ОТЗЫВ

официального оппонента на диссертационную работу Горшенина Андрея Константиновича «Полупараметрические методы анализа неоднородных данных и их применение в задачах математического моделирования», представленную на соискание ученой степени доктора физико-математических наук по специальности 05.13.18 – Математическое моделирование, численные методы и комплексы программ.

Актуальность темы исследования.

Диссертационная работа А. К. Горшенина посвящена научной проблеме, которую можно сформулировать достаточно кратко: развитие статистических методов решения прямых и обратных задач в различных прикладных областях. За этой формулировкой однако стоит целое семейство проблем, которым посвящена данная диссертация: это разработка различных классов смешанных вероятностных моделей и полупараметрических методов анализа статистических данных, исследование их вероятностных свойств, создание эффективных вычислительных алгоритмов оценивания и прогнозирования параметров разрабатываемых моделей, валидация алгоритмов и их применение для решения прикладных задач в самых разных предметных областях.

Актуальность выбранной темы диссертационного исследования определяется в частности тем, что развитие перечисленных выше статистических методов имеет фундаментальное значение для широкого круга как теоретических, так и прикладных задач. В настоящее время серьезное продвижение в научных и технологических задачах невозможно без привлечения и анализа огромных массивов данных, которые как правило неоднородны по структуре и методам их получения. Анализировать такие экстремально большие и разнородные данные традиционными детерминированными методами крайне неэффективно, а зачастую просто невозможно, поэтому естественным подходом здесь являются вероятностные и статистические методы. В практических задачах приходится часто иметь дело и со случайным объемом выборки. Здесь важным инструментом являются предельные теоремы, где в качестве предельных законов для распределений сумм и максимумов либо для неоднородных и нестационарных случайных процессов берутся смеси распределений, которым в данной диссертации уделено особое внимание.

Существует многочисленные применения смешанных вероятностных моделей в таких прикладных задачах, как анализ процессов в турбулентной плазме, при моделировании финансовых инструментов, в технологии обработки изображений, в частности, в медицине, в задачах транспорта газов и аэрозольных частиц в приземном слое атмосферы, и многих других. Следует отметить также, что в реальных прикладных задачах часто можно

наблюдать стохастические системы и зашумленные случайные процессы, вероятностные характеристики которых могут измениться после вмешательства некоторого случайного фактора, изменяющего качественные характеристики системы. Подобные задачи возникают в теории обнаружения, статистическом контроле качества, в технической и медицинской диагностике, при обучении нейронных сетей и т.п. Таким образом, актуальность представленной работы не вызывает сомнений.

Структура диссертации и общая характеристика работы. Диссертация объемом 355 страниц состоит из введения, семи глав, заключения и списка литературы из 458 наименований. Текст содержит также 175 рисунков и 28 таблиц.

Во введении дается краткая характеристика структуры диссертации, обзор литературы, показана актуальность работы, формулируется цель диссертационного исследования, выносимые на защиту результаты, и достаточно подробно, по каждому параграфу, их в диссертации насчитывается 33, дано описание его содержания, что существенно облегчает чтение данной диссертационной работы.

В первой главе представлен класс статистических моделей, которые основаны на выборках, объем которых сам является случайной величиной с обобщенным отрицательным биномиальным законом. Такие модели оказываются весьма эффективными при анализе распределений максимального элемента и суммы всех наблюдений при неограниченном росте объема выборки. Здесь диссертант представил ряд теоретических результатов и конкретных предельных теорем, устанавливающих структуру предельных распределений. В основу разрабатываемых моделей в данной главе и в целом в диссертации взяты смешанные распределения вероятностей. Например, в п.1.2 рассматривается обобщение отрицательного биномиального распределения как смешанного пуассоновского со смешивающим обобщенным гамма-распределением. Здесь получены явные рекуррентные представления для таких распределений. Отдельное рассмотрение посвящено асимптотическому распределению максимальной порядковой статистики в выборке, объем которой является обобщенной отрицательной биномиальной случайной величиной. Получен весьма интересный результат: при некоторых ограничениях на параметры данное распределение является безгранично делимым.

В целом первая глава является чрезвычайно насыщенной в теоретическом отношении. Кроме вышеперечисленных результатов, в ней для отрицательных биномиальных объемов выборок получено аналитическое представление асимптотических распределений порядковых статистик и выборочных квантилей: в этом случае оно является

распределением Стьюдента. Доказан закон больших чисел для сумм с обобщенным отрицательным биномиальным распределением, что обобщает известную теорему Реньи, в котором для слагаемых не предполагается независимость и одинаковая распределенность. Кроме того, доказан новый вариант центральной предельной теоремы для сумм со случайным числом независимых и необязательно одинаково распределенных слагаемых в схеме серий, в которой в качестве предельных распределений возникают произвольные нормальные смеси.

Отметим, что кроме приведенных в диссертации, достаточно разнообразных приложений в конкретных дисциплинах, результаты, связанные со смешанными вероятностными моделями, могут использоваться для эффективного стохастического моделирования таких сложных распределений на компьютере. Действительно, наличие явного представления для смешанного распределения делает возможным применение метода суперпозиции, известного алгоритма моделирования случайных величин.

Во второй главе исследуются аналитические свойства смешанных моделей на основе нормальных и гамма-распределений. Подробно представлен метод скользящего разделения смесей и его использование в качестве базовой процедуры статистического оценивания распределений случайных коэффициентов стохастического дифференциального уравнения Ланжевена. Представлены также две модели возмущений параметров смеси – это модели добавления и расщепления компоненты. Для них приведены результаты относительно асимптотически оптимальных критериев проверки гипотез о числе компонент смеси. Для каждой из моделей выписаны двусторонние оценки, связывающие расстояния Леви между смесями и смешивающими законами. Они обосновывают корректность аппроксимации произвольных сдвиговых нормальных смесей, которые в общем случае не являются идентифицируемыми, конечными аналогами в задаче их статистического разделения.

Алгоритмы анализа данных даны в главе 3. В их основе лежит метод скользящего разделения смесей. Здесь можно перечислить следующие результаты: получены явные линейные и матричные выражения для моментных характеристик конечных нормальных смесей, предложен адаптивный алгоритм выделения полезного сиг нала на фоне шума в смешанных нормальных моделях, получен аналитический вид оценок параметров в линейной и матричной формах, разработан алгоритм последовательной идентификации - определения локальной связности - компонент смесей вероятностных распре делений.
Глава 4 посвящена приложениям разработанных моделей к анализу распределений по размерам пылевых частиц лунного реголита. При этом допускается, что частицы могут как агрегировать, например, в результате спекания, так и дезагрегировать вследствие механического дробления. Приводятся серьезные аргументы и достаточно обширные

статистические данные в пользу того, что распределения частиц по размерам следует искать в виде смеси логнормальных распределений. Здесь диссертант использует результаты своих исследований из раздела 1.4 диссертации.

Другое интересное приложение представлено в главе 5. Здесь усилия автора были направлены на моделирование и анализ процессов, наблюдаемых в экспериментах с плазмой в турбулентном режиме. В разделе 5.1 для решения этой задачи предложено рассматривать, или, точнее сказать, аппроксимировать, спектры флуктуаций турбулентной плазмы с помощью конечных сдвиг-масштабных смесей вероятностных распределений. Для нескольких серий спектров, полученных для разных режимов низкочастотной плазменной турбулентности, продемонстрирована эффективность использования предложенного метода. С его помощью диссертанту удалось решить действительно важные прикладные задачи, а именно: осуществить идентификацию амплитудного спектра с определением формы гармоник в нем и разделением на компоненты, выявить повторяемость стохастических процессов с характерными средними частотами полуширины спектра, а также определить значения таких физических показателей функционирования плазмы, как величина радиального электрического поля и фазовые скорости флуктуаций.

В главе 6 представлен еще один класс прикладных задач, где разработанные автором диссертации статистические методы анализа данных оказываются полезными, - это исследование метеорологических данных по осадкам, в частности, их длительности и интенсивности, и океанологические – это обмен турбулентными потоками тепла между океаном и атмосферой. Здесь особое внимание уделяется выявлению экстремальных наблюдений в рассматриваемых пространственно-временных рядах. Используются как сугубо статистические подходы для оценивания неизвестных параметров, так и набор алгоритмов машинного обучения и нейронных сетей для решения задач заполнения пропусков и прогнозирования. В данной задаче предложено и обосновано использование классических и обобщенных отрицательных биномиальных и гамма-моделей для распределений длительностей дождливых периодов (интервалов времени, в которые осадки регистрировались непрерывно) и соответствующих им объемов осадков. Обнаружено хорошее согласие моделей с реальными данными.

В разделе 6.5 продемонстрировано применение CPC-подхода для анализа статистических закономерностей во временной эволюции тепловых потоков между океаном и атмосферой. Установлен ряд закономерностей во временной изменчивости математического ожидания, дисперсии, коэффициентов асимметрии и эксцесса приращений значений процесса тепловых потоков, а также предложен подход к

определению доли экстремальных наблюдений. 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 между атмосферой и океаном.

Последняя, седьмая глава диссертации, представляет комплексы разработанных программ, с помощью которых был проведен анализ статистических данных и визуализации результатов, описанных в главах 3-6.

В разделе 7.1 представлены графические интерфейсы для запуска CPC-метода и визуального представления его результатов с помощью динамической и диффузионных компонент, моментных характеристик и квантилей, в том числе с помощью анимированных графиков. Эти инструменты созданы с помощью языка программирования пакета MATLAB. Раздел 7.2 посвящен анализу распределений длительностей и объемов осадков, реализующих методы оценивания параметров обобщенных отрицательных биномиальных и гамма-распределений. В разделе 7.3 предложена 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. Наконец, раздел 7.5 посвящен вопросам трансформации отдельных программных решений, в том числе описанных в предшествующих разделах, в научно-образовательные сервисы цифровых платформ.

Научная новизна диссертационной работы заключается в следующем:

1. Впервые разработаны смешанные вероятностные модели для выборок со случайным объемом на основе нового варианта центральной предельной теоремы для сумм со случайным числом независимых и необязательно одинаково распределенных слагаемых величин, с использованием схемы максимума для выборок, объем которых задается обобщенными отрицательными биномиальными распределениями, а также с привлечением закона больших чисел для случайных сумм для моделирования редких событий.
2. Даны доказательства устойчивости в метрике Леви дисперсионно-сдвиговых и конечных сдвиговых смесей нормальных распределений относительно возмущений параметров смешивающего распределения, обосновывающие корректность полупараметрических вычислительных процедур разделения смесей этих семейств

распределений.

3. Разработан комплекс полупараметрических методов анализа неоднородных данных и даны результаты аналитического исследования некоторых их свойств в моделях аддитивного зашумления конечными смесями и округления наблюдений.
4. Предложен полупараметрический подход к статистическому оцениванию распределений случайных коэффициентов стохастических дифференциальных уравнений Ланжевена.
5. Предложена статистическая методология построения моделей сгруппированных скрытых наблюдений при заданных характерных точках их эмпирической функции распределения.
6. Создан комплекс методов и алгоритмов статистической идентификации и классификации экстремальных наблюдений на основе обобщенных отрицательных биномиальных распределений числа наблюдений и обобщенных гамма-моделей для данных.
7. Разработаны и написаны программные продукты для автоматизации обработки массивов неоднородных данных на высокопроизводительных вычислительных комплексах, реализующие разработанные полупараметрические методы. С их помощью решен ряд задач математического моделирования в физике плазмы, роста и формирования частиц, метеорологии, океанологии.

Теоретическая и практическая значимость. Суть работы, использованные методы и полученные диссертантом новые результаты носят как фундаментальный теоретический характер, так и прикладной. Круг исследований очень широк и является ярко выраженным междисциплинарным, и можно сказать еще шире, многодисциплинарным. Разработанные методы анализа данных и алгоритмы моделирования тесно увязаны с теоретическими результатами и полученными в диссертации предельными теоремами. Важно при этом, что эти алгоритмы применены к анализу реальных данных актуальных прикладных задач из самых различных областях науки и технологии.

Обоснованность научных положений. Автором сформулированы математические утверждения о свойствах предложенных моделей и получены строгие математические доказательства целого ряда новых теорем о поведении этих моделей. Большое внимание уделено реализации и валидации предложенных полупараметрических моделей, и создан комплекс программ для анализа данных из различных прикладных задач. Все это убедительно обосновывает научные положения, выдвинутые в диссертации.

Оценка изложения материалов диссертации и автореферата.

Материал, изложенный в диссертации, изложен ясным и логически хорошо организованным языком, он правильно структурирован. Проведенные исследования можно считать завершенными, результаты, полученные в диссертации, являются новыми. Основные результаты по теме диссертации опубликованы в 82 печатных работах, из которых 31 - в журналах, включенных в перечень ВАК, 51 статья – в изданиях, индексированных в базах Web of Science Core Collection и/или Scopus. Получены 39 свидетельств о государственной регистрации компьютерных программ, зарегистрированные в Федеральной службе по интеллектуальной собственности (Роспатент).

Результаты диссертации докладывались на многочисленных российских и международных конференциях, научных семинарах ряда ведущих научных академических институтов и центров, а также были внедрены в следующих организациях: институте общей физики им. А. М. Прохорова Российской академии наук для решения за дач вероятностно-статистического моделирования процессов в экспериментах с турбулентной плазмой в стеллараторе Л-2М, в Институте океанологии им. П. П. Ширшова Российской академии наук для анализа статистических закономерностей в метеорологических и океанологических данных, а также излагаются в ряде тем учебного курса Прикладной многомерный статистический анализ Центра компетенций Националь ной технологической инициативы по технологиям хранения и анализа больших данных на базе МГУ имени М. В. Ломоносова.

По содержанию диссертации и автореферату имеются следующие **замечания**.

1. Общее замечание касается выбора метода статистического анализа для решения прямых и обратных задач, связанных с усвоением экстремально больших объемов данных. В целом, как уже отмечено выше в данном отзыве, полупараметрические методы статистического моделирования являются полезным и эффективным для очень многих физических процессов. Но не всех, в частности и ряда процессов из тех приложений, которые рассмотрены в данной диссертации. Действительно, при изучении процессов, протекающих во времени, например, рассмотренных в диссертации процессов осадкообразования, их длительности и интенсивности, ключевым фактором является корреляция во времени данных процессов. Поэтому естественным в данном случае является привлечение теории случайных процессов. Преимуществом в таком подходе является экономичность описания: например, в случае гауссовских флуктуаций достаточно

восстановить лишь корреляционную функцию. Отметим, что с точки зрения статистических данных серия измерений часто рассматривается как неоднородная последовательность, но является, с точки зрения случайных процессов, однородным случайным процессом в том смысле, что корреляционная функция зависит лишь от разности аргументов.

Ярким примером является и случай моделирования турбулентности. Спектр турбулентных флуктуаций оказывается при таком описании просто преобразованием Фурье от корреляционной функции. А моделирование скорости осуществляется просто на основании рандомизации Фурье представления, где спектр выступает в роли плотности вероятности (ненормированной) распределения фазы и интенсивности пульсаций.

Понятно, что и метод, базирующийся на случайных процессах и полях, имеет свои ограничения и недостатки. Поэтому представляется логичным разрабатывать комбинированные методы, основанные как на корреляционном анализе, так и на статистическом усвоении данных с помощью полупараметрических моделей, предложенных в диссертации.

2. Одним из примеров, когда метод скользящего разделения смесей может столкнуться с серьезными трудностями, является как раз случай, когда изучаемый процесс, протекающий во времени, имеет очень длинные хвосты своей корреляционной функции ("длинные корреляции"), что нередко имеет место в практических задачах. Было бы полезно посмотреть, как справляется с этой проблемой предложенный в диссертации метод.

3. Другая проблема, с которой может столкнуться статистические методы, основанные на полупараметрических моделях, может быть прокомментирована на примере задачи об агрегации--дезагрегации частиц, рассмотренных и в данной диссертации. Дело в том, что в зависимости от скоростей агрегации-дезагрегации могут наступить вырожденные распределения, в частности, процесс может закончиться желированием, то есть образованием одной большой частицы, либо распределением, в котором имеются только первичные мономеры. Эти процессы чрезвычайно трудно поддаются статистическому анализу без предварительного моделирования динамики процессов агрегации-дезагрегации.

4. Следующее замечание касается уравнения Ланжевена со случайными входными функциями дисперсии и дрифта. В диссертации не сказано, при каких предположениях строится метод анализа коэффициентов. Дело в том, что для существования решения уравнения Ланжевена, то есть существования случайного процесса, описываемого уравнением Ланжевена, должны быть выполнены достаточно ограничительные условия на коэффициенты уравнения даже в случае, когда они детерминированы. Тем более сделать

нужные предположения важно в случае, когда эти коэффициенты являются случайными процессами.

5. Не достаточно внимания уделяется в диссертации вопросу о границах применимости разработанных алгоритмов, анализу примеров, где возникли трудности, либо большие ошибки при оценивании параметров.

6. В диссертации встречаются опечатки и стилистические неточности, однако их совсем не много, и они незначительны, так что на фоне очень грамотного и ясного изложения ими легко можно пренебречь.

Оценивая работу в целом, следует отметить, что диссертационная работа является завершенной научно-квалификационной работой, выполнена на высоком научном уровне, а совокупность ее результатов можно квалифицировать как научное достижение в области методов вероятностно-статистического моделирования и анализа данных. Представленные в работе исследования обладают научной новизной и достоверностью, все научные выводы строго обоснованы. Основные положения диссертации достаточно полно освещены в научных публикациях автора, прошли апробацию на многочисленных конференциях и семинарах. Автореферат полностью отражает содержание диссертации. Сделанные мною замечания не влияют на общий весьма высокий уровень оценки диссертационной работы.

Считаю, что диссертация Горшенина А. К. соответствует требованиям, установленным Положением о порядке присуждения ученых степеней, а ее автор заслуживает присуждения ученой степени доктора физико-математических наук по специальности 05.13.18 – Математическое моделирование, численные методы и комплексы программ.

Официальный оппонент
Главный научный сотрудник лаборатории
Стохастических задач ИВМ и МГ СО РАН,
д.ф.-м.н. по специальности 01.01.07 –
вычислительная математика, профессор К.К. Сабельфельд

28.04.2021

Подпись д.ф.-м.н. Сабельфельда К.К. заверяю
Ученый секретарь ИВМ и МГ СО РАН Л.В. Вшивкова

Сабельфельд Карл Карлович

Главный научный сотрудник лаборатории Стохастических задач Федеральное государственное бюджетное учреждение науки Институт вычислительной математики и математической геофизики Сибирского отделения Российской академии наук (ИВМ и МГ СО РАН),

Адрес: 630090, г. Новосибирск, проспект Академика Лаврентьева, 6.

Телефон: (383) 330-77-21

E-mail: karl@osmf.sscc.ru