На правах рукописи

Корчагин Александр Юрьевич

Прогнозирование стохастических процессов с помощью сеточного метода разделения дисперсионно-сдвиговых смесей нормальных законов

Специальность 01.01.05 —
«Теория вероятностей и математическая статистика»

Автореферат

диссертации на соискание учёной степени
кандидата физико-математических наук

Москва — 2015

Работа выполнена на кафедре математической статистики факультета вычислительной математики и кибернетики Московского государственного университета имени М. В. Ломоносова

Научный руководитель:	доктор физико-математических наук, профессор **Королев Виктор Юрьевич**
Официальные оппоненты:	**Беляев Константин Павлович,** доктор физико-математических наук, Институт океанологии РАН им. П. П. Ширшова, ведущий научный сотрудник
	Артюхов Сергей Владимирович, кандидат физико-математических наук, ЗАО «Сбербанк-Технологии», ведущий аналитик
Ведущая организация:	Казанский (Приволжский) федеральный университет

Защита состоится 26 июня 2015 г. в 11:00 на заседании диссертационного совета Д 501.001.44 при Московском государственном университете имени М. В. Ломоносова по адресу: 119991, ГСП-1, Москва, Ленинские горы, МГУ, 2-й учебный корпус, факультет ВМК, аудитория 685. Желающие присутствовать на заседании диссертационного совета должны сообщить об этом за 2 дня до указанной даты по тел. (495) 939-30-10 (для оформления заявки на пропуск).

С диссертацией можно ознакомиться в библиотеке Фундаментальной библиотеке МГУ им. М. В. Ломоносова. С текстом автореферата можно ознакомиться на официальном сайте ВМК МГУ http://cs.msu.ru в разделе "Наука" - "Работа диссертационных советов" - "Д 501.001.44".

Автореферат разослан 2015 года.

Ученый секретарь
диссертационного совета
к.ф.-м.н., доцент О. В. Шестаков

Общая характеристика работы

Актуальность темы. Дисперсионно-сдвиговые смеси нормальных законов активно используются как математические модели статистических закономерностей, наблюдаемых во многих практических задачах. Изначально они вводились в семидесятых-восьмидесятых годах прошлого столетия в работах О.-Е. Барндорфф-Нильсеном и его коллегами как довольно естественные обобщения нормального закона в терминах случайно остановленных процессов броуновского движения с нетривиальным сносом. Наиболее популярными дисперсионно-сдвиговыми смесями нормальных законов являются обобщенные гиперболические распределения, определяемые пятью параметрами.

Эти смеси интересны тем, что хотя формально в них смешивание происходит по обоим параметрам нормальных законов – сдвигу и дисперсии, – но эти параметры связаны жесткой пропорциональной зависимостью, так что фактически смешивающее распределение одномерно. В частности, для обобщенных гиперболических законов смешивающим является обобщенное обратное гауссовское распределение.

В работах [1, 2] был предложен еще один класс специальных дисперсионно-сдвиговых смесей нормальных законов – класс обобщенных дисперсионных гамма-распределений (generalized variance-gamma distributions), который в отличие от обобщенных гиперболических законов содержит распределения, хвосты которых убывают экспоненциально-степенным (вейбулловским) образом. В некоторых случаях такие распределения оказываются более адекватными моделями реально наблюдаемых закономерностей, нежели обобщенные гиперболические законы [3].

Наличие большого числа настраиваемых параметров порождает уверенность в том, что обобщенные гиперболические или обобщенные дисперсионные гамма-распределения являются практически универсальными моделями.

Однако в прикладной теории вероятностей хорошо известен принцип, восходящий, по-видимому, к работе [4], согласно которому та или иная

[1] *Королев В. Ю., Соколов И. А.* Скошенные распределения Стьюдента, дисперсионные гамма-распределения и их обобщения как асимптотические аппроксимации // Информатика и ее применения, 2012. Т. 6. Вып. 1. С. 2–10.

[2] *Закс Л. М., Королев В. Ю.* Обобщенные дисперсионные гамма-распределения как предельные для случайных сумм // Информатика и ее применения, 2013. Т. 7. Вып. 1. С. 105–115.

[3] *Qian Chen, Gerlach R. H.* The two-sided Weibull distribution and forecasting financial tail risk // International Journal of Forecasting, 2013. Vol. 29. No. 4. P. 527–540.

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

модель может считаться в достаточной мере обоснованной только тогда, когда она является *асимптотической аппроксимацией*, то есть когда существует довольно простая предельная схема, например, схема максимума или схема суммирования, и соответствующая предельная теорема, в которой рассматриваемая модель выступает в качестве предельного распределения. В книге [5] прослежена глубокая связь этого принципа с универсальным принципом неубывания энтропии в замкнутых системах. Как известно, нормальное распределение обладает максимальной (дифференциальной) энтропией среди всех распределений, носителем которых является вся числовая прямая, и имеющих конечный второй момент. Если бы моделируемая сложная система была информационно изолирована от окружающей среды, то в соответствии с принципом неубывания энтропии, который в теории вероятностей проявляется в виде предельных теорем [5], наблюдаемые статистические распределения ее характеристик были бы неотличимы от нормального. Но поскольку любая математическая модель по своему определению не может учесть все факторы, влияющие на состояние или эволюцию моделируемой системы, то параметры этого нормального закона изменяются в зависимости от состояния среды, внешней по отношению к моделируемой системе. Другими словами, эти параметры являются случайными и изменяются под влиянием информационных потоков между системой и внешней средой. Таким образом, во многих ситуациях разумные модели статистических закономерностей изменения параметров сложных систем должны иметь вид смесей нормальных законов, частным случаем которых являются дисперсионно-сдвиговые смеси нормальных законов.

В классических задачах математической статистики объем выборки, доступной исследователю, традиционно считается детерминированным и в асимптотических постановках играет роль неограниченно возрастающего *известного* параметра. В то же время, на практике часто возникают ситуации, когда размер выборки не является заранее определенным и может рассматриваться как случайный. Эти ситуации, как правило, связаны с тем, что статистические данные накапливаются в течение фиксированного времени. Это имеет место, в частности, в страховании, когда в течение разных отчетных периодов одинаковой длины (скажем, месяцев) происходит разное число страховых событий – страховых выплат и/или заключений страховых контрактов; в медицине, когда число пациентов с тем или иным заболеванием варьируется от года к году; в технике, когда при испытании на надежность (скажем, при определении наработки на отказ) разных партий

[5] *Gnedenko B. V., Korolev V. Yu.* Random Summation: Limit Theorems and Applications. – Boca Raton: CRC Press, 1996.

приборов, число отказавших приборов в разных партиях будет разным; в информатике при разработке методов оценки «своевременности» завершения программ, включая методы решения задач предсказания времени безотказного функционирования или времени выполнения прикладных программ в случайных вычислительных средах. В таких ситуациях заранее не известное число наблюдений, которые будут доступны исследователю, разумно считать случайной величиной. Другими словами, в таких ситуациях объем выборки не является известным параметром, а сам становится *наблюдением*, то есть статистикой. В силу указанных обстоятельств вполне естественным становится изучение асимптотического поведения распределений статистик достаточно общего вида, основанных на выборках случайного объема, а также поиск удобной и адекватной модели, описывающей статистические закономерности поведения таких статистик.

На естественность такого подхода, в частности, обратил внимание Б. В. Гнеденко в работе [6], в которой рассматривались асимптотические свойства распределений выборочных квантилей, построенных по выборкам случайного объема, и было продемонстрировано, что при замене неслучайного объема выборки случайной величиной асимптотические свойства статистик могут радикально измениться. К примеру, вместо ожидаемого в соответствии с классической теорией нормального закона, могут возникать распределения с произвольно тяжелыми хвостами. В частности, если объем выборки является геометрически распределенной случайной величиной, то вместо ожидаемого в соответствии с классической теорией нормального закона, в качестве асимптотического распределения выборочной медианы возникает распределение Стьюдента с двумя степенями свободы, хвосты которого столь тяжелы, что у него отсутствуют моменты порядков, больших второго.

Литература о статистиках, построенных по выборкам случайного объема, обширна. Их свойства изучены достаточно полно. Однако условия сходимости распределений таких статистик к дисперсионно-сдвиговым смесям нормальных законов были найдены лишь недавно [7, 2]. В работе [8] приведены критерии сходимости распределений статистик, построенных по выборкам случайного объема, к многомерным обобщенным гиперболическим

[6] *Гнеденко Б. В.* Об оценке неизвестных параметров распределения при случайном числе независимых наблюдений // Труды Тбилисского Математического института, 1989. Т. 92. С. 146–150.

[7] *Королев В. Ю.* Обобщенные гиперболические распределения как предельные для случайных сумм // Теория вероятностей и ее применения, 2013. Т. 58. Вып. 1. С. 117-–132.

[8] *Korolev V. Yu., Zeifman A. I.* On convergence of the distributions of statistics constructed from samples with random sizes to normal variance-mean mixtures // Journal of Statist. Planning and Inference, to appear. Available at: arXiv:1410.1518v1, 2014.

законам. Как показано в этой статье, указанная сходимость имеет место тогда и только тогда, когда случайная интенсивность потока информативных событий, в результате которых накапливаются наблюдения, формирующие выборку, имеет асимптотически обобщенное обратное гауссовское распределение. В некоторых случаях обобщенные гамма-распределения с экспоненциально-степенными хвостами лучше описывают статистические закономерности поведения наблюдаемых величин. Вместе с тем, как показано в работе [9],асимптотическое поведение хвостов смесей нормальных законов в определенном смысле совпадает с аналогичным поведением хвостов смешивающих законов. Следовательно, аналогичная асимптотика должна быть присуща хвостам распределений интенсивностей потоков информативных событий. Действительно, как оказалось, обобщенные гамма-распределения заметно лучше согласуются с эмпирическими распределениями числа событий в книгах заявок в высокочастотных системах электронной торговли на финансовых рынках (P-значения при проверке согласия с помощью критерия хи-квадрат примерно равны 0.8), нежели обобщенные обратные гауссовские распределения (аналогичные P-значения примерно равны 0.2). Поэтому поиск критериев сходимости к многомерным дисперсионным гамма-распределениям представляет собой весьма перспективную задачу, решение которой позволяет получить дополнительную информацию о структуре моделируемой системы или моделируемого процесса.

Неотъемлемой составной частью задачи *практического* математического моделирования стохастических процессов или явлений является задача определения параметров используемых математических моделей. Если для описания стохастических процессов или явлений используются смешанные модели, в частности, упоминавшиеся выше дисперсионно-сдвиговые смеси нормальных законов, то задача определения параметров сводится к статистическому разделению смесей – статистическому оцениванию параметров смесей вероятностных распределений. Такая задача, в частности, является одной из самых часто встречающихся практических задач моделирования и исследования волатильности. Она в деталях разобрана, например, в книге [10], где можно найти дальнейшие ссылки на многие работы, посвященные данной тематике.

Для решения задачи разделения смесей вероятностных распределений традиционно используются итерационные процедуры типа EM-алгоритма.

[9] *Антонов С. Н., Кокшаров С. Н.* Об асимптотическом поведении хвостов масштабных смесей нормальных распределений // Статистические методы оценивания и проверки гипотез. – Пермь: Изд-во Пермского университета, 2006. С. 90–105.

[10] *Королев В. Ю.* Вероятностно-статистические методы декомпозиции волатильности хаотических процессов. – М.: Изд-во Московского университета, 2011.

К сожалению, классический EM-алгоритм обладает рядом серьезных недостатков при его применении к смесям нормальных законов. В частности, он демонстрирует крайнюю неустойчивость по отношению к исходным данным и начальным приближениям. Для преодоления этих недостатков предложено много модификаций EM-алгоритма, см., например, [10]. Вместе с тем, в указанной книге предложен и исследован принципиально новый «сеточный» метод приближенного решения задачи разделения смесей. В работе [11] подробно исследованы вопросы сходимости сеточных методов разделения смесей.

В соответствии с подходом к статистическому анализу хаотических стохастических процессов, в частности к решению задачи декомпозиции волатильности таких процессов, развитом в книге [10], в общем случае на практике приходится решать задачу разделения конечных смесей нормальных законов с произвольно большим числом неизвестных параметров (параметров компонент и их весов). И хотя в большинстве приложений возникают смеси не более чем с пятью-семью компонентами, даже при использовании таких смесей, скажем, в задачах анализа и прогнозирования финансовых рисков приходится моделировать траекторию движения точки в пространствах, размерность которых соответственно лежит в пределах от 14 (для пятикомпонентных смесей) до 20 (для семикомпонентных смесей), что существенно увеличивает вычислительные и временные ресурсы, необходимые для практического решения указанных задач. Поскольку во многих ситуациях, например, при прогнозировании на основе высокочастотных данных, эти задачи необходимо решать в режиме, близком к реальному времени, для создания эффективных методов статистического анализа на основе смешанных моделей на первый план выходит проблема снижения размерности решаемой задачи, т. е. параметрического пространства.

Одним из возможных подходов к снижению размерности является априорное сужение классов допустимых смесей. К примеру, при решении многих задач, связанных с анализом процессов атмосферной или плазменной турбулентности, а также процессов, описывающих эволюцию различных финансовых индексов, высочайшую адекватность продемонстрировали модели, основанные на дисперсионно-сдвиговых смесях нормальных законов. Как уже отмечалось, класс таких смесей очень обширен и, в частности, включает в себя обобщенные гиперболические распределения, и обобщенные дисперсионные гамма-распределения. В указанных семействах смесей число

[11] *Назаров А. Л.* Приближенные методы разделения смесей вероятностных распределений. Диссертация на соискание ученой степени кандидата физ.-матем. наук. – М.: Московский государственный университет им. М. В. Ломоносова, 2013.

неизвестных параметров равно пяти или шести (если учитывать неслучайный сдвиг). Как показано в первой главе диссертации, у подобных моделей имеются довольно серьезные теоретические обоснования, а именно, указанные модели являются асимптотическими аппроксимациями в простой предельной схеме случайного суммирования и потому могут успешно применяться для анализа процессов типа остановленных случайных блужданий. Эти выводы подтверждены статистическим анализом высокочастотных финансовых данных, в результате которого выявлен синхронизированный характер изменения интенсивностей потоков заявок в системах электронных торгов, что естественно приводит к синхронизированному поведению параметров сдвига и диффузии в соответствующих моделях вида смесей нормальных законов [12].

Для решения задачи оценивания параметров обобщенных гиперболических распределений традиционно используется метод, предложенный в статье [13] и по сути являющийся классическим EM-алгоритмом, приспособленным к конкретной задаче, и, соответственно, наследующий присущие EM-алгоритмам недостатки. В связи с этим возникает важная задача адаптации упоминавшихся выше сеточных методов для решения задачи статистического разделения произвольных дисперсионно-сдвиговых смесей нормальных законов, решению которой посвящена глава 2 данной диссертации, где на примере обобщенных гиперболических и обобщенных дисперсионных гамма-распределений описывается и изучается принципиально новый метод разделения дисперсионно-сдвиговых смесей нормальных законов.

Эффективно работающие алгоритмы статистического разделения смесей могут быть использованы при решении задачи прогнозирования рисков. А именно, традиционная задача прогнозирования стохастических процессов сводится к построению точечного прогноза возможной его траектории. Вместе с тем во многих случаях исследователь в не меньшей степени заинтересован в решении задач прогнозирования *распределения* значения случайного процесса, что позволяет решать, в частности, задачи прогнозирования финансовых рисков как вероятностей превышения критических порогов рассматриваемым индексом.

Помимо непосредственного исследования распределений, любая финансовая организация заинтересована в получении достаточно достоверных

[12] *Королев В. Ю., Черток А. В., Корчагин А. Ю., Горшенин А. К.* Вероятностно-статистическое моделирование информационных потоков в сложных финансовых системах на основе высокочастотных данных // Информатика и ее применения, 2013 г., том 7, Вып. 1, с. 12–21.

[13] *Protassov R. S.* EM-based maximum likelihood parameter estimation for a multivariate generalized hyperbolic distribution with fixed λ // Statistics, Computing, 2004. Vol. 14. P. 67–77.

прогнозов на основе наблюдаемых данных. Прогнозирование содержит в себе большой спекулятивный фактор, но некоторые жесткие требования к любому осмысленному методу прогнозирования известны заранее: метод должен работать достаточно быстро, чтобы прогноз оставлял время для принятия решения, а также должен показывать хорошие результаты на случайно выбранных исторических данных.

В диссертации продемонстрировано, что указанная задача прогнозирования рисков с помощью смешанных моделей может быть успешно сведена к решению задачи прогнозирования траектории точки, описывающей параметры обобщенного гиперболического или обобщенного дисперсионного гамма-распределения в соответствующем четырех- или пятимерном пространстве.

Целью данной работы является всестороннее изучение специальных вероятностных моделей стохастических процессов и явлений, имеющих вид дисперсионно-сдвиговых смесей нормальных законов, в частности, обобщенных гиперболических и обобщенных дисперсионных гамма-распределений.

Для достижения поставленной цели необходимо было решить следующие **задачи**:

1. доказать критерии сходимости распределений статистик, построенных по выборкам случайного объема, в частности, сумм случайного числа случайных величин, к многомерным дисперсионно-сдвиговым смесям нормальных законов, в частности, к обобщенных гиперболическим и обобщенным дисперсионным гамма-распределениям;

2. разработать эффективный комбинированный методд статистического разделения дисперсионно-сдвиговых смесей нормальных законов, в частности, обобщенных гиперболических и обобщенных дисперсионных гамма-распределений, и изученить его свойства;

3. изучить и продемонстрировать возможности предложенных моделей и методов на примере решения практических задач, в частности, задачи статистической локализации невосполнимых областей головного мозга человека по магнитоэнцефалограммам и миограммам; задачи прогнозирования финансовых рисков; задачи анализа текстовой информации для анализа и предотвращения утечек данных.

Основные положения, выносимые на защиту:

1. Предложено теоретическое обоснование адекватности моделей, имеющих вид дисперсионно-сдвиговых смесей нормальных законов: доказаны предельные теоремы о сходимости распределений многомерных

статистик, построенных по выборкам случайного объема, к многомерным дисперсионно-сдвиговым смесям нормальных законов. В том числе доказаны критерии сходимости распределений случайных сумм независимых многомерных случайных величин к многомерным дисперсионно-сдвиговым смесям нормальных законов, в частности, к многомерным обобщенным гиперболическим и обобщенным дисперсионным гамма-распределениям, а также функциональная предельная теорема о сходимости обобщенных процессов Кокса к процессам Леви с одномерными обобщенными дисперсионными гамма-распределениями.

2. Разработан, реализован, а также теоретически и экспериментально исследован комбинированный сеточный метод разделения дисперсионно-сдвиговых смесей нормальных законов. Этот метод успешно применен к решению задачи отыскания опорных точек для локализации невосполнимых областей головного мозга человека с помощью выявления двигательной активности на основе магнитоэнцефалограмм и миограмм.

3. Разработан, реализован и исследован метод прогнозирования финансовых рисков с помощью приближенного решения задачи статистического разделения дисперсионно-сдвиговых смесей нормальных законов. Проведено тестирование метода на различных финансовых данных. Этот метод также применен в задаче анализа текстовой информации для предотвращения утечек данных.

Научная новизна:

1. Впервые доказаны предельные теоремы о сходимости распределений многомерных случайных последовательностей с независимыми случайными индексами к дисперсионно-сдвиговым смесям нормальных законов. В частности, впервые доказаны критерии сходимости распределений сумм случайного числа независимых многомерных случайных величин, а также многомерных статистик, построенных по выборкам случайного объема, к многомерным обобщенным гиперболическим и многомерным обобщенным дисперсионным гамма-распределениям.

2. Впервые разработан комбинированный метод разделения дисперсионно-сдвиговых смесей нормальных законов и получены теоретические оценки для выбора оптимальных параметров метода. Этот метод впервые применен к решению задачи отыскания опорных точек для локализации невосполнимых областей головного мозга человека с помощью выявления двигательной активности на основе магнитоэнцефалограмм и миограмм, где он продемонстрировал высокую эффективность.

3. Впервые разработан и экспериментально исследован метод прогнозирования финансовых рисков на основе разделения дисперсионно-сдвиговых смесей нормальных законов.

Практическая значимость диссертационной работы состоит в том, что разработанные методы показывают более высокую точность по сравнению с аналогами в ряде практических задач, в частности применительно к задаче выявления двигательной активности на основе магнитоэнцефалограмм и миограмм, а также применительно к задаче анализа текстовой информации для предотвращения утечек данных.

Достоверность обеспечивается корректными доказательствами теорем и подробными описаниями алгоритмов и экспериментов, допускающими воспроизводимость.

Апробация работы. Основные результаты диссертации докладывались на ежегодной научной конференции «Тихоновские чтения» (Москва, 27–31 октября 2014 г.), на XXXII Международном семинаре по проблемам устойчивости стохастических моделей (Тронхейм, Норвегия, июнь 2014 г.), на научно-исследовательском семинаре «Теория риска и смежные вопросы» на факультете ВМК МГУ, на первой научно-практической конференции молодых ученых «Задачи современной информатики» (Москва, ИПИ РАН, декабрь 2014 г.).

Публикации. Основные результаты по теме диссертации изложены в 10 печатных изданиях [1] - [10], в том числе высокорейтинговых журналах; 6 работ изданы в журналах, рекомендованных ВАК, 3 — в тезисах докладов.

Личный вклад автора состоит в получении основных теоретических результатов, программной реализации, экспериментальной апробации. В работе [2] автору принадлежит доказательство теорем о сходимости статистик, построенных по выборкам случайного объема, к многомерным дисперсионно-сдвиговым смесям нормальных законов. В частности, доказаны критерии сходимости к многомерным обобщенным гиперболическим и многомерным обобщенным дисперсионным гамма-распределениям. В работах [8 – 10] автором проведены конкретные расчеты с использованием предложенного им метода разделения смесей. В остальных работах вклад соавторов заключается в следующем. В работе [3] В. Ю. Королев предложил подход к оцениванию границ носителя смешивающего распределения. В работе [4] В. Ю. Королев и О. А. Морева предложили альтернативные подходы к многопроходному выбору сеток. В работе [6] В. Ю. Королев, И. А. Соколов и А. В. Черток исследовали свойства обобщенных гиперболических моделей, а А. Ю. Корчагин исследовал свойства обобщенных дисперсионных

гамма-моделей. В работе [7] Ярошенко И. И. предложил простой метод отыскания параметров распределения из заданного класса.

Объем и структура работы. Диссертация состоит из введения, трех глав, заключения и двух приложений. Полный объем диссертации составляет 113 страниц с 33 рисунками и 20 таблицами.

Содержание работы

Во **введении** обосновывается актуальность исследований, проводимых в рамках данной диссертационной работы, приводится обзор научной литературы по изучаемой проблеме, формулируется цель, ставятся задачи работы.

В первых разделах **Первой главы** содержатся общие сведения о свойствах дисперсионно-сдвиговых смесей нормальных законов.

Дисперсионно-сдвиговой смесью нормальных законов называется функция распределения

$$F(x) = \int_0^\infty \Phi\Big(\frac{x-\beta-\alpha z}{\sigma\sqrt{z}}\Big)\, dG(z), \quad x \in \mathbb{R}. \tag{1.1}$$

$G(z)$ – функция распределения, такая, что $G(0) = 0$.

В соотношении (1.1) смешивание происходит одновременно и по параметру сдвига, и по параметру масштаба, но так как эти параметры в (1.1) связаны жесткой зависимостью, при которой параметры положения (*сдвига*) смешиваемых нормальных законов пропорциональны их *дисперсиям* то фактически смесь (1.1) является однопараметрической. Именно поэтому смеси вида (1.1) называются *дисперсионно-сдвиговыми*.

Аналогично, в многомерном случае распределение $\mathbb{R}^m$-значного случайного вектора $\boldsymbol{Z}$ является многомерной дисперсионно-сдвиговой смесью нормальных законов, если

$$\boldsymbol{Z} \stackrel{d}{=} \boldsymbol{b} + U\,\boldsymbol{a} + \sqrt{U}\,A\,\boldsymbol{Y}, \tag{1.2}$$

где $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^m$, A – вещественная $(m \times m)$-матрица такая, что $\Sigma \equiv AA^\top$ положительно определена, $\boldsymbol{Y}$ – случайный вектор со стандартным m-мерным нормальным распределением $\mathcal{N}_{\boldsymbol{0},\boldsymbol{I}}$, а U – неотрицательная случайная величина, независимая от $\boldsymbol{Y}$. Этот факт иногда записывается в виде $F = \mathcal{N}_{\boldsymbol{b}+z\boldsymbol{a},\, z\Sigma} \circ G$.

В первой главе также описаны два конкретных параметрических семейства дисперсионно-сдвиговых смесей нормальных законов: обобщенные

гиперболические распределения и обобщенные дисперсионные гамма-распределения.

Плотность *обобщенного обратного гауссовского распределения* обозначим $p_{GIG}(x;\nu,\mu,\lambda)$,

$$p_{GIG}(x;\nu,\mu,\lambda)=\frac{\lambda^{\nu/2}}{2\mu^{\nu/2}K_\nu\big(\sqrt{\mu\lambda}\big)}\cdot x^{\nu-1}\cdot\exp\Big\{-\frac{1}{2}\Big(\frac{\mu}{x}+\lambda x\Big)\Big\},\quad x>0. \tag{1.4}$$

Здесь $\nu\in\mathbb{R}$,

$$\begin{aligned}&\mu>0,\ \ \lambda\geqslant 0,\ \ \text{если } \nu<0,\\ &\mu>0,\ \ \lambda>0,\ \ \text{если } \nu=0,\\ &\mu\geqslant 0,\ \ \lambda>0,\ \ \text{если } \nu>0,\end{aligned}$$

$K_\nu(z)$ – модифицированная бесселева функция третьего рода порядка ν,

$$K_\nu(z)=\frac{1}{2}\int_0^\infty y^{\nu-1}\exp\Big\{-\frac{z}{2}\Big(y+\frac{1}{y}\Big)\Big\}dy,\qquad z\in\mathbb{C},\ \operatorname{Re}z>0.$$

В 1977–78 гг. О.-Э. Барндорфф-Нильсен ввел класс **обобщенных гиперболических распределений** как класс специальных дисперсионно-сдвиговых смесей нормальных законов. Пусть $\alpha\in\mathbb{R}$, $\beta\in\mathbb{R}$. Если функцию обобщенного гиперболического распределения с параметрами α, β, ν, μ, λ обозначить $P_{GH}(x;\alpha,\beta,\nu,\mu,\lambda)$, то по определению

$$P_{GH}(x;\alpha,\beta,\nu,\mu,\lambda)=\int_0^\infty\Phi\Big(\frac{x-\beta-\alpha z}{\sqrt{z}}\Big)\,p_{GIG}(z;\nu,\mu,\lambda)dz,\quad x\in\mathbb{R}. \tag{1.5}$$

Для другого важного рассматриваемого класса нам понадобиться определение обобщенного гамма-распределения – это распредление, определяемое плотностью вероятностей вида

$$p_{GG}(x;\nu,\kappa,\delta)=\begin{cases}\dfrac{|\nu|}{\delta^{k|\nu|}\Gamma(\kappa)}x^{\kappa\nu-1}\exp\Big\{-\dfrac{x^\nu}{\delta^{|\nu|}}\Big\}, & x\geqslant 0,\\ 0, & x<0,\end{cases} \tag{1.6}$$

с параметрами $\nu\in\mathbb{R}$, κ, $\delta\in\mathbb{R}^+$, отвечающими соответственно за *степень, форму и масштаб*, где

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

обозначает эйлерову гамма-функцию.

В работе [1] введено пятипараметрическое семейство распределений

$$P_{GVG}(x;a,\sigma,\nu,\kappa,\delta)=\int_0^\infty\Phi\left(\frac{x-au}{\sigma\sqrt{u}}\right)p_{GG}(u;\nu,\kappa,\delta)du, \tag{1.7}$$

где $p_{GG}(u;\nu,\kappa,\delta)$ – плотность GG-распределения (1.6). В статье распределения вида (1.7) названы **обобщенными дисперсионными гамма-распределениями**.

Указанные выше два семейства сопоставляются, демонстрируется что в некоторых ситуациях использование обобщенных дисперсионных гамма-распределений в качестве моделей дает лучшие результаты.

В разделе 1.4 первой главы приводятся известные и доказываются новые предельные теоремы, объясняющие характер смешивающего распределения в конкретных ситуациях и дающие дополнительное обоснование высокой адекватности моделей типа дисперсионно-сдвиговых смесей в рамках асимптотического подхода.

Рассмотрим последовательность серий $\{\boldsymbol{X}_{n,j}\}_{j\geqslant 1}$, $n\in\mathbb{N}$, независимых, но не обязательно одинаково распределенных в каждой серии случайных величин. Для $n,k\in\mathbb{N}$ Положим

$$\boldsymbol{S}_{n,k}=\boldsymbol{X}_{n,1}+...+\boldsymbol{X}_{n,k}. \tag{1.34}$$

Для $n,k\in\mathbb{N}$ пусть $\boldsymbol{a}_{n,k}=(a_{n,k}^{(1)},\ldots,a_{n,k}^{(m)})^{\top}\in\mathbb{R}^m$ – неслучайные векторы и $b_{n,k}\in\mathbb{R}$ – положительные числа. Назначение векторов $\boldsymbol{a}_{n,k}$ и чисел $b_{n,k}$ – обеспечить слабую относительную компактность семейства случайных векторов $\big\{\boldsymbol{Y}_{n,k}\equiv b_{n,k}^{-1}\big(\boldsymbol{S}_{n,k}-\boldsymbol{a}_{n,k}\big)\big\}_{n,k\in\mathbb{N}}$, когда это требуется.

Рассмотрим семейство $\{N_n\}_{n\in\mathbb{N}}$ неотрицательных случайных величин таких, что при каждом $n,k\in\mathbb{N}$ случайные величины N_n независимы от случайных векторов $\boldsymbol{S}_{n,k}$. Особо заметим, что «построчная» независимость случайных векторов $\{\boldsymbol{S}_{n,k}\}_{k\geqslant 1}$ не требуется. Пусть $\boldsymbol{c}_n=(c_n^{(1)},\ldots,c_n^{(m)})^{\top}\in\mathbb{R}^m$ – неслучайные векторы и d_n – положительные числа, $n\in\mathbb{N}$. Наша цель – изучить асимптотическое поведение случайных векторов $\boldsymbol{Z}_n\equiv d_n^{-1}\big(\boldsymbol{S}_{n,N_n}-\boldsymbol{c}_n\big)$ при $n\to\infty$, уделив особое внимание ситуации, в которой предельные распределения для $\boldsymbol{Z}_n$ имеют вид дисперсионно-сдвиговых смесей нормальных законов.

Характеристическую функцию случайного вектора $\boldsymbol{Y}_{n,k}$ обозначим $h_{n,k}(\boldsymbol{t})$, $\boldsymbol{t}\in\mathbb{R}^m$. Пусть $\boldsymbol{Y}$ – $\mathbb{R}^m$-мерный случайный вектор, характеристическая функция которого будет обозначаться $h(\boldsymbol{t})$, $\boldsymbol{t}\in\mathbb{R}^m$. Введем случайные величины $U_n=d_n^{-1}b_{n,N_n}$. Пусть $\boldsymbol{V}_n=(V_n^{(1)},\ldots,V_n^{(m)})^{\top}$, где $V_n^{(k)}=d_n^{-1}(a_{n,N_n}^{(k)}-c_n^{(k)})$ – k-я компонента случайного вектора $d_n^{-1}(\boldsymbol{a}_{n,N_n}-\boldsymbol{c}_n)$. В дальнейшем символ $\boldsymbol{W}_n$ будет обозначать $(m+1)$-мерный случайный вектор $\boldsymbol{W}_n=(U_n,\,\boldsymbol{V}_n^{\top})^{\top}=(U_n,\,V_n^{(1)},\ldots,V_n^{(m)})^{\top}$.

Пусть $\Lambda(\boldsymbol{X}_1,\,\boldsymbol{X}_2)$ – любая вероятностная метрика, метризующая слабую сходимость в пространстве $(m+1)$-мерных случайных векторов. Примером такой метрики является расстояние Леви–Прохорова.

Исследуем слабую сходимость нормированных случайных векторов с независимыми случайными индексами

$$\boldsymbol{Z}_n \Longrightarrow \boldsymbol{Z} \quad (n \to \infty) \tag{1.32}$$

к некоторому случайному вектору $\boldsymbol{Z}$. Нам потребуется следующее дополнительное *условие согласованности*: для любого $T \in (0, \infty)$

$$\lim_{n\to\infty} \mathsf{E} \sup_{\|\boldsymbol{t}\| \leqslant T} \left| h_{n,N_n}(\boldsymbol{t}) - h(\boldsymbol{t}) \right| = 0. \tag{1.29}$$

Теорема 1.6. *Пусть случайные векторы* $\boldsymbol{S}_{n,k}$ *имеют вид* (1.34). *Предположим, что семейство случайных векторов* $\{\boldsymbol{Y}_{n,k}\}_{n,k\in\mathbb{N}}$ *слабо относительно компактно, причем выполнено условие согласованности* (1.29). *Тогда сходимость* (1.32) *нормированных многомерных случайных сумм* $\boldsymbol{Z}_n$ *к некоторому случайному вектору* Z *имеет место с некоторыми* $\boldsymbol{c}_n \in \mathbb{R}^m$ *в том и только в том случае, когда существует слабо относительно компактная последовательность случайных векторов* $\boldsymbol{W}_n^* \equiv (U_n^*, (\boldsymbol{V}_n^*)^\top)^\top \in \mathcal{W}(\boldsymbol{Z}|\,\boldsymbol{Y})$, $n \in \mathbb{N}$ *такая, что*

$$\lim_{n\to\infty} \Lambda(\boldsymbol{W}_n^*,\ \boldsymbol{W}_n) = 0. \tag{1.35}$$

Далее получены критерии сходимости распределений сумм случайного числа независимых многомерных случайных величин, к многомерным дисперсионно-сдвиговым смесям нормальных законов и конкретно к обобщенным гиперболическим и обобщенным дисперсионным гамма-распределениям.

Предположим, что суммы $\boldsymbol{S}_{n,k}$ неслучайного числа случайных векторов асимптотически нормальны в том смысле, что существует положительно определенная симметричная $(m \times m)$-матрица Σ такая, что для любого $T \in (0, \infty)$

$$\lim_{n\to\infty} \mathsf{E} \sup_{\|\boldsymbol{t}\| \leqslant T} \left| h_{n,N_n}(\boldsymbol{t}) - \exp\{-\tfrac{1}{2}\boldsymbol{t}^\top \Sigma \boldsymbol{t}\} \right| = 0, \tag{1.39}$$

где $h_{n,k}(\boldsymbol{t})$ – характеристическая функция нормированных и центрированных сумм $\boldsymbol{Y}_{n,k} = b_{n,k}^{-1}(\boldsymbol{S}_{n,k} - \boldsymbol{a}_{n,k})$.

Предположим, что существуют векторы $\boldsymbol{a}_n \in \mathbb{R}^m$ и $\boldsymbol{b}_n \in \mathbb{R}^m$ такие, что для всех $n, k \in \mathbb{N}$ справедливы соотношения

$$\boldsymbol{a}_{n,k} = d_n^{-1} b_{n,k}^2 \boldsymbol{a}_n, \qquad \boldsymbol{c}_n = d_n \boldsymbol{b}_n, \tag{1.40}$$

причем существуют пределы

$$\boldsymbol{a} = \lim_{n\to\infty} \boldsymbol{a}_n, \quad \boldsymbol{b} = \lim_{n\to\infty} \boldsymbol{b}_n. \tag{1.41}$$

Теорема 1.7. *Предположим, что семейство случайных векторов* $\{\boldsymbol{Y}_{n,k}\}_{n,k\in\mathbb{N}}$ *слабо относительно компактно, центрирующие векторы имеют вид* (1.40) *и удовлетворяют условию* (1.41). *Предположим, что суммы* $\mathbf{S}_{n,k}$ *неслучайного числа случайных векторов асимптотически нормальны в смысле* (1.39). *Тогда распределения центрированных и нормированных случайных сумм сходятся к распределению некоторого случайного вектора* $\boldsymbol{Z}$:

$$\boldsymbol{Z}_n \Longrightarrow Z \quad (n \to \infty)$$

тогда и только тогда, когда существует функция распределения G *такая, что* $G(0) = 0$, *распределение* F *случайного вектора* $\boldsymbol{Z}$ *имеет вид* $F = \mathcal{N}_{\boldsymbol{b}+z\boldsymbol{a},\, z\Sigma} \circ G$ *и*

$$\mathsf{P}(U_n^2 < x) \Longrightarrow G(x) \quad (n \to \infty).$$

В качестве следствия этой теоремы получены условия сходимости распределений сумм случайного числа независимых многомерных случайных величин к обобщенным гиперболическим и обобщенным дисперсионным гамма-распределениям.

Теорема 1.8. *Предположим, что семейство случайных векторов* $\{\boldsymbol{Y}_{n,k}\}_{n,k\in\mathbb{N}}$ *слабо относительно компактно, центрирующие векторы имеют вид* (1.40) *и удовлетворяют условию* (1.41). *Предположим, что суммы* $\boldsymbol{S}_{n,k}$ *неслучайного числа случайных векторов асимптотически нормальны в смысле* (1.39).
(*a*) *Распределения центрированных и нормированных случайных сумм сходятся к многомерному обобщенному дисперсионному гамма-распределению* $F(d\boldsymbol{x}; \boldsymbol{a}, \boldsymbol{b}, \Sigma, \nu, \kappa, \delta)$ *при* $n \to \infty$ *тогда и только тогда, когда*

$$\mathsf{P}(U_n^2 < x) \Longrightarrow P_{GG}(x; \nu, \kappa, \delta) \quad (n \to \infty).$$

(*b*) *Распределения центрированных и нормированных случайных сумм сходятся к многомерному обобщенному гиперболическому распределению* $F(d\boldsymbol{x}; \boldsymbol{a}, \boldsymbol{b}, \Sigma, \nu, \mu, \lambda)$ *при* $n \to \infty$ *тогда и только тогда, когда*

$$\mathsf{P}(U_n^2 < x) \Longrightarrow P_{GIG}(x; \nu, \mu, \lambda) \quad (n \to \infty).$$

В разделе 1.3 первой главы предложена простая предельная схема, основанная на элементарных случайных блужданиях типа обобщенных процессов Кокса, в рамках которой происходит формирование конкретных моделей типа подчиненных винеровских процессов – процессов Леви, с обобщенными дисперсионными гамма конечномерными распределениями.

Рассматривается последовательность обобщенных дважды стохастических пуассоновских процессов (обобщенных процессов Кокса) вида

$$Q_n(t) = \sum\nolimits_{i=1}^{N_1^{(n)}(\Lambda_n(t))} X_{n,i}, \quad t \geqslant 0, \tag{1.18}$$

где $\{N_1^{(n)}(t),\, t \geqslant 0\}_{n\geqslant 1}$ – последовательность пуассоновских процессов с единичными интенсивностями; при каждом $n = 1, 2, ...$ случайные величины $X_{n,1}, X_{n,2}, ...$ одинаково распределены; при каждом $n \geqslant 1$ случайные величины $X_{n,1}, X_{n,2}, ...$ и процесс $N_1^{(n)}(t),\ t \geqslant 0$, независимы; при каждом $n = 1, 2, ...$ процесс $\Lambda_n(t),\ t \geqslant 0$, является неубывающим положительным процессом Леви, независимым от процесса

$$Z_n(t) = \sum\nolimits_{i=1}^{N_1^{(n)}(t)} X_{n,i}, \quad t \geqslant 0, \tag{1.19}$$

причем $\Lambda_n(0) = 0$.

Предполагается, что выполнены следующие условия. Пусть $0 < m_n^\beta \equiv \mathsf{E}|X_{n,1}|^\beta < \infty$ при некотором $\beta \in [1, 2]$, существуют $\delta \in (0, 1]$, $\delta_1 \in (0, 1]$ и $C_n \in (0, \infty)$ такие, что для каждого $t \in (0, 1]$ справедливо неравенство

$$\mathsf{E}\Lambda_n^\delta(t) \leqslant (C_n t)^{\delta_1}, \tag{1.20}$$

причем

$$K \equiv \sup_n C_n^{\delta_1/\delta} m_n^\beta < \infty. \tag{1.27}$$

В разделе 1.3 показано, что распределения Вейбулла с $\nu \leqslant 1$ безгранично делимы и приведены необходимые и достаточные условия сходимости обобщенных процессов Кокса к некоторым процессам Леви с одномерными обобщенными дисперсионными гамма-распределениями, в частности, к подчиненному винеровскому процессу, в котором субординатор является процессом Леви–Вейбулла с $\nu \leqslant 1$.

Теорема 1.4. *Пусть обобщенные процессы Кокса $Q_n(t)$ (см. (1.18)) управляются неубывающими положительными процессами Леви $\Lambda_n(t)$, удовлетворяющими условию (1.20) с некоторыми $\delta, \delta_1 \in (0, 1]$. Предположим, что случайные величины $\{X_{n,j}\}_{j\geqslant 1}$, $n = 1, 2, ...$, удовлетворяют условиям*

$$k_n \mathsf{E}(X_{n,1} - a_n)^2 \mathbf{1}(|X_{n,1} - a_n| \geqslant \epsilon) \longrightarrow 0 \ \ и \ \ k_n a_n \longrightarrow a,\ k_n \sigma_n^2 \longrightarrow \sigma^2 \tag{1.28}$$

с некоторыми $k_n \in \mathbb{N}$, $a \in \mathbb{R}$ и $0 < \sigma^2 < \infty$. Также предположим, что условие (1.27) выполнено с $\beta = 2$. Тогда обобщенные процессы Кокса $Q_n(t)$ слабо сходятся в пространстве Скорохода $\mathcal{D}$ к подчиненному винеровскому процессу $W(U(t))$, в котором субординатор $U(t)$ является процессом Леви–Вейбулла с $\nu \leqslant 1$ в том и только том случае, когда

$$\mathsf{P}\big(\Lambda_n(1) < k_n x\big) \xrightarrow{d} P_{GG}(x; \nu, 1, \delta)$$

с теми же самыми k_n.

Во второй главе предлагается принципиально новый метод разделения дисперсионно-сдвиговых смесей нормальных законов, в частности, на примере исследуемых ранее обобщенных гиперболических и обобщенных дисперсионных гамма-распределений.

Традиционно для решения задач разделения смесей вероятностных законов используются EM (Expectation Maximization) алгоритмы. К сожалению, классический EM-алгоритм обладает рядом серьезных недостатков. В частности, он демонстрирует неустойчивость по отношению к исходным данным и начальным приближениям.

Для преодоления этих недостатков предложено много модификаций EM-алгоритма, см., например, [10]. Вместе с тем, в указанной книге предложен и исследован принципиально новый «сеточный» метод приближенного решения задачи разделения смесей. В работе [11] подробно исследованы вопросы сходимости сеточных методов разделения смесей.

На практике приходится решать задачу разделения конечных смесей нормальных законов с произвольно большим числом неизвестных. Хотя в большинстве приложений возникают смеси не более чем с пятью-семью компонентами, даже при использовании таких смесей размерность задачи лежит в пределах от 14 до 20, что существенно увеличивает вычислительные и временные ресурсы, необходимые для практического решения указанных задач. Одним из возможных подходов к снижению размерности является априорное сужение классов допустимых смесей. В частности, класс *дисперсионно-сдвиговых смесей нормальных законов* показал высочайшую адекватность при решении многих задач, связанных с анализом финансовых индексов, атмосферной и плазменной турбулентности. Среди прочих семейств, входящих в данный класс, отметим упомянутые ранее обобщенные гиперболические распределения и семейство обобщенных дисперсионных гамма-распределений.

В разделе 2.2 описаны основные идеи метода. На первом этапе на положительной полупрямой выделяется основная часть носителя смешивающего распределения: ограниченный интервал, вероятность которого, вычисленная в соответствии со смешивающим распределением, практически равна единице. На этот интервал накидывается конечная сетка, содержащая, возможно, очень много *известных* узлов $u_1, \ldots, u_K$. Считая параметр сдвига β равным нулю, приблизим искомое обобщенное гиперболическое распределение конечной смесью нормальных законов:

$$P_{GH}(x;\, \alpha, 0, \nu, \mu, \lambda) \approx \sum\nolimits_{i=1}^{K} p_i \Phi\Big(\frac{x - \alpha u_i}{\sqrt{u_i}}\Big), \quad x \in \mathbb{R}. \tag{2.2}$$

Пусть $p_1^{(m)}, \ldots, p_{K-1}^{(m)}, \alpha^{(m)}$ – оценки параметров $p_1, \ldots, p_{K-1}$ и α на m-й итерации, $p_K^{(m)} = 1 - p_1^{(m)} - \ldots - p_{K-1}^{(m)}$.
Обозначим: $\varphi_{ij}^{(m)} = \frac{1}{\sqrt{u_i}}\varphi\Big(\frac{x_j - \alpha^{(m)} u_i}{\sqrt{u_i}}\Big)$, $g_{ij}^{(m)} = \frac{p_i^{(m)}\varphi_{ij}^{(m)}}{\sum_{r=1}^{K} p_r^{(m)}\varphi_{rj}^{(m)}}$, где $i = 1, \ldots, K$; $j = 1, \ldots, n$.

Итерационный процесс определяется следующим образом.

$$p_i^{(m+1)} = \frac{1}{n}\sum\nolimits_{j=1}^{n} g_{ij}^{(m)}, \quad i = 1, \ldots, K. \tag{2.3}$$

$$\alpha^{(m+1)} = \frac{\overline{x}}{\sum_{i=1}^{K} u_i p_i^{(m+1)}}, \tag{2.4}$$

Теорема 2.1. *Пусть узлы $u_1, \ldots, u_K$ сетки различны, неотрицательны и известны. Итерационный процесс* (2.3) – (2.4) *является монотонным, то есть каждая его итерация не уменьшает целевую сеточную функцию правдоподобия*

$$L(p_1, \ldots, p_K, \alpha; x_1, \ldots, x_n) = \prod\nolimits_{j=1}^{n} \Big[\sum\nolimits_{i=1}^{K} \frac{p_i}{\sqrt{u_i}}\varphi\Big(\frac{x_j - \alpha^{(m)} u_i}{\sqrt{u_i}}\Big)\Big].$$

На втором этапе применяется какой-либо стандартный метод подгонки обобщенного обратного гауссовского распределения $P_{GIG}(z;\, \nu, \mu, \lambda)$ к эмпирическим данным типа гистограммы $(u_1, p_1), \ldots, (u_K, p_K)$. Например, параметры ν, μ и λ можно оценить, минимизируя соответствующую статистику хи-квадрат. Или же, например, можно решить задачу наименьших квадратов

$$(\nu^*, \mu^*, \lambda^*) = \arg\min_{\nu,\mu,\lambda} \sum\nolimits_{i=1}^{K} \left[p_i - \int_{\frac{1}{2}(u_{i-1}+u_i)}^{\frac{1}{2}(u_i+u_{i+1})} p_{GIG}(u;\, \nu, \mu, \lambda)du\right]^2,$$

где $u_0 = 0$, $u_{K+1} = \infty$. Также хорошие результаты показал метод поиска наилучшего распределения в смысле минимизации расстояния Кульбака-Лейблера, который в данном случае эквивалентен максимизации правдоподобия полученной гистограммы в выбранных классах распределений.

При применении указанного двухэтапного метода в динамическом режиме крайне важным становится вопрос о выборе наиболее эффективных и быстродействующих численных процедур и их параметров. В частности, исключительную важность приобретает правильный выбор сетки на первом этапе. В разделе 2.3 диссертации получены теоретические оценки для верхней границы сетки, а в разделе 2.7 приведены дополнительные практические рекомендации.

Предложенный алгоритм был в первую очередь протестирован на большом количестве искуственно сгенерированных выборок с целью оценки качества работы метода на тех данных, природа которых заранее известна.

Для тестирование использовалась задача оценивания параметров обобщенных гиперболических распределений с использованием указанного выше алгоритма выбора сетки с умеренным числом узлов $K = 40$. Для вычислений использовались выборки объемов $n = 1000$ и $n = 10000$ с разными наборами параметров.

В разделе 2.5 приведены результаты численных экспериментов на реальных данных – применения метода к двум известным биржевым индикаторам. Разделение смесей проводилось в режиме скользящего окна с целью изучения динамики данных индикаторов. В качестве семейств подбираемых распределений использовались все те же GH- и GVG-распределения. Из полученных результатов следует, что оба семейства хорошо описывают рассмотренные данные, при этом обобщенные дисперсионные гамма-распределения показывают лучшие результаты по сравнению с обобщенными гиперболическими распределениями при проверке критерия согласия хи-квадрат.

В разделе 2.6 второй главы рассмотрено применение предложенного метода в задаче выявления двигательной активности в головном мозге человека. Формально эта задача сводится к максимально точному определению начала изменения характера слабого полезного сигнала (возникновения ненулевого тренда) на фоне сильного (негауссовского) шума. Для решения этой задачи в разделе 2.6 предложены различные версии комбинированного сеточного алгоритма разделения дисперсионно-сдвиговых смесей, реализуемого в скользящем режиме. В частности,

а) Версии, использующие только первый этап комбинированного метода (оценивание смешивающего распределения) с последующим сравнением дискретных аппроксимаций смешивающих распределений с помощью критерия однородности хи-квадрат. Для этого данные миограммы специально преобразуются и в качестве исходного ряда используются приращения динамической компоненты волатильности (см. [10]).

б) Версии, использующие оба этапа комбинированного метода с оцениваем параметров обобщенных гиперболических распределений, подгоняемых к приращениям миограммы в скользящем режиме. В таких версиях решение о появлении тренда делается по изменениям вектора параметров в нескольких метриках.

Предложенные методы продемонстрировали высокую эффективность и в ряде случаев более высокую точность по сравнению с известными методами.

Третья глава посвящена описанию алгоритма прогнозирования параметров дисперсионно-сдвиговых смесей в общем виде, в частности, для задачи оценки рисков.

У многих специалистов в области практических задач анализа финансовых рынков сложилось вполне обоснованное мнение, что анализировать и прогнозировать нужно не значения наблюдаемых процессов, а их *распределения*. В частности, одной из важнейших практических задач является проблема оценки и прогнозирования рисков, тесно связанная с изучением поведения хвостов распределений наблюдаемых процессов.

Минимальные требования к любому осмысленному методу прогнозирования заключаются в его быстроте и корректности результатов на случайно выбранных исторических данных. В качестве входных данных для метода прогнозирования будем использовать результат работы модифицированного сеточного метода разделения дисперсионно-сдвиговых смесей нормальных законов, то есть ряд параметров распределений, посчитанных для $\hat{N}$ известных окон с историческими наблюдениями, $\theta_1, \theta_2, \ldots, p_N$, где каждое $\theta_i = (\alpha_i, \beta_i, \nu_i, \mu_i, \lambda_i)^T$.

Задача прогнозирования – получить оценки $\theta_{n+1}, \theta_{n+2}, \ldots$ для окон, частично или полностью состоящих из будущих наблюдений.

Рассмотрим соотношение:

$$\tilde{\theta}_{i+1} = F_1\theta_i + F_2\theta_{i-1} + \ldots + F_R\theta_{i-R+1} + \varepsilon, \tag{3.3}$$

где $R \in \mathbb{N}$ – заранее фиксированный параметр, имеющий смысл порядка прогноза,
$F_j \in R^{5\times 5}$ – матрицы-регрессоры.

Процедура поиска параметров модели имеет вид:

$$(F_1, \ldots, F_R) = \arg\min \sum\nolimits_{i=R+1}^{N-1} \left(p_{i+1} - \tilde{p}_{i+1}\right)^2. \tag{3.4}$$

По сути рассматривается типичная авторегрессионная модель, где поиск матриц F_j проводится путем обучения модели с использованием минимизации суммарной остаточной суммы квадратов на $\hat{N} - R$ предсказаниях модели по известных данных.

В этой главе также предложен подход к определению точности получаемых прогнозов, а также приведены результаты практического применения метода на реальных финансовых данных, в частности, прогнозирования

приращений двух известных биржевых индексов. Проведен анализ точности получаемых прогнозов с использованием различных квантилей и метрик.

В заключение третьей главы приведено описание применения алгоритма прогнозирования к задаче анализа текстовой информации с целью предотвращения утечек данных.

Одной из самых актуальных задач информационной безопасности для корпоративного сегмента является обнаружение *внутренних угроз*, в частности, своевременное обнаружение утечек информации. Для решения подобных задач существует класс так называемых DLP-систем (от англ. Data Loss Prevention, или Data Leak Prevention).

Предложенный в главе 3 метод прогнозирования показал хорошие результаты при его применении для решения задачи способом, описанным в [14]. В рамках предложенной процедуры используется простой и легко интерпретируемый подход к сравнению прогнозов с реальными данными на основе критерия согласия хи-квадрат, который позволяет принять обоснованное решение о возможной утечке данных.

Публикации автора по теме диссертации

1. *Корчагин А. Ю.* О сходимости случайных сумм независимых случайных векторов к многомерным обобщенным дисперсионным гамма-распределениям. Системы и средства информатики, М.: ИПИ РАН, 2015 г., том 25, №1, С. 131–146.

2. *Королев В. Ю., Корчагин А. Ю., Зейфман А. И.* О сходимости распределений статистик, построенных по выборкам случайного объема, к многомерным обобщенным дисперсионным гамма-распределениям // Доклады Академии наук, 2015. Т. 462. Вып. 4, с. 10–24.

3. *Королев В. Ю., Корчагин А. Ю.* Модифицированный сеточный метод разделения дисперсионно-сдвиговых смесей нормальных законов // Информатика и ее применения, 2014 г., том 8, Вып. 4, сс. 11–19.

4. *Корчагин А. Ю., Ярошенко И. И.* О практическом использовании модифицированного сеточного метода разделения дисперсионно-сдвиговых смесей нормальных законов // Труды первой научно-практической конференции молодых ученых "Задачи современной информатики – М.: ИПИ РАН, С. 34–38.

[14] *Машечкин И. В., Петровский М. И., Царев Д. В.* Применение методов интеллектуального анализа текстовой информации для предотвращения утечек данных // Программирование, 2015. N. 1. С. 32–43.

5. *Королев В. Ю., Корчагин А. Ю., Морева О. А.* Непараметрическое оценивание функции плотности смесей вероятностных законов с помощью EM-алгоритма. // Системы и средства информатики, М.: ИПИ РАН, 2012г, том 22, Вып. 2, с. 197–227.

6. *Королев В. Ю., Черток А. В., Корчагин А. Ю., Горшенин А. К.* Вероятностно-статистическое моделирование информационных потоков в сложных финансовых системах на основе высокочастотных данных // Информатика и ее применения, 2013 г., том 7, Вып. 1, с. 12–21.

7. *Королев В. Ю., Корчагин А. Ю., Соколов И.А., Черток А. В.* О работах в области моделирования информационных потоков в современных высокочастотных финансовых приложениях // Системы и средства информатики, М.: ИПИ РАН, 2014г, том 24, Вып. 4, с. 63–85.

8. *Korolev V. Yu., Chertok A. V., Korchagin A. Yu., Zeifman A. I.* Modeling high-frequency order flow imbalance by functional limit theorems for two-sided risk processes // Applied Mathematics and Computation (New York), издательство Elsevier BV (Netherlands), 2014 г., том 253, с. 224–241.

9. *Chertok A. V., Korolev V. Yu., Korchagin A. Yu.* On order flow modeling with Cox processes. // XXXII International Seminar on Stability Problems for Stochastic Models, Book of Abstracts. 2014. Moscow, IPI RAN, p. 23 – 24

10. *Gorshenin A. K., Korolev V. Yu, Zeifman A. I., Shorgin S. Ya, Chertok A. V., Evstafyev A. I., Korchagin A. Yu.* Modelling stock order flows with non-homogeneous intensities from high-frequency data // AIP Conference Proceedings, 2013 г., International Symposium on Computational Models for Life Sciences, Vol. 1559, P. 2394–2397.