot Q_{r,1}\right)^{1/\lambda\gamma}$.

(iii) Если $\gamma \in (0,1]$ и $r \in (0,1]$, то $M_{\lambda,\gamma,\mu,r} \stackrel{d}{=} \Pi_\lambda \left(S_{\gamma,1} Z_{r,1}^{1/\gamma} \right)^{-1/\lambda}$.

(iv) Если $r \in (0,1]$ и $\lambda\gamma \in (0,1]$, то

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

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

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

$$\mathbb{E}M_{\lambda,\gamma,\mu,r}^\delta = \Gamma\Big(r + \frac{\delta}{\lambda\gamma}\Big)\Gamma\Big(1 - \frac{\delta}{\lambda}\Big)\Big(\mu^{\delta/\lambda\gamma}\Gamma(r)\Big)^{-1}.$$

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*

$$F(x) = 1 - \frac{c}{ax^\lambda + c}, \quad x \geqslant 0. \tag{1.1}$$

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

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

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

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

При выполнении условий теоремы 1.2 получен явный вид предельной ф. р. и для произвольных порядковых статистик (теорема 1.7). Кроме того в предположении существования плотностей у элементов выборки, выборочные квантили имеют распределение Стьюдента в качестве предельного (теорема 1.8).

В §1.3 также доказан закон больших чисел для сумм с обобщенным отрицательным биномиальным распределением (обобщение теоремы Реньи). Здесь и далее символ $\Longrightarrow$ обозначает слабую сходимость.

Теорема 1.9. *Пусть для с. в.* $X_1, X_2, \ldots$ *(не обязательно независимых и одинаково распределенных) при* $n \to \infty$ *выполнено условие* $n^{-\beta} \sum_{j=1}^{n} X_j \Longrightarrow a$ *для некоторых конечных параметров* $\beta > 0$ *и* $a > 0$. *Пусть величины* $r > 0$, γ *и* $\mu > 0$ *произвольны. Пусть для каждого* $n \in \mathbb{N}$ $N_{r,\gamma,\mu/n^\gamma}$ *– с. в., имеющая обобщенное отрицательное биномиальное распределение, независимая от последовательности* $X_1, X_2, \ldots$. *Тогда*

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

Результаты этого раздела используются далее 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 предельных распределений возникают произвольные нормальные смеси. Пусть $\{X_{n,j}\}_{j \geqslant 1}$, $n \in \mathbb{N}$, схема серий построчно независимых необязательно одинаково распределенных с. в. с ф. р. $F_{n,j}(x)$.

Обозначим $S_{n,k} = X_{n,1} + ... + X_{n,k}, \quad n, k \in \mathbb{N}$. Независимость строк $\{S_{n,k}\}_{k \geqslant 1}$ не предполагается. Пусть $\mu_{n,j} = \mathbb{E}X_{n,j}$, $\sigma^2_{n,j} = \mathbb{D}X_{n,j}$, причем $0 < \sigma^2_{n,j} < \infty$, $n, j \in \mathbb{N}$. Дисперсию с. в. $S_{n,k}$ обозначим $B^2_{n,k} = \sigma^2_{n,1} + ... + \sigma^2_{n,k}$.

Теорема 1.10. *Пусть выполнено случайное условие Линдеберга*

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

Тогда сходимость $Z_n = \frac{S_{n,N_n}-c_n}{d_n} \Longrightarrow Z = Y \cdot U + V$ *при* $n \to \infty$, *где* $(U_n = \frac{b_{n,N_n}}{d_n}, V_n = \frac{a_{n,N_n}-c_n}{d_n}) \Longrightarrow (U, V)$ *и с. в.* Y *не зависит от пары* (U, V), *имеет место для некоторых* $a_{n,k} \in \mathbb{R}$, $b_{n,k} > 0$, $c_n \in \mathbb{R}$ *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* $\lim\limits_{n\to\infty} L_2\left((U_n, V_n), (U'_n, V'_n)\right) = 0$, *где* L_2 – *некоторая вероятностная метрика, метризующая слабую сходимость, и* $\mathbb{P}(Z < x) = \mathbb{E}\Phi\left(\frac{x-V'_n}{U'_n}\right)$, $x \in \mathbb{R}$.

Данный результат используется в дальнейшем в главе 4 для обоснования вида моделей для размеров частиц лунного реголита.

Вторая глава посвящена исследованию аналитических свойств моделей на основе конечных нормальных и гамма-распределений. В §2.1 для статистической оценки распределений случайных коэффициентов в стохастическом дифференциальном уравнении Ланжевена вида $dX(t) = a(t)dt + b(t)dW$, которое определяет некоторый случайный процесс $X(t)$, где $W(t)$ – винеровский процесс, а коэффициенты сдвига (дрейфа) $a(t)$ и масштаба (диффузии) $b(t)$ – случайны. 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 $i \geqslant 1$. Тогда можно использовать дискретную аппроксимацию $\mathbb{P}\left(X(t_i) - X(t_{i-1}) < x\right) \approx \sum_{k=1}^{K} p_k \Phi\left(\frac{x-a_k}{b_k}\right)$, то есть модель конечной смеси нескольких нормальных распределений с параметрами, изменяющимися при переходе от t_i к t_{i+1}. Для их статистического оценивания предложено использовать метод скользящего разделения смесей (СРС). На основе получаемых оценок коэффициентов возможно содержательно расширять признаковое пространство в методах машинного обучения за счет использования характеристик адекватных математических моделей. Соответствующие примеры рассмотрены в §5.2 для экспериментальных данных в физике плазмы.

В §2.2 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, полученная автором в кандидатской диссертации. 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.

В §2.2 и §2.3 рассмотрены две важные модели возмущений параметров смеси – добавления и расщепления компоненты – и приведены полученные автором в кандидатской диссертации 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 эти результаты 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).

Предположим, что каждое из независимых наблюдений $X_1, \ldots, X_n$ имеет распределение типа конечной сдвиговой смеси нормальных законов $G(x) = \mathbb{E}\Phi(x-V)$, где $\Phi(\cdot)$ – ф. р. стандартного нормального закона и V – дискретная с. в., принимающая значения a_i с вероятностями p_i. 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. Для каждой из них в даном параграфе выписаны в явном виде двусторонние оценки, связывающие расстояния Леви, которое будет обозначаться $L(\cdot,\cdot)$, между смесями и смешивающими законами. В качестве примера рассмотрим один из полученных результатов для модели добавления компоненты, в которой наблюдения имеют распределение $G_p(x) = (1-p)\sum\limits_{i=1}^{k} p_i\Phi(x-a_i)+p\Phi(x-a)$, 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 можно считать, что выполнены соотношения $a_0 \leqslant a \leqslant a_1 \leqslant a_2 \leqslant \ldots \leqslant a_k$, означающие, что рассматриваются конечные математические ожидания. Поэтому параметр a_0 может полагаться известным. Данной модели соответствует дискретная случайная величина V_p, принимающая значения a_i с вероятностями

$p_i(1-p)$ и a с вероятностью p.

Теорема 2.8. *В модели добавления компоненты*

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

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

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

Теоремы 2.8–2.11 обосновывают корректность аппроксимации произвольных сдвиговых нормальных смесей, которые в общем случае не являются идентифицируемыми, конечными аналогами в задаче их статистического разделения.

В §2.5 получены результаты об устойчивости дисперсионно-сдвиговых смесей нормальных законов вида

$$\Phi_{\alpha,\sigma,F_A}(x)=\int\limits_0^\infty \Phi\left(\frac{x-\alpha u}{\sigma\sqrt{u}}\right)dF_A(u),\quad \alpha\in\mathbb{R},\quad \sigma>0,$$

где $F_A\,(u)$ – ф. р. положительной с вероятностью единица с. в., относительно возмущений смешивающего распределения.

Теорема 2.12. *Предположим, что F_A и F_B – ф. р. с точками роста, расположенными на неотрицательной полуоси, и по крайней мере F_A имеет плотность, ограниченную некоторым числом $0<a<\infty$. Тогда $L(\Phi_{\alpha,\sigma,F_A},\Phi_{\alpha,\sigma,F_B})\leqslant 2(1+a)L(F_A,F_B)$.*

Таким образом, близость смешивающих распределений в смысле расстояния Леви необходимо влечет и близость соответствующих смесей. Полученные результаты могут быть использованы для обоснования вычислительных процедур разделения дисперсионно-сдвиговых смесей нормальных законов.

В §2.6 разработаны теоретические подходы к устранению ошибок в специальной смешанной модели округления данных. Пусть $X_1,X_2,\ldots$ – независимые одинаково распределенные с. в. с неизвестным математическим ожиданием $E_X<+\infty$; $\varepsilon_1,\varepsilon_2,\ldots$ – независимые одинаково распределенные с. в. с математическим ожиданием $E_\varepsilon<+\infty$; $X_1,X_2,\ldots$ и $\varepsilon_1,\varepsilon_2,\ldots$ являются независимыми; $Y_j=\left[X_j+\varepsilon_j+\frac{1}{2}\right]$ 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) с математическим ожиданием $E_Y<+\infty$. В данных предположениях 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 для дисперсии (теоремы 2.14 и 2.16). Приведем формулировки только некоторых результатов.

Теорема 2.13. *Пусть случайные величины* ε_j, $j = 1, 2, \ldots$, *имеют распределение типа конечной* k*-компонентной смеси нормальных законов с параметрами* $\mathbf{a}$, $\boldsymbol{\sigma}$ *и* $\mathbf{p}$. *Тогда* $|E_Y - E_X| \leqslant A + \left(1 + \frac{1}{4\pi^2\sigma^2}\right) e^{-2\pi^2\sigma^2}$ *где* $A = \max(|a_1|, \ldots, |a_k|)$, $\sigma = \min(\sigma_1, \ldots, \sigma_k)$.

Теорема 2.14. *В условиях и обозначениях теоремы 2.13 и в предположении, что случайные величины* $X_j \overset{\text{п. н.}}{=} E_X$, $j = 1, 2, \ldots$, *доверительный интервал для* E_X *уровня* $1 - \alpha$, $0 < \alpha < 1$, *имеет вид* $\left[\hat{E}_X - f(\mathbf{a}, \boldsymbol{\sigma}, \alpha, n), \hat{E}_X + f(\mathbf{a}, \boldsymbol{\sigma}, \alpha, n)\right]$, *где* $\hat{E}_X = \frac{1}{n}\sum\limits_{j=1}^{n}\left[E_X + \varepsilon_j + \frac{1}{2}\right]$, $f(\mathbf{a}, \boldsymbol{\sigma}, \alpha, n) = \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}}\left(\sqrt{A^2 + \Sigma^2} + \frac{1}{2}\right) + A + \frac{1}{\pi}\left(1 + \frac{1}{4\pi^2\sigma^2}\right) e^{-2\pi^2\sigma^2}$, $z_{1-\frac{\alpha}{2}}$ – $\left(1 - \frac{\alpha}{2}\right)$*-квантиль стандартного нормального распределения и* $\Sigma = \max(\sigma_1, \ldots, \sigma_k)$.

Соответствующие соотношения во всех случаях зависят только от «экстремальных» значений параметров смесей, но не от числа компонент и весов в распределении зашумляющих наблюдений.

В **третьей главе** 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). Приведем одну из них.

Теорема 3.2. *Моменты случайной величины* Z_n *с распределением типа конечной нормальной смеси для использования в CPC-методе в матричной записи имеют следующий вид:*

- *математическое ожидание:* $\mathbb{E}Z_n = \mathbf{p}_n\,\mathbf{a}_n^T$;
- *дисперсия:* $\mathbb{D}Z_n = \mathbf{p}_n\left(D_{\mathbf{a}_n}\,\mathbf{a}_n^T + D_{\boldsymbol{\sigma}_n}\,\boldsymbol{\sigma}_n^T\right) - (\mathbf{p}_n\,\mathbf{a}_n^T)^2$;
- *коэффициент ассиметрии:*

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

- *коэффициент эксцесса:*

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

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

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

Эти результаты существенным образом используются при анализе процессов в физике турбулентной плазмы в §5.2 и §5.3, а также в океанологии в §6.5.

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 (теоремы 3.3 и 3.4). Введем следующие обозначения: $\widetilde{A} = \widetilde{\mathbf{a}}\,\mathbf{1}_{\widetilde{k}\times 1}$, $\widetilde{\Sigma} = \widetilde{\boldsymbol{\sigma}}\,\mathbf{1}_{\widetilde{k}\times 1}$, $\mathcal{E} = \bigoplus_{r=1}^{k} \mathbf{1}_{\widetilde{k}\times 1}$, $\widetilde{\mathbf{a}} = (a_1, \ldots, a_{\widetilde{k}})$, $\widetilde{\boldsymbol{\sigma}} = (\sigma_1^2, \ldots, \sigma_{\widetilde{k}}^2)$, $\widehat{\mathbf{p}}_r = (\widehat{p}_{(r-1)\widetilde{k}+1}, \ldots, \widehat{p}_{r\widetilde{k}})$, $\widehat{\mathbf{a}}_r = (a_{(r-1)\widetilde{k}+1}, \ldots, a_{r\widetilde{k}})$, $\widehat{\boldsymbol{\sigma}}_r = (\sigma^2_{(r-1)\widetilde{k}+1}, \ldots, \sigma^2_{r\widetilde{k}})$, $\widetilde{\mathbf{p}}_r^{-1} = (\widetilde{p}_1^{-1}, \ldots, \widetilde{p}_{\widetilde{k}}^{-1})$, $r = \overline{1,k}$; $\mathbf{p} = (p_1, \ldots, p_k)$, $\mathbf{a} = (a_1, \ldots, a_k)$, $\boldsymbol{\Sigma} = (\sigma_1, \ldots, \sigma_k)$, $\widehat{\mathbf{p}} = (\widehat{\mathbf{p}}_1 \cdots \widehat{\mathbf{p}}_k)$, $\widehat{\mathbf{a}} = (\widehat{\mathbf{a}}_1 \cdots \widehat{\mathbf{a}}_k)$, $\widehat{\boldsymbol{\sigma}} = (\widehat{\boldsymbol{\sigma}}_1 \cdots \widehat{\boldsymbol{\sigma}}_k)$. Символ $\bigoplus$ обозначает прямую сумму соответствующих матриц, таким образом, $\mathcal{E}$ имеет блочно-диагональную структуру (элементы – векторы из единиц размера $\widetilde{k}$). В теореме ниже символ $\circ$ обозначает произведение Адамара.

Теорема 3.4. *Оценки метода наименьших квадратов (МНК) параметров неизвестного смешанного распределения сигнала X на фоне смешанного гаусовского шума имеют вид:*

$$\mathbf{p} = \widetilde{k}^{-1}\left[\left(\widetilde{\mathbf{p}}_1^{-1}\,\widetilde{\mathbf{p}}_2^{-1} \cdots \widetilde{\mathbf{p}}_k^{-1}\right) \circ \widehat{\mathbf{p}}\right]\,\mathcal{E}, \quad \mathbf{a} = \widetilde{k}^{-1}\left(\widehat{\mathbf{a}}\,\mathcal{E} - \widetilde{A}\,\mathbf{1}_{1\times k}\right),$$
$$\boldsymbol{\Sigma} = \widetilde{k}^{-1}\left(\widehat{\boldsymbol{\sigma}}\,\mathcal{E} - \widetilde{\Sigma}\,\mathbf{1}_{1\times k}\right),$$

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. Важную роль в данной процедуре играют методы получения оценок максимального правдоподобия – они требуют тонкой настройки и оказывают существенное влияние на результаты анализа. 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.

В §3.3 разработан алгоритм последовательной идентификации (определения локальной связности) компонент смесей вероятностных распределений. В его основу положена комбинация жадного

алгоритма для поиска числа компонент и одного из методов кластеризации (например, k-средних или нечеткая). Данный метод используется для статистического определения числа формирующих процессов в турбулентной плазме в разделе 5.2.2, а также для статистического оценивания распределений случайных коэффициентов СДУ Ланжевена для потоков тепла между океаном и атмосферой в разделе 6.5.3. Предложенная процедура может быть естественным образом расширена на случай многомерных смешанных распределений.

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 его высокая эффективность.

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.

В **четвертой главе** 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 §4.1 теоретические результаты §1.4 существенно используются для обоснования корректности использование логнормальных моделей 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 (см. рису-

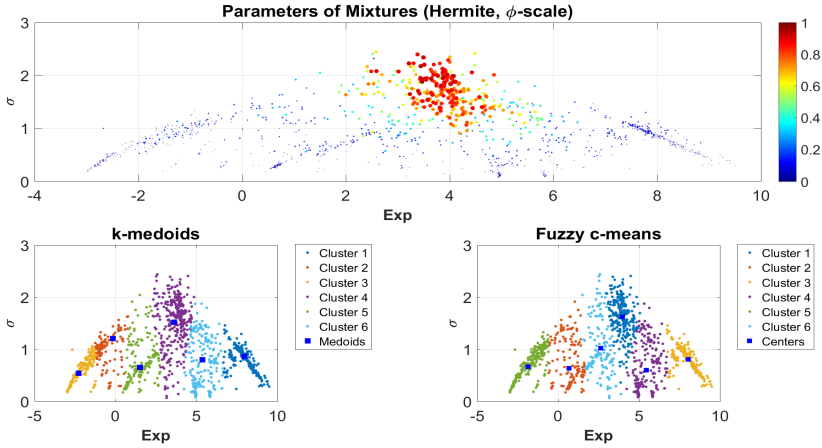

Рис. 1. Кластеризация параметров аппроксимирующих смесей (бутстреп, ϕ-шкала)

нок 1), предложенных моделей может оказаться перспективным инструментом выявления структуры подобных реальных данных с учетом физико-химической интерпретации результатов. Разработанные методы представлены в §4.5 в виде алгоритма 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));
8:          // Имитационное моделирование выборок
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
15:         [ChiParam,ChiPval]=ChiApprox(PhiSize(i), Values(i));
16:     // Функция кластеризации параметров
17:     Clusters←RegolithClustering(Params);
18:     Plot(RegolithSamples, Params, Clusters);          // Визуализация
19:     return ;
```

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.

В **пятой главе** 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, на основании которого удалось решить важные для прикладной области задачи: 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.

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. (см. рисунок 2).

В §5.3 представлены методы прогнозирования значений момент-

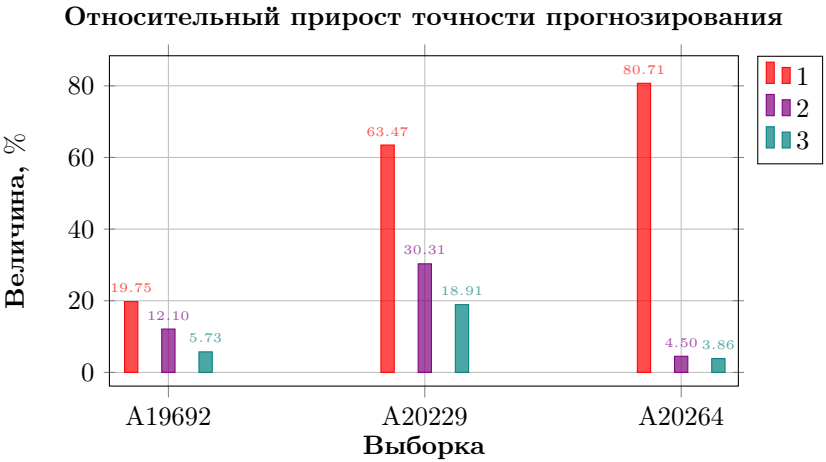

Рис. 2. Прирост точности за счет расширения признакового пространства моментами для приращений относительно конфигураций для данных (1), с выборочными (2) и модельными (3) моментами

ных 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.

Шестая глава посвящена разработке вероятностных моделей на основе теоретических результатов §1.3 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.

В §6.1 на основе k-ичной дискретизации исходных непрерывных данных об объемах осадков решена задача построения вероятност-

ных 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.

В §6.2 решена задача выбора в достаточной степени универсальных с точки зрения эффективности применения в произвольных географических регионах методов машинного обучения для заполнения пропусков в пространственно-временных метеорологических данных. 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. За счет тонкой настройки и дополнительного расширения признакового пространства, могут быть получены и более высокие значения точности, в том числе и иными методами машинного обучения, например, случайными лесами. Созданные инструменты могут быть успешно использованы и для иных видов наблюдений, например данных экологического мониторинга окружающей среды.

В §6.3 предложено и обосновано использование вероятностных моделей на основе классических и обобщенных отрицательных биномиальных и гамма-распределений для длительностей «дождливых» периодов (интервалов времени, в которые осадки регистрировались непрерывно) и соответствующих им объемов осадков. Продемонстрировано высокое соответствие моделей с реальными данными. Разработан эффективный метод функционального оценивания параметров GNB- и GG-распределений.

Рассмотрим GNB-распределение в качестве примера. Пусть построена гистограмма для исходных данных – длительностей «дождливых» периодов. Они могут принимать только целочисленные зна-

чения, 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.

Предложение 6.1. *Для поиска оценок $\widehat{r}$, $\widehat{\gamma}$ и $\widehat{\mu}$ параметров GNB-распределения необходимо решить одну из следующих оптимизационных задач:*

– *в метрике ℓ^1:* $\arg\min\limits_{r,\gamma,\mu} \sum\limits_{k=1}^{N_b} \left| \frac{1}{k!} \int\limits_0^\infty e^{-z} z^k f^{GG}_{r,\gamma,\mu}(z)dz - h_k \right|$;

– *в метрике ℓ^2:* $\arg\min\limits_{r,\gamma,\mu} \sqrt{\sum\limits_{k=1}^{N_b} \left(\frac{1}{k!} \int\limits_0^\infty e^{-z} z^k f^{GG}_{r,\gamma,\mu}(z)dz - h_k \right)^2}$;

– *в метрике ℓ^∞:* $\arg\min\limits_{r,\gamma,\mu} \max\limits_{k=\overline{1,N_b}} \left| \frac{1}{k!} \int\limits_0^\infty e^{-z} z^k f^{GG}_{r,\gamma,\mu}(z)dz - h_k \right|$.

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.

В §6.4 разработаны статистические методы и алгоритмы обнаружения и идентификации экстремальных наблюдений в различных временных рядах на примере осадков и их интенсивностей. 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.

А именно, 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 $T_l^\gamma = V_{i_1}^\gamma + V_{i_2}^\gamma + \ldots + V_{i_l}^\gamma$, $T^\gamma = V_1^\gamma + V_2^\gamma + \ldots + V_M^\gamma$.

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 $SR_{GG} = \frac{(M-l)T_l^\gamma}{l(T^\gamma - T_l^\gamma)}$, 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 метода скользящего окна (см. алгоритм 6.9).

Алгоритм 6.9. Статистическая проверка аномальности

```
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); i ← 1;
6:     for all (Windows) do
7:         m ← Windows_i;
8:         for j=1, M − m + 1 do
9:             ExtrInd_{j:j+m−1} ← SR_GG(Vols_{j:j+m−1}, γ, l=1);
10:        [abs, abs_GG, int, int_GG, rel, rel_GG] ← Identification(ExtrInd);
11:        PlotExtr(abs, abs_GG, int, int_GG, rel, rel_GG); // Визуализация
12:    return γ;
```

Задавая ширину окна равной $m \leqslant M$ и сдвигая каждый раз на один элемент в направлении астрономического времени, с помощью статистики SR_{GG}, полагая в ней $l = 1$, можно последовательно проверить экстремальность каждого объема относительно остальных в описанном выше смысле.

Тогда каждое наблюдение считается: абсолютно экстремальным, если оказывается аномальным во всех m случаях; промежуточным экстремумом, если он признается аномальным более чем в половине случаев (то есть не меньше чем на $\lceil m/2 \rceil$ положениях окна); относительно экстремальным, если оказывается аномальным хотя бы один раз, но не более, чем в половине случаев; стандартным, если они не было распознано как экстремальное ни на одном из положений окна. На рисунке 3 приведено сравнение результатов анализа осадков в Элисте с помощью данной процедуры (отмечены маркерами) и выводов на основе модифицированного метода превышения порогового значения в восходящем и нисходящем вариантах (красная и зеленая линии, соответственно).

С использованием асимптотического распределения экстремальных наблюдений $F_{\lambda,\mu,r}(x)$ (см. теорему 1.2) разработан подход к определению экстремальных суточных объемов как превышающих

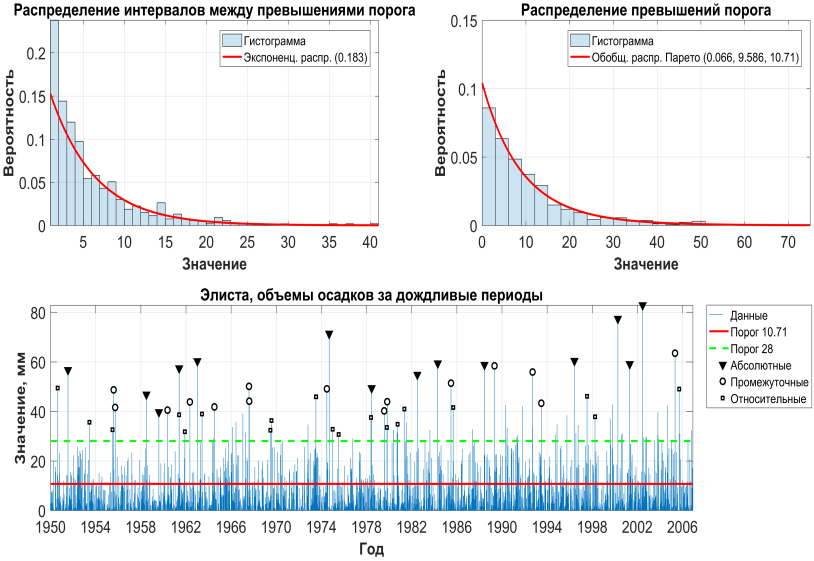

Рис. 3. Сравнение методов определения экстремальности данных

квантили выбранных уровней данного распределения. Предложены несколько процедур для оценивания его параметров. Приведем одну из них, основанную на методе наименьших квадратов.

Предложение 6.8. *При известном значении параметра r МНК-оценки величин λ и μ имеют следующий вид:*

$$\widehat{\mu}_{LS} = \exp\left\{\frac{1}{m-1}\Big(\sum_{j=1}^{m-1}\log\frac{j^{1/r}}{m^{1/r}-j^{1/r}} - \widehat{\lambda}_{LS}\sum_{j=1}^{m-1}\log X^*_{(j)}\Big)\right\},$$

$$\widehat{\lambda}_{LS} = \sum_{j=1}^{m-1}\log X^*_{(j)}\left(\left(\log\frac{j^{1/r}}{m^{1/r}-j^{1/r}}\right)^{m-1} - \sum_{k=1}^{m-1}\log\frac{k^{1/r}}{m^{1/r}-k^{1/r}}\right)\times$$

$$\times\left((m-1)\sum_{j=1}^{m-1}\left(\log X^*_{(j)}\right)^2 - \left(\sum_{j=1}^{m-1}\log X^*_{(j)}\right)^2\right)^{-1}.$$

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.

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 рисунке 4 приведен пример определения статистической структуры процесса теплообмена (верхний график: эволюция компонент) и вклада каждой из структурных составляющих в общее развитие процесса во времени (нижний график: веса компонент).

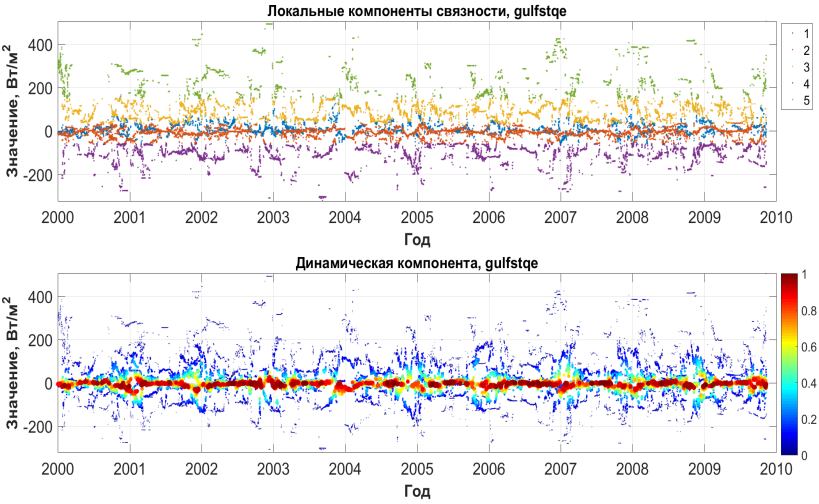

Рис. 4. Оценки распределения сдвига (Гольфстрим, явные потоки)

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.

В **седьмой главе** 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`.

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.

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.

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.

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.

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

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

1. ***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.**

2. *Горшенин А. К.* Программа бутстреп-анализа спектров. Свидетельство о государственной регистрации программ для ЭВМ №2012617918 от 31.08.2012.
3. *Горшенин А. К.* Программа трехмерной визуализации плотностей и параметров распределений. Свидетельство о государственной регистрации программ для ЭВМ №2012660096 от 09.11.2012.
4. ***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.**
5. *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.
6. ***Горшенин А. К.*** **Визуализация результатов для метода скользящего разделения смесей // Информатика и ее применения, 2014. Т. 8. Вып. 4. С. 78–84.**
7. *Горшенин А. К.* Программный модуль анализа спектров с помощью смесей гамма-распределений. Свидетельство о государственной регистрации программ для ЭВМ №2014612083 от 18.02.2014.
8. *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.
9. *Горшенин А. К.* Программный модуль вероятностного анализа спектров на основе логарифмических преобразований. Свидетельство о государственной регистрации программ для ЭВМ №2014661370 от 29.10.2014.
10. *Горшенин А. К.* Средство визуализации результатов для метода скользящего разделения смесей. Свидетельство о государственной регистрации программ для ЭВМ №2014661369 от 29.10.2014.
11. *Горшенин А. К.* Программный модуль «Ядро CPC-метода». Свидетельство о государственной регистрации программ для ЭВМ №2015618673 от 13.08.2015.
12. *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.
13. ***Горшенин А. К.*** **Концепция онлайн-комплекса для стохастического моделирования реальных процессов // Информатика и ее применения, 2016. Т. 10. Вып. 1. С. 72–81.**
14. *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.

15. *Горшенин А. К.* Управляющий модуль для CPC-метода. Свидетельство о государственной регистрации программ для ЭВМ №2016613924 от 11.04.2016.
16. *Горшенин А. К.* Программный модуль динамической визуализации эволюции параметров CPC-метода. Свидетельство о государственной регистрации программ для ЭВМ №2016613925 от 11.04.2016.
17. *Горшенин А. К.* Оптимизированный модуль графического вывода для CPC-метода. Свидетельство о государственной регистрации программ для ЭВМ №2016618859 от 09.08.2016.
18. *Горшенин А. К.* Программный модуль анализа статистических характеристик осадков. Свидетельство о государственной регистрации программ для ЭВМ №2016618864 от 09.08.2016.
19. ***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.**
20. ***Горшенин А. К.*** **Анализ вероятностно-статистических характеристик осадков на основе паттернов // Информатика и ее применения, 2017. Т. 11. Вып. 4. С. 38–46.**
21. *Горшенин А. К.* Программный модуль статистического анализа физических экспериментальных данных. Свидетельство о государственной регистрации программ для ЭВМ №2017617451 от 04.07.2017.
22. *Горшенин А. К.* Программный модуль поиска порогового значения для объемов и интенсивностей осадков. Свидетельство о государственной регистрации программ для ЭВМ № 2017662539 от 10.11.2017.
23. *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.
24. ***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.**
25. ***Горшенин А. К.*** **Развитие сервисов цифровых платформ для преодоления нефинансовых барьеров // Информатика и ее применения, 2018. Т. 12. Вып. 4. С. 109–115.**
26. *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.
27. *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.
28. *Горшенин А. К.* Программа скользящего разделения конечных смесей гамма-распределений с оптимизацией на основе векторных вычисле-

ний. Свидетельство о государственной регистрации программ для ЭВМ № 2018619795 от 10.08.2018.

29. *Горшенин А. К.* Программа классификации экстремальных объемов осадков. Свидетельство о государственной регистрации программ для ЭВМ № 2018619796 от 10.08.2018.
30. *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.
31. *Горшенин А. К.* Программный модуль визуализации точности обучения нейронных сетей. Свидетельство о государственной регистрации программ для ЭВМ № 2018619923 от 14.08.2018.
32. *Горшенин А. К.* Программа статистического анализа распределений объемов осадков за дождливые периоды с графическим пользовательским интерфейсом. Свидетельство о государственной регистрации программ для ЭВМ № 2018661221 от 04.09.2018.
33. *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.
34. *Горшенин А. К.* Программа двухэтапного определения аномальных интенсивностей осадков. Свидетельство о государственной регистрации программ для ЭВМ № 2018665545 от 06.12.2018.
35. *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.
36. *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.
37. *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.
38. *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.
39. *Горшенин А. К.* Программа аппроксимации вероятностных распределений размеров частиц лунного реголита. Свидетельство о государственной регистрации программ для ЭВМ № 2019664472 от 07.11.2019.
40. *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.
41. ***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.

42. ***Горшенин А. К., Данилович Е. С., Хромов Д. Р.*** **Система управления обучением ELIS. Пользовательский интерфейс и функциональные возможности // Системы и средства информатики, 2017. Т. 27. Вып. 2. С. 70–84.**
43. *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.
44. *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.
45. *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.
46. *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.
47. *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.
48. *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.
49. ***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.**
50. *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.
51. ***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.**
52. ***Горшенин А. К., Королев В. Ю., Малахов Д. В., Скворцо-***

ва Н. Н. **Об исследовании плазменной турбулентности на основе анализа спектров // Компьютерные исследования и моделирование, 2012. Т. 4. Вып. 4. – С. 793–802.**

53. ***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.**

54. ***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.**

55. ***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.**

56. *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.

57. *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.

58. ***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.**

59. ***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.**

60. ***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.**

61. *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.

62. ***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.**

63. *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.

64. ***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.**

65. *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.

66. ***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.**

67. ***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.**

68. ***Королев В. Ю., Горшенин А. К.*** **О распределении вероятностей экстремальных осадков // Доклады Академии Наук, 2017. Т. 477. Вып. 5. С. 604–609.**

69. *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.

70. ***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.**

71. *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.

72. *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.

73. *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 кон-

ференция по физике плазмы и управляемому термоядерному синтезу, 9-13 февраля 2015 г., Звенигород. Сборник тезисов докладов – М.: ЗАО НТЦ «ПЛАЗМАИОФАН», 2015. С. 79.

74. *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.

75. ***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.)**

76. ***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.**

77. *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.

78. ***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.)**

79. ***Gorshenin A. K.*** **Investigation of Parameters of Meteorological Models Based on Patterns // CEUR Workshop Proceedings, 2018. Vol. 2177. P. 4–10.**

80. ***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.**

81. ***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.**

82. ***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.**

83. *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.

84. ***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.)

85. ***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.**

86. ***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.**

87. ***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.**

88. ***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.**

89. ***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.**

90. ***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.**

91. ***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.**

92. ***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.**

93. ***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.**

94. ***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.**

95. ***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 29th European Conference on Modelling and Simulation (May 26-29, 2015, Albena (Varna), Bulgaria). – Dudweiler, Germany: Digitaldruck Pirrot GmbH. – P. 631–636.**

96. ***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.)**

97. ***Gorshenin A., Kuzmin V.*** **Online system for the construction of structural models of information flows // Proceedings of the 7th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT). – Piscataway, NJ, USA: IEEE, 2015. P. 216–219.**

98. ***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.)**

99. ***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.**

100. ***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.**

101. ***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.**

102. ***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.**

103. ***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.**

104. *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.

105. ***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.

106. *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.

107. ***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.)**

108. ***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.)**

109. ***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.**

110. ***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.).**

111. ***Korolev V., Gorshenin A., Korchagin A., Zeifman A.*** **Generalized gamma distributions as mixed exponential laws and related limit theorems // Proceedings of 31st European Conference on Modelling and Simulation (May 23-26, 2017, Budapest, Hungary). – Dudweiler, Germany: Digitaldruck Pirrot GmbHP. – P. 642–648.**

112. ***Korolev V. Yu., Sokolov I. A., Gorshenin A. K.*** **Max-compound Cox processes. I // Journal of Mathematical Sciences, 2019. Vol. 237. Вып. 6. P. 789–803.**

113. *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.

114. ***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.**

115. ***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.)**

116. ***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.

117. ***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.**

В списке выделены работы, опубликованные в изданиях из перечня ВАК и/или индексируемые в базах Web of Science Core Collection, Scopus.