МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
имени М. В. Ломоносова
Факультет вычислительной математики и кибернетики

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

Диссертация на соискание учёной степени
кандидата физико-математических наук

Научный руководитель:
доктор физико-математических наук,
профессор Королев В.Ю.

Москва — 2015

Оглавление

Введение

Актуальность темы исследования. 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 [30,31,33] 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 [24,62] 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 [37].

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 [16], согласно которому та или иная 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 [42] прослежена глубокая связь этого принципа. Как известно, нормальное распределение об-

ладает 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 [42], 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 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 [18], в которой рассматривались асимптотические свойства распределений выборочных квантилей, построенных по выборкам случайного объема, и было продемонстрирова-

но, 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 [62, 63]. В работе [47] приведены критерии сходимости распределений статистик, построенных по выборкам случайного объема, к многомерным обобщенным гиперболическим 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 [12], 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 [59], 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. 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, [59]. Вместе с тем, в указанной книге предложен и исследован принципиально новый «сеточный» метод приближенного решения задачи разделения смесей. В работе [60] 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 [59], 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 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 про-

цессов 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 [6].

Для решения задачи оценивания параметров обобщенных гиперболических распределений традиционно используется метод, предложенный в статье [64] 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 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.

Используемые подходы и методы. В данной работе используются методы характиристических функций, методы многомерного статистического анализа, прямые вероятностные методы. Помимо этого, используются методы, разработанные непосредственно в диссертации: комбинированых двухэтапный сеточный метод разделения смесей, метод прогнозирования типа авторегрессии.

В методе разделения смесей используется поход снижения размерности исходной задачи путем априорного сужение классов допустимых смесей. Подход к отысканию модели для задачи прогнозирования состоит в решении стандартной задачи минимизации остаточной суммы квадратов. Для настройки алгоритмов используется подход, основанный на многократном прогоне алгоритмов.

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. А именно:

– доказательство критериев 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;

– разработка эффективного комбинированного метода статистического разделения дисперсионно-сдвиговых смесей нормальных законов, в частности, обобщенных гиперболических и обобщенных дисперсионных гамма-распределений, и изучение его свойств;

– демонстрация возможностей 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.

Краткое содержание диссертации. В первой главе приведено описание общих свойств дисперсионно-сдвиговых смесей нормальных законов, а также описаны два конкретных параметрических семейства дисперсионно-сдвиговых смесей нормальных законов: обобщенные гиперболические распределения и обобщенные дисперсионные гамма-распределения. Эти семейства сопоставляются, демонстрируется что в некоторых ситуациях использование обобщенных дисперсионных гамма-распределений в качестве моделей дает лучшие результаты. В этой главе также приводятся предельные теоремы, объясняющие характер смешивающего распределения в конкретных ситуациях и дающие дополнительное обоснование высокой адекватности моделей типа дисперсионно-сдвиговых смесей в рамках асимптотического подхода.

Помимо этого, в первой главе предложена простая предельная схема, основанная на элементарных случайных блужданиях, в рамках которой происходит формирование моделей типа подчиненных винеровских процессов. Приводятся необходимые и достаточные условия сходимости обобщенных процессов Кокса к процессам Леви с одномерными

обобщенными дисперсионными гамма-распределениями. Заключительная часть этой главы посвящена описанию критериев 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 обобщенным гиперболическим и обобщенным дисперсионным гамма-распределениям.

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. Также в этой главе изучаются основные свойства метода и предложены практические рекомендации по его использованию. В частности, при использовании этого метода в динамическом режиме крайне важным становится вопрос о выборе наиболее эффективных и быстродействующих численных процедур и их параметров. Приведены результаты работы метода как на искусственно сгенерированных выборках, так и на реальных данных. В частности, рассмотрено применение предложенного метода в задаче выявления двигательной активности в головном мозге человека.

Третья глава посвящена описанию алгоритма прогнозирования параметров дисперсионно-сдвиговых смесей в общем виде, в частности, для задачи оценки рисков. Предложен подход к определению точности получаемых прогнозов, а также приведены результаты практического применения метода на реальных финансовых данных. Помимо этого, алгоритм прогнозирования применен к задаче анализа текстовой информации с целью предотвращения утечек данных.

Основные результаты.

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 помо-

щью 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.

Апробация диссертации. 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 (Тронхайм, 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.

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. Список литературы содержит 82 наименования.

Глава 1

Дисперсионно-сдвиговые смеси нормальных законов и их свойства

1.1 Используемые обозначения

Введем обозначения, которые в дальнейшем будут использоваться без дополнительных комментариев. Пусть $m \in \mathbb{N}$. Векторы $\boldsymbol{x} = (x^{(1)}, \ldots, x^{(m)})^\top$ считаются элементами m-мерного евклидова пространства $\mathbb{R}^m$, верхний индекс $^\top$ означает транспонирование вектора или матрицы. Скалярное произведение в $\mathbb{R}^m$ будет обозначаться $\langle \cdot, \cdot \rangle$:

$$\langle \boldsymbol{x},\, \boldsymbol{y} \rangle = \boldsymbol{x}^\top \boldsymbol{y} = x^{(1)} y^{(1)} + \ldots + x^{(m)} y^{(m)}.$$

Как обычно, евклидова норма вектора $\boldsymbol{x}$ обозначается $\|\boldsymbol{x}\| = \langle \boldsymbol{x},\, \boldsymbol{x} \rangle^{1/2}$. Если A – вещественная квадратная $(m \times m)$-матрица, то ее определитель обозначается $\det(A)$. Тождественная (единичная) $(m \times m)$-матрица обозначается $\boldsymbol{I}$. Чтобы различать число нуль и нулевой вектор, используются обозначения $0 \in \mathbb{R}$ и $\mathbf{0} = (0, \ldots, 0)^\top \in \mathbb{R}^m$.

Тот факт, что случайный вектор $\boldsymbol{X}$ имеет m-мерное нормальное распределение с вектором средних $\boldsymbol{a}$ и ковариационной матрицей Σ, будем обозначать $\boldsymbol{X} \sim \mathcal{N}_{\boldsymbol{a},\Sigma}$. Это означает, что существует вектор $\boldsymbol{a} \in \mathbb{R}^m$ и неотрицательно определенная симметричная матрица Σ размерности $m \times m$, такие что характеристическая функция $\mathfrak{f}_{\boldsymbol{X}}$ вектора $\boldsymbol{X}$ имеет вид

$$\mathfrak{f}_{\boldsymbol{X}}(\boldsymbol{t}) = \mathsf{E} \exp\{i \langle \boldsymbol{t}, \boldsymbol{X} \rangle\} = \exp\{i \boldsymbol{a}^\top \boldsymbol{t} - \tfrac{1}{2} \boldsymbol{t}^\top \Sigma \boldsymbol{t}\}, \qquad \boldsymbol{t} \in \mathbb{R}^m.$$

Плотность невырожденного нормального распределения $\mathcal{N}_{\boldsymbol{a},\Sigma}$ имеет вид

$$\varphi_{\boldsymbol{a},\Sigma}(\boldsymbol{x}) = \frac{1}{(2\pi)^{m/2} |\Sigma|^{1/2}} \exp\{-\tfrac{1}{2} (\boldsymbol{x} - \boldsymbol{a})^\top \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{a})\}, \qquad x \in \mathbb{R}^m,$$

где $|\Sigma|$ – определитель матрицы Σ, а Σ^{-1} – матрица, обратная к Σ. Одномерную стандартную нормальную функцию распределения будем обозначать $\Phi(x)$,

$$\Phi(x) = \int\limits_{-\infty}^{x} \varphi(z)dz, \quad \varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\Big\{ -\frac{x^2}{2} \Big\}, \quad x \in \mathbb{R}.$$

Будем считать, что все случайные величины и векторы, упоминаемые ниже, заданы на одном и том же вероятностном пространстве $(\Omega, \mathfrak{F}, \mathsf{P})$. Символы $\mathfrak{B}_m$ и $\mathfrak{B}_+$ соответственно обозначают борелевские σ-алгебры подмножеств $\mathbb{R}_m$ и $\mathbb{R}_+ \equiv [0, \infty)$. Символами $\stackrel{d}{=}$ и $\Longrightarrow$ соответственно будут обозначаться совпадение распределений и сходимость по распределению. Распределение случайного вектора $\boldsymbol{X}$ будет обозначаться $\mathcal{L}(\boldsymbol{X})$.

Семейство $\{\boldsymbol{X}_j\}_{j\in\mathbb{N}}$ $\mathbb{R}^m$-значных случайных векторов называется *слабо относительно компактным*, если каждая последовательность его элементов содержит подпоследовательность, сходящуюся по распределению. Иногда вместо *«слабая относительная компактность»* будет использоваться термин *«слабая компактность»*. В конечномерном случае, рассматриваемом в данной работе, слабая относительная компактность семейства $\{\boldsymbol{X}_j\}_{j\in\mathbb{N}}$ эквивалентна его плотности (tightness)

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

(см., например, [28]).

1.2 Определение и некоторые свойства дисперсионно-сдвиговых смесей нормальных законов

1.2.1 Определение и элементарные свойства дисперсионно-сдвиговых смесей нормальных законов

Данная глава посвящена описанию общих свойств дисперсионно-сдвиговых смесей нормальных законов, а также описанию двух конкретных параметрических семейств дисперсионно-сдвиговых смесей нормальных законов: обобщенных гиперболических распределений и обобщенных дисперсионных гамма-распределений. В этой главе предпринята попытка сопоставить эти семейства и продемонстрировать, что в некоторых ситуациях использование обобщенных дисперсионных гамма-распределений в качестве моделей наблюдаемых статистических закономерностей дает лучшие результаты. Также приводятся предельные теоремы, объясняющие характер смешивающего распределения в конкретных ситуациях и дающих дополнительное обоснование высокой адекватности моделей типа дисперсионно-сдвиговых смесей в рамках асимптотического подхода.

Одномерные дисперсионно-сдвиговые смеси нормальных законов

Понятие дисперсионно-сдвиговой смеси нормальных законов (normal variance-mean mixture) введено в семидесятых-восьмидесятых годах прошлого столетия в работах О.-Е. Барндорфф-Нильсена и его коллег [30,31,33] как довольно гибкое обобщение нормального распределения. Сначала рассмотрим одномерный случай.

Пусть $\beta \in \mathbb{R}$, $\alpha \in \mathbb{R}$, $0 < \sigma < \infty$, $G(x)$ – функция распределения, все точки роста которой сосредоточены на $\mathbb{R}_+$. Дисперсионно-сдвиговой смесью нормальных законов называется функция распределения

$$F(x) = \int\limits_0^\infty \Phi\Big(\frac{x-\beta-\alpha z}{\sigma\sqrt{z}}\Big)\, dG(z), \quad x \in \mathbb{R}. \tag{1.1}$$

Обратим внимание, что 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*.

Если Y и U – независимые случайные величины с функциями распределения $\Phi(x)$ и $G(x)$ (при этом, очевидно, $\mathsf{P}(U \geqslant 0) = 1$), а Z – случайная величина, удовлетворяющая соотношению

$$Z \stackrel{d}{=} \sigma Y\sqrt{U} + \alpha U + \beta,$$

то $\mathsf{P}(Z < x) = F(x)$, $x \in \mathbb{R}$.

Легко убедиться, что если $\mathsf{E}U < \infty$, то

$$\mathsf{E}Z = \beta + \alpha\mathsf{E}U,$$

а если при этом и $\mathsf{E}U^2 < \infty$, то

$$\mathsf{E}Z^2 = \beta^2 + (\sigma^2 + 2\alpha\beta)\mathsf{E}U + \alpha^2\mathsf{E}U^2, \quad \mathsf{D}Z = \alpha^2\mathsf{D}U + \sigma^2\mathsf{E}U.$$

При вычислении моментов более высоких порядков можно использовать следующее соотношение между характеристическими функциями $\mathfrak{f}_Z$ и $\mathfrak{f}_U$ случайных величин Z и U:

$$\mathfrak{f}_Z(t) = e^{it\beta}\mathfrak{f}_U\big(\alpha t + \tfrac{1}{2}i\sigma^2t^2\big), \quad t \in \mathbb{R}. \tag{1.2}$$

Чтобы убедиться в справедливости (1.2), достаточно заметить, что по теореме Фубини

$$\mathfrak{f}_Z(t) = \mathsf{E}e^{itZ} = \mathsf{E}\exp\{it(\beta + \sigma\sqrt{U}\cdot Y + \alpha U)\} = \int\limits_0^\infty \mathsf{E}\exp\{it\beta + \sigma\sqrt{u}\cdot Y + \alpha u)\}\, dG(u) =$$

$$= e^{it\beta}\int\limits_0^\infty \exp\left\{it\alpha u - \tfrac{1}{2}\sigma^2 t^2 u\right\} dG(u) = e^{it\beta}\int\limits_0^\infty \exp\left\{iu\big(\alpha t + \tfrac{1}{2}i\sigma^2 t^2\big)\right\} dG(u) =$$

$$= e^{it\beta}\mathfrak{f}_U\big(\alpha t + \tfrac{1}{2}i\sigma^2 t^2\big), \quad t \in \mathbb{R}.$$

Многомерные дисперсионно-сдвиговые смеси нормальных законов

Многомерные дисперсионно-сдвиговые смеси нормальных законов также были введены О.-Е. Барндорфф-Нильсеном и его коллегами в работах [30, 31, 33]. По аналогии с одномерной ситуацией, будем говорить, что распределение $\mathbb{R}^m$-значного случайного вектора $\boldsymbol{Z}$ является многомерной дисперсионно-сдвиговой смесью нормальных законов, если $\boldsymbol{Z} \stackrel{d}{=} \boldsymbol{b} + U\,\boldsymbol{a} + \sqrt{U}\,A\,\boldsymbol{Y}$, где $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^m$, A – вещественная $(m \times m)$-матрица такая, что матрица $\Sigma \equiv AA^\top$ 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 $\boldsymbol{Y}$. Эквивалентно, будем говорить, что вероятностная мера F на $(\mathbb{R}^m, \mathfrak{B}_m)$ является многомерной дисперсионно-сдвиговой смесью нормальных законов, если

$$F(d\boldsymbol{x}) = \int\limits_0^\infty \mathcal{N}_{\boldsymbol{b}+z\boldsymbol{a},\, z\Sigma}(d\boldsymbol{x})G(dz), \tag{1.3}$$

где смешивающим распределением G является вероятностная мера на $(\mathbb{R}_+, \mathfrak{B}_+)$. Этот факт иногда будем записывать в виде $F = \mathcal{N}_{\boldsymbol{b}+z\boldsymbol{a},\, z\Sigma} \circ G$.

1.2.2 Обобщенные гиперболические распределения и некоторые их свойства

Одномерные обобщенные гиперболические 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 $\nu \in \mathbb{R}$,

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

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

$$K_\nu(z) = \frac{1}{2}\int\limits_0^\infty y^{\nu-1}\exp\Big\{-\frac{z}{2}\Big(y + \frac{1}{y}\Big)\Big\}dy, \quad z \in \mathbb{C},\ \operatorname{Re} z > 0.$$

Соответствующую функцию распределения обозначим $P_{GIG}(x;\nu,\mu,\lambda)$,

$$P_{GIG}(x;\nu,\mu,\lambda)=\begin{cases}0, & x<0,\\ \int\limits_0^x p_{GIG}(z;\nu,\mu,\lambda)dz, & x\geqslant 0.\end{cases}$$

Как сказано в статье [53], обобщенное обратное гауссовское распределение введено в 1946 г. Этьеном Альфеном (Étienne Halphen), который использовал его для описания объема воды, проходящего ежемесячно через гидроэлектростанции. В [53] обобщенное обратное гауссовское распределение названо *распределением Альфена*. К сожалению, автору не удалось обнаружить оригинальные работы самогó Альфена. В 1973 г. это распределение было переоткрыто Гербертом Зихелем [54], который использовал его в качестве смешивающего закона при рассмотрении специальных смешанных пуассоновских распределений (*распределений Зихеля*, см., например, [22]) как дискретных распределений с тяжелыми хвостами. В 1977 г. эти распределения снова переоткрыл О.-Э. Барндорфф-Нильсен [30,31], который, в частности, использовал их для описания распределения размеров частиц.

Класс обобщенных обратных гауссовских распределений довольно обширен и содержит, в частности, как распределения с экспоненциально убывающими хвостами (гамма-распределение ($\mu=0$, $\nu>0$)), так и распределения с хвостами, убывающими степенным образом (обратное гамма-распределение ($\lambda=0$, $\nu<0$), обратное гауссовское распределение ($\nu=-\frac{1}{2}$) и его предельный при $\lambda\to 0$ случай – распределение Леви (устойчивое распределение с характеристическим показателем, равным $\frac{1}{2}$, сосредоточенное на неотрицательной полуоси – распределение времени достижения стандартным винеровским процессом единичного уровня)).

В 1977–78 гг. О.-Э. Барндорфф-Нильсен [30, 31] 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

$$P_{GH}(x;\alpha,\beta,\nu,\mu,\lambda)=\int\limits_0^\infty \Phi\Big(\frac{x-\beta-\alpha z}{\sqrt{z}}\Big)\,p_{GIG}(z;\nu,\mu,\lambda)dz,\quad x\in\mathbb{R}. \tag{1.5}$$

В оригинальном определении обобщенных гиперболических распределений в работах [30, 31] масштабный параметр β полагался равным единице, но присутствовал еще параметр сдвига $\mu\in\mathbb{R}$, которым в (1.5) без потери общности пренебрегаем, так как его легко ввести в модель, переобозначив $x\mapsto x-\mu$.

Несложно убедиться, что плотность $p_{GH}(x;\alpha,\beta,\nu,\mu,\lambda)$ обобщенного гиперболического распределения имеет вид

$$p_{GH}(x;\alpha,\beta,\nu,\mu,\lambda)=\int\limits_0^\infty \frac{1}{\sqrt{z}}\varphi\Big(\frac{x-\beta-\alpha z}{\sqrt{z}}\Big)p_{GIG}(z;\nu,\mu,\lambda)\,dz=$$

$$=\int\limits_0^\infty \frac{1}{\sqrt{2\pi z}}\exp\Big\{-\frac{(x-\beta-\alpha z)^2}{2z}\Big\}\frac{\lambda^{\nu/2}z^{\nu-1}}{2\mu^{\nu/2}K_\nu\big(\sqrt{\mu\lambda}\big)}\exp\Big\{-\frac{1}{2}\Big(\frac{\mu}{z}+\lambda z\Big)\Big\}dz=$$

$$=\frac{\lambda^{\nu/2}(\lambda+\alpha^2)^{\nu/2-1/4}}{2\sqrt{2\pi}\mu^{\nu/2}K_\nu\big(\sqrt{\mu\lambda}\big)}\cdot\big(\mu+(x-\beta)^2\big)^{1/4-\nu/2}\exp\{\alpha(x-\beta)\}\,K_{\nu-1/2}\Big(\sqrt{\big(\mu+(x-\beta)^2\big)\big(\lambda+\alpha^2\big)}\Big).$$

Класс обобщенных гиперболических распределений очень широк и содержит, в частности,

(a) симметричные и скошенные (skew) распределения Стьюдента (в том числе распределение Коши), которым в представлении (1.5) соответствуют смешивающие обратные гамма-распределения;

(b) дисперсионные гамма-распределения (Variance Gamma (VG) distributions) (в том числе симметричные и несимметричные распределения Лапласа), которым в представлении (1.5) соответствуют смешивающие гамма-распределения;

(c) нормальные\\обратные нормальные (NIG) распределения, которым в представлении (1.5) соответствуют смешивающие обратные нормальные распределения;

(d) и многие другие типы распределений.

Обобщенные гиперболические распределения продемонстрировали высочайшую адекватность при их применении для описания статистических закономерностей поведения различных характеристик сложных открытых систем, в частности турбулентных систем и финансовых рынков. Публикации, посвященные моделям, основанным на обобщенных гиперболических распределениях, исчисляются сотнями. Достаточно упомянуть лишь канонические работы [29, 31, 32, 34, 35, 38–41, 50, 51, 67]. Согласно расхожему мнению, столь высокая адекватность обобщенных гиперболических моделей может быть формально объяснена большим числом настраиваемых параметров, позволяющим подогнать какую угодно модель к каким угодно данным. Среди статистиков хорошо известно высказывание Ж. Бертрана «Give me four parameters and I shall describe an elephant; with five, it will wave its trunk» (цитируется по статье Л. ЛеКама [49]). Это обстоятельство, конечно же, играет свою роль, однако на самом деле модели типа (1.5) в большинстве случаев адекватны по гораздо более естественным глубоким причинам, обсуждению которых в значительной мере посвящена данная глава.

Многомерные обобщенные гиперболические распределения

В соответствии со сказанным выше, по аналогии с одномерной ситуацией, будем говорить, что распределение $\mathbb{R}^m$-значного случайного вектора $\boldsymbol{Z}$ является многомерным обобщенным гиперболическим, если $\boldsymbol{Z} \stackrel{d}{=} \boldsymbol{b} + U\,\boldsymbol{a} + \sqrt{U}A\,\boldsymbol{Y}$, где $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^m$, A – вещественная $(m \times m)$-матрица такая, что матрица $\Sigma \equiv AA^\top$ положительно определена, $\boldsymbol{Y}$ – случайный вектор со стандартным m-мерным нормальным распределением $\mathcal{N}_{\boldsymbol{0},\boldsymbol{I}}$, а U – независимая от $\boldsymbol{Y}$ случайная величина, имеющая обобщенное обратное гауссовское распределение. Аналогично, будем говорить, что вероятностная мера $F(d\boldsymbol{x};\, \boldsymbol{a},\, \boldsymbol{b}, \Sigma, \nu, \mu, \lambda)$ на $(\mathbb{R}^m,\, \mathfrak{B}_m)$ задает многомерное обобщенное гиперболическое распределение, если $F(B;\, \boldsymbol{a},\, \boldsymbol{b}, \Sigma, \nu, \mu, \lambda) = (\mathcal{N}_{\boldsymbol{b}+z\boldsymbol{a},\, z\Sigma} \circ P_{GIG}(dz; \nu, \mu, \lambda)(B)$ для любого $B \in \mathfrak{B}_m$, то есть

$$F(d\boldsymbol{x};\, \boldsymbol{a},\, \boldsymbol{b}, \Sigma, \nu, \mu, \lambda) = \int\limits_0^\infty \mathcal{N}_{\boldsymbol{b}+z\boldsymbol{a},\, z\Sigma}(d\boldsymbol{x}) P_{GIG}(dz; \nu, \mu, \lambda).$$

1.2.3 Обобщенные дисперсионные гамма-распределения и некоторые их свойства

Одномерные обобщенные дисперсионные гамма-распределения

Гамма-распределение и обратное гамма-распределение являются частными представителями класса обобщенных гамма-распределений, важная роль которых в моделировании и анализе стохастической структуры информационных потоков описана в книге [23]. Обобщенные гамма-распределения (GG-распределения) были впервые описаны как единое семейство в 1962 г. в работе [56] в качестве семейства вероятностных моделей, включающего в себя одновременно гамма-распределения и распределения Вейбулла.

Обобщенным гамма-распределением называется распределение, 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

$$\Gamma(\kappa) = \int\limits_0^\infty x^{\kappa-1} e^{-x} dx$$

обозначает эйлерову гамма-функцию. Функцию распределения, соответствующую плотности $p_{GG}(x; \nu, \kappa, \delta)$, обозначим $P_{GG}(x; \nu, \kappa, \delta)$, $x \geqslant 0$.

Семейство GG-распределений включает в себя практически все наиболее популярные абсолютно непрерывные распределения. В частности, семейство GG-распределений содержит следующие распределения:

1. Гамма-распределение ($\nu = 1$):

$$p_{GG}(x;\kappa,\theta) = \frac{1}{\Gamma(\kappa)}\theta^{\kappa}x^{\kappa-1}e^{-\theta x}, \quad x \geqslant 0,\ \kappa > 0,\ \theta > 0;$$

1.1. Показательное (экспоненциальное) распределение ($\nu = 1$, $\kappa = 1$):

$$p_{GG}(x;\theta) = \theta e^{-\theta x}, \quad x \geqslant 0,\ \theta > 0;$$

1.2. Распределение Эрланга ($\nu = 1$, $\kappa \in \mathbb{N}$):

$$p_{GG}(x;\kappa,\theta) = \frac{1}{\Gamma(\kappa)}\theta^{\kappa}x^{\kappa-1}e^{-\theta x}, \quad x \geqslant 0,\ \kappa > 0,\ \theta > 0;$$

1.3. Распределение Хи-квадрат ($\nu = 1$, $\delta = 2$):

$$p_{GG}(x;n) = \frac{1}{2\Gamma(\frac{n}{2})}\Big(\frac{x}{2}\Big)^{n/2-1}e^{-x/2}, \quad x \geqslant 0,\ n \in \mathbb{N};$$

2. Распределение Накагами ($\nu = 2$):

$$p_{GG}(x;\mu,\lambda) = \frac{2(\lambda\mu)^{\mu}}{\Gamma(\mu)}x^{2\mu-1}e^{-\lambda\mu x^2}, \quad x \geqslant 0,\ \mu > 0,\ \lambda > 0;$$

2.1. Полунормальное распределение (распределение максимума винеровского процесса на отрезке $[0,1]$) ($\nu = 2$, $\kappa = \frac{1}{2}$):

$$p_{GG}(x;\delta) = \sqrt{\frac{2}{\pi\delta}}\exp\Big\{-\frac{x^2}{2\delta^2}\Big\}, \quad x \geqslant 0,\ \delta > 0;$$

2.2. Распределение Рэлея ($\nu = 2$, $\kappa = 1$):

$$p_{GG}(x;\delta) = \frac{x}{\delta^2}\exp\Big\{-\frac{x^2}{2\delta^2}\Big\}, \quad x \geqslant 0,\ \delta > 0;$$

2.3. Хи-распределение ($\nu = 2$, $\delta = \sqrt{2}$):

$$p_{GG}(x;n) = \frac{1}{2^{n/2-1}\Gamma(\frac{n}{2})}x^{n-1}\exp\Big\{-\frac{x^2}{2}\Big\}, \quad x \geqslant 0,\ n \in \mathbb{N};$$

2.4. Распределение Максвелла (распределение модулей скоростей движения молекул в разреженном газе) ($\nu = 2$, $\kappa = 3/2$):

$$p_{GG}(x;\delta) = \sqrt{\frac{2}{\pi}}\frac{x^2}{\delta^3}\exp\Big\{-\frac{x^2}{2\delta^2}\Big\}, \quad x \geqslant 0,\ \delta > 0;$$

3. Распределение Вейбулла–Гнеденко ($\kappa = 1$):

$$p_{GG}(x;\eta,\mu) = \frac{\eta x^{\eta-1}}{\mu^\eta}\exp\Big\{-\Big(\frac{x}{\mu}\Big)^\eta\Big\}, \quad x \geqslant 0,\ \eta > 0,\ \mu > 0;$$

4. Обратное гамма-распределение ($\nu = -1$):

$$p_{GG}(x;\mu,\lambda) = \frac{1}{\mu\lambda\Gamma(\lambda)}\Big(\frac{\mu\lambda}{x}\Big)^{\lambda+1}\exp\Big\{-\frac{\mu\lambda}{x}\Big\}, \quad x \geqslant 0,\ \lambda > 0,\ \mu > 0;$$

4.1. Распределение Леви ($\nu = -1$, $\kappa = 1/2$):

$$p_{GG}(x;\mu) = \sqrt{\frac{\mu}{2\pi}}\frac{1}{x^{3/2}}\exp\Big\{-\frac{\mu}{2x}\Big\}, \quad x \geqslant 0,\ \mu > 0;$$

5. Логнормальное распределение ($\kappa \to \infty$):

$$p_{GG}(x;\mu,\delta) = \frac{1}{\delta x\sqrt{2\pi}}\exp\Big\{-\frac{(\log x - \mu)^2}{2\delta^2}\Big\}, \quad x \geqslant 0,\ \mu \in \mathbb{R},\ \delta > 0.$$

Широкая применимость GG-распределений обусловлена возможностью их использования в качестве адекватных асимптотических аппроксимаций, поскольку практически все они выступают в качестве предельных в различных предельных теоремах теории вероятностей. А именно:

– показательное распределение выступает в качестве предельного как в схеме максимума (минимума) (см., например, [19]), так и в схеме геометрического суммирования, описывая распределение времени восстановления в прореженных процессах восстановления, выступающих моделями потоков редких событий (см., например, [45]);

– гамма-распределение является безгранично делимым и потому выступает в качестве предельного для распределений сумм независимых равномерно предельно малых случайных величин. При этом распределение Эрланга возникает как допредельное распределение суммы независимых экспоненциально распределенных случайных величин, что в терминах случайной интенсивности может означать, что если случайная интенсивность потока поступления запросов имеет гамма-распределение со значимым параметром формы, то при обработке этих запросов в основном задействованы механизмы последовательной обработки информации;

– распределение Вейбулла–Гнеденко принадлежит к так называемому первому типу предельных распределений экстремальных порядковых статистик (минимума или максимума) (см., например, [19]), что в терминах случайной интенсивности может означать, что если случайная интенсивность потока поступления запросов имеет распределение Вейбулла–Гнеденко со значимым параметром степени, то при обработке этих запросов в основном задействованы механизмы параллельной обработки информации;

– полунормальное распределение (распределение модуля стандартной нормальной случайной величины) возникает как предельное для максимальных частичных сумм независимых случайных величин (см., например, [20]);

– распределение Леви принадлежит к классу устойчивых законов и потому является предельным для сумм независимых одинаково распределенных случайных величин. Оно также является распределением времени достижения стандартным винеровским процессом (процессом броуновского движения) фиксированного уровня;

– логнормальное распределение выступает в качестве предельного для распределения размера частиц при дроблении (см., например, [21]).

Эти свойства GG-распределений обосновывают, в частности, целесообразность моделирования с их помощью распределения случайной интенсивности потока запросов в информационных системах. Это семейство также широко используется в других прикладных задачах в самых разных областях, см., например, [23].

Хорошо известные четырехпараметрические семейства скошенных распределений Стьюдента и дисперсионных гамма-распределений являются подклассами введенного в работе [24] пятипараметрического семейства распределений

$$P_{GVG}(x;a,\sigma,\nu,\kappa,\delta)=\int\limits_0^\infty \Phi\left(\frac{x-au}{\sigma\sqrt{u}}\right)p_{GG}(u;\nu,\kappa,\delta)du, \tag{1.7}$$

где $p_{GG}(u;\nu,\kappa,\delta)$ – плотность GG-распределения (1.6). В статье [24] распределения вида (1.7) названы *обобщенными дисперсионными гамма-распределениями*.

Задача поиска универсальной модели статистических закономерностей во многих областях, в частности в финансовой математике или в физике плазмы, подобна задаче поиска «философского камня» в алхимии и поэтому не имеет точного решения. Однако, основываясь на вышеперечисленных аналитических и асимптотических свойствах представителей семейства GG-распределений и предельных теоремах для сумм независимых случайных величин как теоретико-вероятностной формализации принципа неубывания неопределенности в сложных системах [42], можно утверждать, что семейство обобщенных дисперсионных гамма-распределений является *практически* универсальным для многих задач. Оно

представляется еще более гибкой моделью, нежели обобщенные гиперболические распределения, так как класс обобщенных гамма-распределений в определенном смысле шире класса обобщенных обратных гауссовских распределений, поскольку, в отличие от последнего, он содержит распределения вейбулловского (экспоненциально-степенного) типа с произвольным показателем степени в экспоненте.

Возможное поведение хвостов обобщенных дисперсионных гамма-распределений

Для определенности в данном разделе будем считать, что $\alpha = \beta = 0$. Дисперсионно-сдвиговая смесь нормальных законов с такими параметрами является чисто масштабной смесью нормальных законов (и стало быть, соответствующее ей распределение симметрично). В работе [12] показано, что асимптотическое поведение хвостов масштабных смесей нормальных законов всецело определяется аналогичным поведением хвостов смешивающих законов. В частности, если все точки роста функции распределения $G(x)$ лежат на $\mathbb{R}_+$,

$$F(x) = \int_0^\infty \Phi\Big(\frac{x}{\sqrt{u}}\Big)\, dG(u), \quad x \in \mathbb{R},$$

$L(x)$ – медленно меняющаяся функция, $0 < \rho < 2$, $0 < \gamma < 2$, то для того чтобы

$$\liminf_{x\to\infty} \frac{-\ln[1 - F(x)]}{x^\rho L(x)} = \frac{1}{\gamma},$$

необходимо и достаточно, чтобы

$$\liminf_{x\to\infty} \frac{-\ln[1 - G(x)]}{x^{\frac{\rho}{2-\rho}} \big[L\big(x^{\frac{1}{2-\rho}}\big)\big]^{\frac{2}{2-\rho}}} = \tfrac{1}{2}\gamma^{\frac{2}{\rho-2}}.$$

К примеру, для того чтобы хвосты смеси убывали вейбулловским образом, необходимо и достаточно, чтобы точно так же (с точностью до параметра масштаба) вели себя хвосты смешивающего распределения. Такая ситуация возможна в рамках семейства обобщенных дисперсионных гамма распределений (поскольку распределение Вейбулла принадлежит семейству обобщенных гамма-распределений), и невозможна в рамках обобщенных гиперболических законов. В то же время, как показывает статистический анализ некоторых данных, например, связанных с финансовыми рынками, именно вейбулловский характер убывания присущ хвостам наблюдаемых распределений [36, 37, 55].

Таким образом, семейство обобщенных дисперсионных гамма-распределений представляется более универсальным, нежели обобщенные гиперболические распределения. Этот вывод подтверждается статистическим анализом конкретных данных, см. ниже.

Многомерные обобщенные дисперсионные гамма-распределения

По аналогии с обобщенными гиперболическими распределениями, будем говорить, что распределение $\mathbb{R}^m$-значного случайного вектора $\boldsymbol{Z}$ является многомерным обобщенным гиперболическим, если $\boldsymbol{Z} \stackrel{d}{=} \boldsymbol{b} + U\,\boldsymbol{a} + \sqrt{U}A\,\boldsymbol{Y}$, где $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^m$, A – вещественная $(m \times m)$-матрица такая, что матрица $\Sigma \equiv AA^\top$ положительно определена, $\boldsymbol{Y}$ – случайный вектор со стандартным m-мерным нормальным распределением $\mathcal{N}_{\boldsymbol{0},\boldsymbol{I}}$, а U – независимая от $\boldsymbol{Y}$ случайная величина, имеющая обобщенное гамма-распределение. Эквивалентно, будем говорить, что вероятностная мера $F_{GVG}(d\boldsymbol{x}; \boldsymbol{a}, \boldsymbol{b}, \Sigma, \nu, \kappa, \delta)$ на $(\mathbb{R}^m, \mathfrak{B}_m)$ задает многомерное обобщенное гиперболическое распределение, если $F_{GVG}(B; \boldsymbol{a}, \boldsymbol{b}, \Sigma, \nu, \kappa, \delta) = (\mathcal{N}_{\boldsymbol{b}+z\boldsymbol{a},\, z\Sigma} \circ P_{GG}(dz; \nu, \kappa, \delta)(B)$ для любого $B \in \mathfrak{B}_m$, то есть

$$F_{GVG}(d\boldsymbol{x}; \boldsymbol{a}, \boldsymbol{b}, \Sigma, \nu, \kappa, \delta) = \int_0^\infty \mathcal{N}_{\boldsymbol{b}+z\boldsymbol{a},\, z\Sigma}(d\boldsymbol{x}) P_{GG}(dz; \nu, \kappa, \delta).$$

1.2.4 Дисперсионно-сдвиговые смеси нормальных законов как предельные для распределений случайных сумм независимых одинаково распределенных случайных величин

В данном разделе приведен обзор ассимптотических свойств дисперсионно-сдвиговых смесей нормальных законов других авторов для понимания контекста. Важно отметить, что этот раздел не содержит результатов диссертации.

В прикладной теории вероятностей принято, что ту или иную модель можно считать в достаточной мере обоснованной (адекватной) только тогда, когда она является *асимптотической аппроксимацией*, то есть когда существует довольно простая предельная схема (например, схема суммирования) и соответствующая предельная теорема, в которой рассматриваемая модель выступает в качестве предельного распределения [16]. В первоисточниках упомянутые выше обобщенные гиперболические модели вводились чисто умозрительно как распределения процесса броуновского движения со случайным временем, в каждый момент имеющим то или иное обобщенное обратное гауссовское распределение. Лишь в статье [33] со ссылкой на работу А. Реньи [52] имеется довольно расплывчатое предположение, что дисперсионно-сдвиговые смеси нормальных законов могут быть предельными для сумм случайного числа случайных величин.

Однако, несмотря на то, что свойства обобщенных гиперболических распределений изучены довольно полно, до недавнего времени не было дано корректного доказательства того факта, что обобщенные гиперболические распределения выступают в качестве предельных в простейшей схеме случайно остановленных случайных блужданий. А значит, приводимая в некоторых работах аргументация, связывающая смешивание в модели (1.5) со случайным характером волатильности при применении обобщенных гиперболических распределений в финансовой математике, не имела строгого формального обоснования.

Возможно, причина в том, что в схеме «нарастающих» сумм, рассматривавшейся в [52], полное решение указанной задачи невозможно. Его можно получить лишь, рассматривая случайные суммы в рамках асимптотической схемы серий. Основополагающей работой в этом направлении стала работа Б. В. Гнеденко и Х. Фахима [17].

«Асимптотическое» обоснование некоторых из упомянутых выше моделей было дано лишь недавно в статьях [24, 25], где показано, что скошенные распределения Стьюдента и дисперсионные гамма-распределения могут выступать в качестве предельных в довольно простых предельных теоремах для регулярных статистик, построенных по выборкам случайного объема, в частности в схеме случайного суммирования случайных величин, и, следовательно, могут считаться *естественными* асимптотическими аппроксимациями для распределений многих процессов, например, сходных с неоднородными случайными блужданиями.

В статье [63] приведена общая теорема о необходимых и достаточных условиях сходимости распределений сумм случайного числа независимых одинаково распределенных случайных величин к однопараметрическим сдвиг-масштабным смесям нормальных законов и в качестве следствия из нее получены необходимые и достаточные условия сходимости распределений случайных сумм независимых одинаково распределенных случайных величин к обобщенным гиперболическим распределениям. На примере довольно общего и просто интерпретируемого частного случая – специальных случайных блужданий с непрерывным временем, порожденных обобщенными дважды стохастическими пуассоновскими процессами, – там же приведены оценки скорости этой сходимости. В работе [62] результаты статьи [63] перенесены на обобщенные дисперсионные гамма-распределения.

Пусть $\{X_{n,j}\}_{j\geqslant 1}$, $n = 1, 2, \ldots,$ – семейство последовательностей одинаково распределенных в каждой последовательности (при каждом фиксированном n) случайных величин. Пусть $\{N_n\}_{n\geqslant 1}$ – последовательность целочисленных неотрицательных случайных величин таких, что при каждом $n \geqslant 1$ случайные величины $N_n, X_{n,1}, X_{n,2}, \ldots$ независимы. Положим

$$S_{n,k} = X_{n,1} + \ldots + X_{n,k}.$$

Для простоты обозначений будем полагать $\sum_{j=1}^{0} \ldots = 0$.

Расстояние Леви, которое, как известно, метризует слабую сходимость в пространстве функций распределения, будем обозначать $L(\,\cdot\,,\,\cdot\,)$,

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

Каждой паре функций распределения $(F,\,H)$ поставим в соответствие множество $\mathcal{M}(F|H)$, содержащее все функции распределения $Q(x)$ с $Q(0) = 0$, обеспечивающие представление характеристической функции, соответствующей функции распределения F, в виде степенно́й смеси характеристической функции, соответствующей функции распреде-

ления H:

$$\int_{-\infty}^{\infty} e^{itx} dF(x) = \int_{0}^{\infty} h^x(t) dQ(x), \quad t \in \mathbb{R},$$

где

$$h(t) = \int_{-\infty}^{\infty} e^{itx} dH(x), \quad t \in \mathbb{R}.$$

Везде далее сходимость будет подразумеваться при $n \to \infty$.

Лемма 1.1 [42]. *Предположим, что существуют последовательность натуральных чисел* $\{k_n\}_{n\geqslant 1}$ *и функция распределения* $H(x)$ *такие, что*

$$\mathsf{P}\big(S_{n,k_n} < x\big) \Longrightarrow H(x).$$

Предположим, что $N_n \to \infty$ *по вероятности. Для того чтобы имела место сходимость*

$$\mathsf{P}\big(S_{n,N_n} < x\big) \Longrightarrow F(x)$$

распределений случайных сумм к некоторой функции распределения $F(x)$*, необходимо и достаточно, чтобы существовала слабо компактная последовательность функций распределения* $\{Q_n^*(x)\}_{n\geqslant 1}$ *такая, что выполняются условия*

(i) $Q_n^*(x) \in \mathcal{M}(F|H),\ n = 1, 2, \ldots,$

(ii) $L(Q_n^*, Q_n) \longrightarrow 0,$

где $Q_n(x) = \mathsf{P}(N_n < k_n x)$, $x \in \mathbb{R}$.

Напомним определение идентифицируемости смесей распределений вероятностей, предложенное в работе [57]. Для целей данной статьи достаточно рассмотреть смеси распределений из однопараметрических семейств. Пусть функция $H(x; y)$ определена на плоскости $\mathbb{R} \times \mathbb{R}$. Предположим, что функция $H(x; y)$ измерима по y при каждом фиксированном $x \in \mathbb{R}$ и является функцией распределения как функция аргумента x при каждом фиксированном $y \in \mathbb{R}$. Пусть $\mathcal{Q}$ – некоторое семейство функций распределения. Обозначим

$$\mathcal{F} = \left\{ F(x) = \int_{-\infty}^{\infty} H(x; y)\, dQ(y),\ x \in \mathbb{R} :\ Q \in \mathcal{Q} \right\}. \tag{1.8}$$

Семейство $\mathcal{F}$ называется *идентифицируемым*, если из равенства

$$\int_{-\infty}^{\infty} H(x; y)\, dQ_1(y) = \int_{-\infty}^{\infty} H(x; y)\, dQ_2(y), \qquad x \in \mathbb{R},$$

с $Q_1 \in \mathcal{Q},\ Q_2 \in \mathcal{Q}$ вытекает, что $Q_1(y) \equiv Q_2(y)$.

Рассмотрим некоторые достаточные условия идентифицируемости смесей распределений из однопараметрических семейств. Хорошо известно, что в общем случае сдвиг-масштабные смеси нормальных законов не являются идентифицируемыми. Однако однопараметрические сдвиг-масштабные смеси нормальных законов типа (1.1) обладают этим свойством.

Семейство функций распределения $\{H(x;y) : y > 0\}$ называется *аддитивно замкнутым*, если для любых $y_1 > 0$, $y_2 > 0$ справедливо соотношение

$$H(x;y_1) * H(x;y_2) \equiv H(x;y_1+y_2). \tag{1.9}$$

Здесь символ $*$ обозначает свертку. Иногда свойство (1.9) семейств распределений вероятностей называется *воспроизводимостью* по параметру y.

Лемма 1.2 [57]. *Предположим, что множество $\mathcal{Q}$ состоит из всех функций распределения $Q(y)$ с $Q(0) = 0$. Пусть семейство функций распределения $\{H(x;y) : y > 0\}$ аддитивно замкнуто. Тогда семейство смесей* (1.8) *является идентифицируемым.*

Теорема 1.1 [63]. *Предположим, что существуют последовательность натуральных чисел $\{k_n\}_{n\geqslant 1}$ и число $\alpha \in \mathbb{R}$ такие, что*

$$\mathsf{P}\big(S_{n,k_n} < x\big) \Longrightarrow \Phi(x-\alpha). \tag{1.10}$$

Предположим, что $N_n \to \infty$ по вероятности. Для того чтобы имела место сходимость распределений случайных сумм к некоторой функции распределения $F(x)$:

$$\mathsf{P}\big(S_{n,N_n} < x\big) \Longrightarrow F(x), \tag{1.11}$$

необходимо и достаточно, чтобы существовала функция распределения $Q(x)$ такая, что $Q(0) = 0$,

$$F(x) = \int_0^\infty \Phi\Big(\frac{x-\alpha z}{\sqrt{z}}\Big) dQ(z), \tag{1.12}$$

и

$$\mathsf{P}(N_n < xk_n) \Longrightarrow Q(x). \tag{1.13}$$

Замечание 1.1. Условие (1.10) выполняется в следующей довольно общей ситуации. Предположим, что случайные величины $X_{n,j}$ имеют конечные дисперсии. Также предположим, что величины $X_{n,j}$ могут быть представлены в виде

$$X_{n,j} = X^*_{n,j} + \alpha_n,$$

где $\alpha_n \in \mathbb{R}$, а $X^*_{n,j}$ – случайная величина с $\mathsf{E}X^*_{n,j} = 0$, $\mathsf{D}X^*_{n,j} = \sigma_n^2 < \infty$, так что $\mathsf{E}X_{n,1} = \alpha_n$ и $\mathsf{D}X_{n,1} = \sigma_n^2$. Предположим, что $\alpha_n k_n \to a$ и $k_n\sigma_n^2 \to 1$ при $n \to \infty$. Тогда вследствие хорошо известного результата о необходимых и достаточных условиях сходимости к нормальному закону распределений сумм независимых случайных величин с конечными дисперсиями в схеме серий (см., например, [16]) можно заметить, что соотношение (1.10) имеет место тогда и только тогда, когда выполнено условие Линдеберга: для любого $\varepsilon > 0$

$$\lim_{n\to\infty} k_n \mathsf{E}(X^*_{n,1})^2 \mathbb{I}(|X^*_{n,1}| \geqslant \varepsilon) = 0$$

(здесь $\mathbb{I}(A)$ – индикаторная функция множества (события) A), то есть квадратичные хвосты распределений слагаемых должны убывать достаточно быстро.

Следствие 1.1 [62, 63]. *Предположим, что существуют последовательность натуральных чисел $\{k_n\}_{n\geqslant 1}$ и число $\alpha \in \mathbb{R}$ такие, что имеет место сходимость* (1.10). *Предположим, что $N_n \to \infty$ по вероятности.*
(*a*) *Для того чтобы имела место сходимость распределений случайных сумм к обобщенным дисперсионным гамма-распределениям*:

$$\mathsf{P}\big(S_{n,N_n} < x\big) \Longrightarrow P_{GVG}(x; \alpha, \sigma, \nu, \kappa, \delta), \tag{1.14}$$

необходимо и достаточно, чтобы

$$\mathsf{P}(N_n < xk_n) \Longrightarrow P_{GG}(x; \nu, \kappa, \delta), \tag{1.15}$$

где $P_{GG}(x; \nu, \kappa, \delta)$ – функция распределения обобщенного гамма-распределения, соответствующая плотности $p_{GG}(x; \nu, \kappa, \delta)$ (см. (1.6)).
(*b*) *Для того чтобы имела место сходимость распределений случайных сумм к обобщенным гиперболическим распределениям*:

$$\mathsf{P}\big(S_{n,N_n} < x\big) \Longrightarrow P_{GH}(x; \alpha, \beta, \nu, \mu, \lambda), \tag{1.16}$$

необходимо и достаточно, чтобы

$$\mathsf{P}(N_n < xk_n) \Longrightarrow P_{GIG}(x; \nu, \mu, \lambda), \tag{1.17}$$

где $P_{GIG}(x; \nu, \mu, \lambda)$ – функция распределения обобщенного обратного гауссовского распределения, соответствующая плотности $p_{GIG}(x; \nu, \mu, \lambda)$ (см. (1.4)).

Замечание 1.2. В соотношениях (1.10), (1.15) и (1.16) предельные функции распределения непрерывны. Поэтому в этих соотношениях сходимость по распределению эквивалентна равномерной сходимости функций распределения.

1.3 Сходимость неоднородных случайных блужданий, порожденных обобщенными процессами Кокса, к обобщенным дисперсионным гамма-процессам Леви

1.3.1 Предварительные замечания. Цель исследования

Результаты, приведенные в предыдущем разделе, можно интерпретировать как утверждения о сходимости одномерных распределений случайно остановленных случайных блужданий. В то же время, во многих работах, описывающих математические модели, скажем, процессов, протекающих на финансовых рынках, существенную роль играют процессы типа подчиненных винеровских процессов, приращения которых имеют распределения с более тяжелыми хвостами, нежели у нормального распределения (приращений классического процесса броуновского движения). По аналогии с результатами предыдущих разделов было бы целесообразно предложить достаточно простую предельную схему, основанную на элементарных случайных блужданиях, в рамках которой происходит формирование моделей типа подчиненных винеровских процессов. Данной теме посвящен текущий раздел.

Выше был описан класс специальных дисперсионно-сдвиговых смесей нормальных законов – класс обобщенных дисперсионных гамма-распределений (generalized variance-gamma distributions), который содержит распределения, хвосты которых, в частности, могут убывать экспоненциально-степенным (вейбулловским) образом. Как показано, например, в работах [6, 37], в некоторых случаях такие распределения оказываются более адекватными моделями реально наблюдаемых закономерностей, нежели, например, хорошо известные обобщенные гиперболические законы [30, 31], широко и успешно применяемые для описания статистических закономерностей поведения стохастических систем. В работе [46] приведены условия сходимости неоднородных случайных блужданий, порожденных обобщенными процессами Кокса, к предельным процессам Леви вида симметричных подчиненных винеровских процессов. В данном разделе эти результаты обобщаются на несимметричный случай и приводятся условия сходимости к процессам Леви с одномерными обобщенными дисперсионными гамма-распределениями, в частности, к подчиненным винеровским процессам с субординатором, являющимся процессом Леви–Вейбулла.

Слабая сходимость процессов в пространстве Скорохода в этом разделе будет обозначаться символом $\Longrightarrow$. Символы $\stackrel{d}{=}$ и $\stackrel{d}{\longrightarrow}$ соответственно будут обозначать совпадение распределений и сходимость по распределению, то есть поточечную сходимость функций распределения в точках непрерывности предельной функции распределения.

1.3.2 Сходимость обобщенных процессов Кокса к процессам Леви

Без существенного ограничения общности ниже будем рассматривать случайные процессы в пространстве Скорохода $\mathcal{D} = \mathcal{D}[0,1]$. Чтобы ввести разумную асимптотику, формализующую «неограниченный» рост интенсивности скачков и позволяющую построить аппроксимацию, соответствующую условиям «большой нагрузки», зафиксируем момент времени t и введем вспомогательный «бесконечно большой» целочисленный параметр n. Всюду далее, если не оговорено иное, будет предполагаться, что $n \to \infty$. Рассмотрим 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 $\Lambda_n(0) = 0$. Предположим, что 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

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

Здесь и далее для определенности считаем, что $\sum_{i=1}^{0} = 0$.

Из (1.18) и (1.19) несложно видеть, что $Q_n(t) = Z_n(\Lambda_n(t))$. Поскольку при каждом $n \geqslant 1$ $Z_n(t)$ и $\Lambda_n(t)$ являются независимыми процессами Леви, то суперпозиция $Q_n(t) = Z_n(\Lambda_n(t))$ также является процессом Леви. Поэтому $Q_n(t_2) - Q_n(t_1) \stackrel{d}{=} Q_n(t_2 - t_1)$ для любых $0 \leqslant t_1 < t_2 < \infty$ и любого $n \geqslant 1$.

Обозначим $a_n = \mathsf{E}X_{n,1}$ и предположим, что

$$0 < m_n^{\beta} \equiv \mathsf{E}|X_{n,1}|^{\beta} < \infty \tag{1.21}$$

при некотором $\beta \in [1,2]$.

Чтобы установить слабую сходимость случайных процессов $Q_n(t)$ в пространстве Скорохода $\mathcal{D}$, сначала необходимо найти предельное распределение с. в. $Q_n(t)$ при каждом $t > 0$.

Пусть $t = 1$. Обозначим $N_n = N_1^{(n)}(\Lambda_n(1))$. Предположим, что при каждом $k_n \in \mathbb{N}$ имеет место сходимость

$$\mathsf{P}(X_{n,1} + ... + X_{n,k_n} < x) \xrightarrow{d} H(x), \tag{1.22}$$

где $H(x)$ – некоторая безгранично делимая функция распределения.

Также предположим, что

$$\mathsf{P}\big(\Lambda_n(1) < k_n x\big) \xrightarrow{d} \mathsf{P}(U < x), \tag{1.23}$$

где U – неотрицательная случайная величина, распределение которой не сосредоточено в нуле. Так как $\Lambda_n(t)$ – процесс Леви, то случайная величина U безгранично делима.

Лемма 1.2 [42]. *Положим $N_n = N_1^{(n)}(\Lambda_n)$, $n \geqslant 1$, где $\{N_1^{(n)(t)},\ t \geqslant 0\}$, $n = 1, 2, \ldots$ – стандартные пуассоновские процессы, а Λ_n, $n = 1, 2, \ldots$ – положительные случайные величины, независимые при каждом $n \geqslant 1$. Тогда $\mathsf{P}(N_n < k_n x) \xrightarrow{d} A(x)$ для некоторой неограниченно возрастающей последовательности k_n вещественных чисел и функции распределения $A(x)$ в том и только том случае, когда $\mathsf{P}(\Lambda_n < k_n x) \xrightarrow{d} A(x)$.*

Из леммы 1.2 вытекает, что сходимость (1.23) эквивалентна тому, что

$$\mathsf{P}\big(N_n < k_n x\big) \xrightarrow{d} \mathsf{P}(U < x). \tag{1.24}$$

По теореме переноса Гнеденко–Фахима [17] условия (1.22) и (1.24) влекут сходимость

$$Q_n(1) = X_{n,1} + ... + X_{n,N_n} \xrightarrow{d} Q, \tag{1.25}$$

где Q – случайная величина с характеристической функцией

$$\mathfrak{f}(s) = \int_0^\infty \big(h(s)\big)^u \, d\mathsf{P}(U < u),$$

где $h(s)$ – характеристическая функция, соответствующая функции распределения $H(x)$. Заметим, что функция распределения $H(x)$ может не быть симметричной, т. е. может не удовлетворять условию $H(-x) = 1 - H(x)$ для всех $x \geqslant 0$.

Пусть Y – безгранично делимая случайная величина с функцией распределения $H(x)$. Так как обе случайные величины Y и U безгранично делимы, можно определить независимые процессы Леви $Y(t)$ и $U(t)$, $t \geqslant 0$, такие, что $Y(1) \stackrel{d}{=} Y$ и $U(1) \stackrel{d}{=} Y$. Тогда можно показать, что $\mathfrak{f}(s) = \mathsf{E}e^{isQ} = \mathsf{E}\exp\big\{isY(U(1))\big\}$, $s \in \mathbb{R}$, т. е. $Q \stackrel{d}{=} Y(U(1))$. Несложно убедиться, что случайная величина Q безгранично делима, так что можно определить процесс Леви $Q(t)$, $t \geqslant 0$, такой, что $Q(1) \stackrel{d}{=} Q$, причем $Q(t) \stackrel{d}{=} Y(U(t))$.

Поскольку в соответствии с (1.25)

$$Q_n(1) = \sum_{i=1}^{N_n} X_{n,i} \Longrightarrow Q(1),$$

а $Q_n(t)$ и $Q(t)$ являются процессами Леви, используя (1.24), мы можем заключить, что для любого $t > 0$

$$Q_n(t) = \sum\nolimits_{i=1}^{N_{n,1}(\Lambda_n(t))} X_{n,i} \xrightarrow{d} Q(t). \tag{1.26}$$

Более того, почти все траектории процессов $Q_n(t)$ и $Q(t)$ принадлежат пространству Скорохода $\mathcal{D}$.

В дополнение к условиям (1.20) и (1.21) потребуем, чтобы

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

Точно так же, как это сделано в работах [46] и [48] можно показать, что условия (1.26) и (1.27) гарантируют выполнение условий теоремы 15.6 в [13], из которой в свою очередь, вытекает

Теорема 1.2 [46]. *Пусть обобщенные процессы Кокса $Q_n(t)$ (см. (1.18)) управляются неубывающими положительными процессами Леви $\Lambda_n(t)$, удовлетворяющими условиям* (1.20) *и* (1.23) *с некоторыми $\delta, \delta_1 \in (0, 1]$ и $k_n \in \mathbb{N}$. Предположим, что случайные величины $\{X_{n,j}\}_{j \geqslant 1}$, $n = 1, 2, ...$, удовлетворяют условиям* (1.22) *с теми же самыми k_n и* (1.21) *с некоторым $\beta \in [1, 2]$. Также предположим, что выполнено условие* (1.27). *Тогда обобщенные процессы Кокса $Q_n(t)$ слабо сходятся в пространстве Скорохода $\mathcal{D}$ к процессу Леви $Q(t)$ такому, что*

$$\mathsf{E} \exp\{isQ(1)\} = \int_0^{\infty} \big(h(s)\big)^u \, d\mathsf{P}(U < u), \quad s \in \mathbb{R},$$

где $h(s)$ – характеристическая функция, соответствующая функции распределения $H(x)$ в (1.22).

Вообще говоря, в теореме 1.2 речь идет о хорошо изученной сходимости семимартингалов со стационарными приращениями, см., например, [44]. Однако специальная структура рассматриваемых здесь процессов типа суперпозиции позволяет ослабить некоторые условия, налагаемые в общем случае. В частности, в следствии VII.3.6 в [44] требуется, чтобы $\delta = \delta_1 = 1$.

Некоторые следствия теоремы 1.2 приведены в работах [46] и [48], где соответственно приведены условия сходимости обобщенных процессов Кокса к симметричным устойчивым процессам Леви и обобщенным гиперболическим процессам Леви.

1.3.3 Сходимость обобщенных процессов Кокса к подчиненным винеровским процессам

Обозначим $\sigma_n^2 = \mathsf{D}X_{n,1}$. Из классической теории предельных теорем хорошо известно, что, если условия

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

выполнены для некоторых $a \in \mathbb{R}$, $0 < \sigma^2 < \infty$ и любого $\epsilon > 0$, то сходимость (1.22) имеет место с $H(x) \equiv \Phi\big(\sigma^{-1}(x-a)\big)$. В таком случае функция распределения $F(x)$ предельной случайной величины $Q(1)$ в теореме 1.2 является дисперсионно-сдвиговой смесью нормальных законов. Как уже отмечалось выше, такие смеси являются идентифицируемыми [62, 63]. Из результатов этих работ и теоремы 1.2 вытекает следующий результат.

Теорема 1.3 [48]. *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* (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 некоторому процессу Леви $Q(t)$ в том и только том случае, когда существует неотрицательная случайная величина U такая, что*

$$\mathsf{P}\big(Q(1) < x\big) = \int_0^\infty \Phi\Big(\frac{x-au}{\sigma\sqrt{u}}\Big)d\mathsf{P}(U < u), \quad x \in \mathbb{R},$$

и условие (1.23) *выполняется с теми же самыми k_n.*

1.3.4 Сходимость обобщенных процессов Кокса к процессам Леви с одномерными обобщенными дисперсионными гамма-распределениями

Класс дисперсионно-сдвиговых смесей нормальных законов вида очень богат. Из его представителей наибольшей популярностью в прикладных исследованиях пользуются обобщенные гиперболические законы, в которых смешивающая с. в. U имеет обобщенное обратное гауссово распределение. Однако, как уже отмечалось, в некоторых случаях бóльшую адекватность демонстрируют такие смеси, в которых с. в. U имеет обобщенное гамма-распределение. В частности, в главе 2 приведены примеры таких данных.

Широкая применимость GG-распределений обусловлена возможностью их использования в качестве адекватных асимптотических аппроксимаций, поскольку практически все они выступают в качестве предельных в различных предельных теоремах теории вероятностей. При некоторых комбинациях параметров GG-распределения являются безгранично делимыми.

Из результатов работы [62] и леммы 1.2 вытекает, что сходимость конечномерных распределений обобщенных процессов Кокса к обобщенным дисперсионным гамма-распределениям имеет место тогда и только тогда, когда выполнено условие (1.23), в котором случайная величина U имеет GG-распределение. Если же говорить о слабой сходимости обобщенных процессов Кокса в пространстве Скорохода, то для непосредственного применения теоремы 1.2 к выводу условий такой сходимости необходимо, чтобы случайная величина U была еще и безгранично делимой.

Проиллюстрируем сказанное на примере условий сходимости неоднородных случайных блужданий, порожденных обобщенными процессами Кокса, к подчиненному винеровскому процессу с субординатором, являющимся процессом Леви–Вейбулла. Напомним, что в наших обозначениях распределение Вейбулла–Гнеденко задается GG-плотностью $p_{GG}(x;\nu,1,\delta)$, $x>0$.

Лемма 1.3. *Если $\nu \leqslant 1$, то распределение Вейбулла–Гнеденко безгранично делимо.*

Доказательство. Как показано в работе [26], распределение Вейбулла–Гнеденко с $\nu \leqslant 1$ может быть представлено в виде смешанного показательного распределения. Но в работе [43] показано, что все смешанные показательные законы безгранично делимы. Лемма доказана.

Согласно лемме 1.3, если $\nu \leqslant 1$, то распределение Вейбулла–Гнеденко безгранично делимо, и можно определить процесс Леви $U(t)$, $t \geqslant 0$, так, чтобы $\mathsf{P}\big(U(1)<x\big) = F(x;\nu,1,\delta)$, $x\in\mathbb{R}$. Такой процесс будем называть процессом Леви–Вейбулла. Как показано в работе [12], на качественном уровне асимптотическое поведение хвостов смесей нормальных законов совпадает с аналогичным поведением хвостов смешивающих законов. Поэтому хвосты конечномерных распределений подчиненного винеровского процесса с субординатором, являющимся процессом Леви–Вейбулла, убывают экспоненциально-степенным образом (см. раздел 1.2.3). При этом случай малых значений параметра $\nu\in(0,1]$ представляет особый интерес, поскольку распределения Вейбулла–Гнеденко с такими параметрами занимают промежуточное место между распределениями с экспоненциальным убыванием хвостов (показательное распределение, гамма-распределение) и «тяжелохвостыми» распределениями со степенным убыванием хвостов типа Ципфа–Парето.

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 (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

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

с теми же самыми k_n.

1.4 Сходимость распределений статистик, построенных по выборкам случайного объема, к многомерным обобщенным дисперсионным гамма-распределениям

1.4.1 Введение. Обозначения и вспомогательные результаты

Как было сказано во введении, в работе [47] приведены критерии сходимости распределений статистик, построенных по выборкам случайного объема, к многомерным обобщенным гиперболическим законам. В этой главе показано, что 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.

Как было отмечено ранее, 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.

В Главе 2 приведен пример одного из многих случаев, когда обобщенные гамма-распределения заметно лучше согласуются с эмпирическими распределениями (демонстрируют гораздо более высокие P-значения при проверке критирия Хи-квадрат, см. раздел 2.6.3), нежели обобщенные обратные гауссовские распределения. Поэтому поиск критериев сходимости к многомерным дисперсионным гамма-распределениям представляет собой весьма перспективную задачу.

Пусть $\{\boldsymbol{S}_{n,k} = (S_{n,k}^{(1)},\dots,S_{n,k}^{(m)})^\top\}$, $n,k \in \mathbb{N}$, – последовательность серий $\mathbb{R}^m$-значных случайных векторов. Для $n,k \in \mathbb{N}$ пусть $\boldsymbol{a}_{n,k} = (a_{n,k}^{(1)},\dots,a_{n,k}^{(m)})^\top \in \mathbb{R}^m$ – 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 $\boldsymbol{c}_n = (c_n^{(1)},\dots,c_n^{(m)})^\top \in \mathbb{R}^m$ – 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 $\boldsymbol{W}_n = (U_n,\, \boldsymbol{V}_n^\top)^\top = (U_n,\, V_n^{(1)}, \ldots, V_n^{(m)})^\top$.

Для расстояния полной вариации между распределениями случайных векторов $\boldsymbol{X}$ и $\boldsymbol{Y}$ будем использовать обозначение $\Delta(\boldsymbol{X},\, \boldsymbol{Y})$:

$$\Delta(\boldsymbol{X},\, \boldsymbol{Y}) = \int\limits_{\mathbb{R}^m} |\mathcal{L}(\boldsymbol{X}) - \mathcal{L}(\,\boldsymbol{Y})|(d\boldsymbol{x}).$$

1.4.2 Условия сходимости распределений многомерных случайных последовательностей с независимыми случайными индексами

Чтобы получить нетривиальные предельные распределения для $\boldsymbol{Z}_n$, в схеме серий, рассматриваемой в данном разделе, нам потребуется следующее дополнительное *условие согласованности*: для любого $T \in (0, \infty)$

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

Несложно убедиться, что условие согласованности (1.29) вытекает из условия

$$\lim_{n\to\infty} \mathsf{E}\Delta(\,\boldsymbol{Y}_{n,N_n},\, \boldsymbol{Y}) = 0.$$

Действительно,

$$\mathsf{E} \sup_{\|\boldsymbol{t}\|\leqslant T} |h_{n,N_n}(\boldsymbol{t}) - h(\boldsymbol{t})| = \sum_{k=1}^{\infty} \mathsf{P}(N_n = k) \sup_{\|\boldsymbol{t}\|\leqslant T} |h_{n,k}(\boldsymbol{t}) - h(\boldsymbol{t})| =$$

$$= \sum_{k=1}^{\infty} \mathsf{P}(N_n = k) \sup_{\|\boldsymbol{t}\|\leqslant T} \left| \int\limits_{\mathbb{R}^m} e^{i\langle \boldsymbol{t},\boldsymbol{x}\rangle} \mathcal{L}(\,\boldsymbol{Y}_{n,k})(d\boldsymbol{x}) - \int\limits_{\mathbb{R}^m} e^{i\langle \boldsymbol{t},\boldsymbol{x}\rangle} \mathcal{L}(\,\boldsymbol{Y})(d\boldsymbol{x}) \right| =$$

$$= \sum_{k=1}^{\infty} \mathsf{P}(N_n = k) \sup_{\|\boldsymbol{t}\|\leqslant T} \left| \int\limits_{\mathbb{R}^m} e^{i\langle \boldsymbol{t},\boldsymbol{x}\rangle} \big[\mathcal{L}(\,\boldsymbol{Y}_{n,k}) - \mathcal{L}(\,\boldsymbol{Y})\big](d\boldsymbol{x}) \right| \leqslant$$

$$\leqslant \sum_{k=1}^{\infty} \mathsf{P}(N_n = k) \int\limits_{\mathbb{R}^m} \big|\mathcal{L}(\boldsymbol{Y}_{n,k}) - \mathcal{L}(\boldsymbol{Y})\big|(d\boldsymbol{x}) = \mathsf{E}\Delta(\boldsymbol{Y}_{n,N_n}, \boldsymbol{Y}).$$

Напомним, распределение $\mathbb{R}^m$-значного случайного вектора $\boldsymbol{X}$ является многомерной дисперсионно-сдвиговой смесью нормальных законов, если $\boldsymbol{X} \stackrel{d}{=} \boldsymbol{a} + U\boldsymbol{b} + \sqrt{U}A\boldsymbol{Y}$, где $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^m$, A – вещественная $(m \times m)$-матрица такая, что матрица $\Sigma \equiv AA^\top$ 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 $\boldsymbol{Y}$. Эквивалентно, будем говорить, что вероятностная мера F на $(\mathbb{R}^m, \mathfrak{B}_m)$ является многомерной дисперсионно-сдвиговой смесью нормальных законов, если

$$F(d\boldsymbol{x}) = \int_0^\infty \mathcal{N}_{\boldsymbol{b}+z\boldsymbol{a},\, z\Sigma}(d\boldsymbol{x})G(dz), \tag{1.30}$$

где смешивающим распределением G является вероятностная мера на $(\mathbb{R}_+, \mathfrak{B}_+)$. Как и выше, этот факт иногда будем записывать в виде $F = \mathcal{N}_{\boldsymbol{b}+z\boldsymbol{a},\, z\Sigma} \circ G$.

Следуя работе [47], для фиксированных случайных векторов $\boldsymbol{Z}$ и $\boldsymbol{Y}$ с характеристическими функциями $f(\boldsymbol{t})$ и $h(\boldsymbol{t})$ введем множество $\mathcal{W}(\boldsymbol{Z}|\,\boldsymbol{Y})$, содержащее все $(m+1)$-мерные случайные векторы $\boldsymbol{W} = (U,\ \boldsymbol{V}^\top)^\top$ с $U \in \mathbb{R}$ и $\boldsymbol{V} \in \mathbb{R}^m$, такие, что характеристическая функция $f(\boldsymbol{t})$ может быть представлена в виде

$$f(\boldsymbol{t}) = \mathsf{E}h(U\boldsymbol{t})e^{i\langle \boldsymbol{t},\, \boldsymbol{V}\rangle}, \quad \boldsymbol{t} \in \mathbb{R}^m, \tag{1.31}$$

и $\mathsf{P}(U \geqslant 0) = 1$. Какими бы ни были случайные векторы $\boldsymbol{Z}$ и $\boldsymbol{Y}$, множество $\mathcal{W}(\boldsymbol{Z}|\,\boldsymbol{Y})$ всегда не пусто, поскольку оно очевидно содержит вектор $(0, \boldsymbol{Z}^\top)^\top$. Множество $\mathcal{W}(\boldsymbol{Z}|\,\boldsymbol{Y})$ может содержать более одного элемента. Несложно видеть, что представление (1.31) эквивалентно тому, что $\boldsymbol{Z} \stackrel{d}{=} U\,\boldsymbol{Y} + \boldsymbol{V}$.

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 (см., например, [13, 58]).

Теорема 1.5 ([47]). *Пусть семейство случайных величин $\{U_n\}_{n\in\mathbb{N}}$ слабо относительно компактно. Предположим, что выполнено условие согласованности* (1.29)*. Тогда слабая 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 $\boldsymbol{Z}$ имеет место при некоторых $\boldsymbol{c}_n \in \mathbb{R}^m$ тогда и только тогда, 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

$$\lim_{n\to\infty} \Lambda(\boldsymbol{W}_n^*,\ \boldsymbol{W}_n) = 0. \tag{1.33}$$

1.4.3 Общая теорема о сходимости распределений сумм случайного числа независимых неодинаково распределенных многомерных случайных величин

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

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

Как будет показано в данном разделе, если $\boldsymbol{S}_{n,k}$ – сумма независимых случайных величин, то условие слабой относительной компактности последовательности $\{U_n\}_{n\in\mathbb{N}}$, использованное выше для описания общего случая, может быть заменено, вообще говоря, менее ограничительным условием слабой относительной компактности семейства $\{\boldsymbol{Y}_{n,k}\}_{n,k\in\mathbb{N}}$. Действительно, пусть, к примеру, нормы случайных векторов $\boldsymbol{S}_{n,k}$ имеют моменты некоторого порядка $\delta>0$. Тогда, если положить $b_{n,k}=(\mathsf{E}\|\boldsymbol{S}_{n,k}-\boldsymbol{a}_{n,k}\|^{\delta})^{1/\delta}$, то по неравенству Маркова

$$\lim_{R\to\infty}\sup_{n,k\in\mathbb{N}}\mathsf{P}\big(\|\boldsymbol{Y}_{n,k}\|>R\big)\leqslant\lim_{R\to\infty}\frac{1}{R^{\,\delta}}=0,$$

то есть семейство $\{\boldsymbol{Y}_{n,k}\}_{n,k\in\mathbb{N}}$ слабо относительно компактно.

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*

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

Доказательство. Достаточно показать, что в рассматриваемой ситуации условие (18) влечет слабую относительную компактность семейства $\{U_n\}_{n\in\mathbb{N}}$ и сослаться на приведенную выше теорему 1. В дальнейшем симметризацию случайного вектора $\boldsymbol{X}$ будем обозначать $\boldsymbol{X}^{(s)}$, $\boldsymbol{X}^{(s)}=\boldsymbol{X}-\boldsymbol{X}'$, где $\boldsymbol{X}'$ – случайный вектор, независимый от $\boldsymbol{X}$ и такой, что $\boldsymbol{X}'\stackrel{d}{=}\boldsymbol{X}$. Для $q\in(0,1)$ пусть $\ell_n(q)$ – точная нижняя грань q-квантилей случайной величины N_n, $n\in\mathbb{N}$. Предположим, что при каждом n случайные элементы $N_n, X_{n,1}^{(s)}, X_{n,2}^{(s)},...$ независимы в совокупности и введем случайные векторы

$$\boldsymbol{Q}_n=\frac{1}{d_n}\sum_{j=1}^{N_n}\boldsymbol{X}_{n,j}^{(s)},\quad n\in\mathbb{N}.$$

Используя элементарно доказываемое неравенство симметризации

$$\mathsf{P}\big(\boldsymbol{X}^{(s)} \geqslant R\big) \leqslant 2\,\mathsf{P}(\|\boldsymbol{X}-\boldsymbol{a}\| \geqslant \tfrac{R}{2}),$$

справедливое для любого случайного вектора $\boldsymbol{X}$ и любых $\boldsymbol{a} \in \mathbb{R}^m$, $R>0$ (см., например, [27]), для любых $R>0$ и $n \in \mathbb{N}$ получаем

$$\mathsf{P}(\|\boldsymbol{Q}_n\| \geqslant R) = \sum_{k=1}^{\infty} \mathsf{P}(N_n=k)\mathsf{P}\biggl(\biggl\|\frac{1}{d_n}\sum_{j=1}^{k}\boldsymbol{X}_{n,j}^{(s)}\biggr\| \geqslant R\biggr) \leqslant$$

$$\leqslant 2\sum_{k=1}^{\infty} \mathsf{P}(N_n=k)\mathsf{P}\biggl(\biggl\|\frac{1}{d_n}\Bigl(\sum_{j=1}^{k}\boldsymbol{X}_{n,j}-\boldsymbol{c}_n\Bigr)\biggr\| \geqslant \frac{R}{2}\biggr) = 2\,\mathsf{P}\biggl(\biggl\|\frac{1}{d_n}\Bigl(\sum_{j=1}^{N_n}\boldsymbol{X}_{n,j}-\boldsymbol{c}_n\Bigr)\biggr\| \geqslant \frac{R}{2}\biggr)$$

Поэтому

$$\lim_{R\to\infty}\sup_n \mathsf{P}(\|\boldsymbol{Q}_n\| \geqslant R) \leqslant 2\lim_{R\to\infty}\sup_n \mathsf{P}\biggl(\biggl|\frac{1}{d_n}\Bigl(\sum_{j=1}^{N_n}\boldsymbol{X}_{n,j}-\boldsymbol{c}_n\Bigr)\biggr\| \geqslant \frac{R}{2}\biggr) = 0$$

в силу (1.32). Следовательно, последовательность случайных векторов $\{\boldsymbol{Q}_n\}_{n\in\mathbb{N}}$ слабо относительно компактна.

Теперь докажем, что

$$C(q) \equiv \sup_n \frac{b_{n,\ell_n(q)}}{d_n} < \infty \tag{1.36}$$

для каждого $q \in (0,1)$. С этой целью используем многомерный аналог неравенства Леви

$$\mathsf{P}\biggl(\max_{1\leqslant m\leqslant k}\biggl\|\sum_{j=1}^{m}\boldsymbol{X}_j^{(s)}\biggr\| \geqslant R\biggr) \leqslant 2\,\mathsf{P}\biggl(\biggl\|\sum_{j=1}^{k}\boldsymbol{X}_j^{(s)}\biggr\| \geqslant R\biggr),$$

справедливого для любых независимых случайных векторов $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_k$ и любого $R>0$ (см., например, [15], с. 455), с помощью которого для произвольного $q \in (0,1)$ получим следующую цепочку неравенств:

$$2\,\mathsf{P}(\|\boldsymbol{Q}_n\| \geqslant R) = 2\sum_{k=1}^{\infty} \mathsf{P}(N_n=k)\mathsf{P}\biggl(\biggl\|\frac{1}{d_n}\sum_{j=1}^{k}\boldsymbol{X}_{n,j}^{(s)}\biggr\| \geqslant R\biggr) \geqslant$$

$$\geqslant 2\sum_{k\geqslant\ell_n(q)} \mathsf{P}(N_n=k)\mathsf{P}\biggl(\biggl\|\frac{1}{d_n}\sum_{j=1}^{k}\boldsymbol{X}_{n,j}^{(s)}\biggr\| \geqslant R\biggr) \geqslant \sum_{k\geqslant\ell_n(q)} \mathsf{P}(N_n=k)\mathsf{P}\biggl(\biggl\|\frac{1}{d_n}\sum_{j=1}^{\ell_n(q)}\boldsymbol{X}_{n,j}^{(s)}\biggr\| \geqslant R\biggr) =$$

$$= \mathsf{P}\big(N_n \geqslant \ell_n(q)\big)\mathsf{P}\biggl(\biggl\|\frac{1}{d_n}\sum_{j=1}^{\ell_n(q)}\boldsymbol{X}_{n,j}^{(s)}\biggr\| \geqslant R\biggr) = (1-q)\mathsf{P}\biggl(\biggl|\frac{1}{d_n}\sum_{j=1}^{\ell_n(q)}\boldsymbol{X}_{n,j}^{(s)}\biggr\| \geqslant R\biggr).$$

Таким образом, слабая компактность семейства случайных векторов $\{\boldsymbol{Q}_n\}_{n\in\mathbb{N}}$, установленная выше, при каждом $q \in (0,1)$ влечет слабую компактность семейства $\{\boldsymbol{Q}_n^{(q)}\}_{n\in\mathbb{N}}$,

где

$$\boldsymbol{Q}_n^{(q)} = \frac{1}{d_n}\sum_{j=1}^{\ell_n(q)} \boldsymbol{X}_{n,j}^{(s)}, \quad n \in \mathbb{N}.$$

Предположим, что (1.36) не имеет места. В таком случае существуют некоторое $q^* \in (0,1)$ и некоторая последовательность $\mathcal{N}$ натуральных чисел такие, что

$$\frac{b_{n,\ell_n(q^*)}}{d_n} \longrightarrow \infty, \quad n \to \infty,\ n \in \mathcal{N}. \tag{1.37}$$

В соответствии с условиями теоремы, семейство случайных векторов $\big\{\boldsymbol{Y}_{n,k} = (\boldsymbol{S}_{n,k} - \boldsymbol{a}_{n,k})/b_{n,k}\big\}_{n,k\in\mathbb{N}}$ слабо относительно компактно. Поэтому можно выбрать подпоследовательность $\mathcal{N}_1 \subseteq \mathcal{N}$ так, чтобы

$$\boldsymbol{Y}_{n,\ell_n(q^*)} = \frac{1}{b_{n,\ell_n(q^*)}}\Bigg(\sum_{j=1}^{\ell_n(q^*)} \boldsymbol{X}_{n,j} - \boldsymbol{a}_{n,\ell_n(q^*)}\Bigg) \Longrightarrow \boldsymbol{Y}, \quad n \to \infty,\ n \in \mathcal{N}_1, \tag{1.38}$$

где $\boldsymbol{Y}$ – некоторый случайный вектор. Обозначим r-е компоненты случайных векторов $\boldsymbol{Q}_n^{(q^*)}$, $\boldsymbol{Y}_{n,\ell_n(q^*)}^{(s)}$ и $\boldsymbol{Y}^{(s)}$ через $Q_{n;r}^{(q^*)}$, $Y_{n,\ell_n(q^*);r}^{(s)}$ и $Y_r^{(s)}$ соответственно:

$$\boldsymbol{Q}_n^{(q^*)} = \big(Q_{n;1}^{(q^*)}, ..., Q_{n;m}^{(q^*)}\big)^\top, \quad \boldsymbol{Y}_{n,\ell_n(q^*)}^{(s)} = \big(Y_{n,\ell_n(q^*);1}^{(s)}, ..., Y_{n,\ell_n(q^*);m}^{(s)}\big)^\top, \quad \boldsymbol{Y}^{(s)} = \big(Y_1^{(s)}, ..., Y_m^{(s)}\big)^\top.$$

Тогда из (1.37) и (1.38) вытекает что для любого $R \in \mathbb{R}$ и любого $r \in \{1, ..., m\}$

$$\mathsf{P}(Q_{n;r}^{(q^*)} \leqslant R) = \mathsf{P}\Bigg(Y_{n,\ell_n(q^*);r}^{(s)} \leqslant \frac{d_n R}{b_{n,\ell_n(q^*)}}\Bigg) \longrightarrow \mathsf{P}(Y_r^{(s)} \leqslant 0) \geqslant \tfrac{1}{2}, \quad n \to \infty,\ n \in \mathcal{N}_1,$$

что противоречит слабой относительной компактности семейства $\{\boldsymbol{Q}_n^{(q^*)}\}$, установленной выше. Таким образом, условие (1.36) выполнено для любого $q \in (0,1)$.

Несложно убедиться, что $N_n \stackrel{d}{=} \ell_n(\upsilon)$, где υ – случайная величина с равномерным распределением на $(0,1)$. Поэтому с учетом (1.36) для любых $R \geqslant 0$ и $n \in \mathbb{N}$ имеем

$$\mathsf{P}(U_n \geqslant R) = \mathsf{P}\Big(\frac{b_{n,\ell_n(\upsilon)}}{d_n} \geqslant R\Big) = \int\limits_0^1 \mathbb{I}\Big(\frac{b_{n,\ell_n(q)}}{d_n} \geqslant R\Big)dq \leqslant$$

$$\leqslant \int\limits_0^1 \mathbb{I}\big(C(q) \geqslant R\big)dq = \mathsf{P}\big(C(\upsilon) \geqslant R\big),$$

так что

$$\lim_{R\to\infty} \sup_n \mathsf{P}\big(U_n \geqslant R\big) = \lim_{R\to\infty} \mathsf{P}\big(C(\upsilon) \geqslant R\big) = 0,$$

то есть последовательность $\{U_n\}_{n\in\mathbb{N}}$ слабо компактна. Тем самым теорема доказана.

1.4.4 Сходимость распределений сумм случайного числа независимых неодинаково распределенных многомерных случайных величин к дисперсионно-сдвиговым смесям, в частности, к многомерным GVG-распределениям

Изучим вопрос о том, при каких условиях предельными распределениями для сумм случайного числа независимых неодинаково распределенных многомерных случайных величин могут быть дисперсионно-сдвиговые смеси нормальных законов, и в частности, многомерные GVG-распределения.

Напомним, что характеристическая функция m-мерного нормального распределения с нулевым вектором математических ожиданий и ковариационной матрицей Σ имеет вид

$$\mathfrak{f}_{\mathbf{0},\Sigma}(\boldsymbol{t}) = \exp\{-\frac{1}{2}\boldsymbol{t}^\top \Sigma \boldsymbol{t}\}, \quad \boldsymbol{t} \in \mathbb{R}^m.$$

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 центрирующие векторы $\boldsymbol{a}_{n,k}$ и $\boldsymbol{c}_n$ в определенном смысле пропорциональны нормирующим константам $b_{n,k}$ и d_n. А именно, 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

$$\mathsf{P}(U_n^2 < x) \Longrightarrow G(x) \quad (n \to \infty). \tag{1.42}$$

Доказательство. Выведем теорему 1.7 в качестве следствия теоремы 1.6. Во-первых, заметим, что условие (1.39) – это не что иное как условие согласованности (1.29) с $h(\boldsymbol{t}) \equiv \exp\{-\frac{1}{2}\boldsymbol{t}^\top \Sigma \boldsymbol{t}\}$.

Во-вторых, заметим, что очевидным образом каждая одномерная проекция многомерной дисперсионно-сдвиговой смеси нормальных законов является одномерной дисперсионно-сдвиговой смесью нормальных законов. Не так давно в работе [63] было показано, что одномерные дисперсионно-сдвиговые смеси нормальных законов идентифицируемы, то есть, если $a \in \mathbb{R}$, $\sigma > 0$, $\mathsf{P}(Y < x) \equiv \Phi(x)$, а U_1 и U_2 – две неотрицательные случайные величины, то тождество

$$\mathsf{E}\Phi\Big(\frac{x - aU_1^2}{\sigma U_1}\Big) \equiv \mathsf{E}\Phi\Big(\frac{x - aU_2^2}{\sigma U_2}\Big)$$

влечет, что $U_1 \stackrel{d}{=} U_2$. Следовательно, множество $\mathcal{W}(\boldsymbol{Z}|\, \boldsymbol{Y})$ содержит не более одного вектора вида $\boldsymbol{W} = (\sigma U,\, (U^2\boldsymbol{a} + \boldsymbol{b})^\top)^\top$. Это означает, что в рассматриваемом случае условие (1.35) сводится к (1.40). Теорема доказана.

В качестве элементарного следствия теоремы 1.7 получаем утверждение, описывающее условия сходимости распределений случайных сумм к дисперсионным гамма-распределениям или обобщенным гиперболическим 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*

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

1.4.5 Предельная теорема для статистик, построенных по выборкам случайного объема

Пусть $\{\boldsymbol{X}_{n,j}\}_{j\geqslant 1}$, $n\in\mathbb{N}$, – последовательность серий независимых в каждой серии (но необязательно одинаково распределенных) случайных векторов со значениями в $\mathbb{R}^r$, $r\in\mathbb{N}$. Для $n,k\in\mathbb{N}$ пусть $\boldsymbol{T}_{n,k}=\boldsymbol{T}_{n,k}(\boldsymbol{X}_{n,1},...,\boldsymbol{X}_{n,k})$ – статистика, то есть измеримая функция аргументов $\boldsymbol{X}_{n,1},...,\boldsymbol{X}_{n,k}$, принимающая значения в $\mathbb{R}^m$. Для каждого $n\geq 1$ определим случайный вектор $\boldsymbol{T}_{n,N_n}$, полагая $\boldsymbol{T}_{n,N_n}(\omega)\equiv\boldsymbol{T}_{n,N_n(\omega)}(\boldsymbol{X}_{n,1}(\omega),...,\boldsymbol{X}_{n,N_n(\omega)}(\omega))$, $\omega\in\Omega$.

Пусть θ_n – $\mathbb{R}^m$-значные векторы, $n\in\mathbb{N}$. В этом разделе будем считать, что случайные векторы $\boldsymbol{S}_{n,k}$ имеют вид $\boldsymbol{S}_{n,k}=\boldsymbol{T}_{n,k}-\theta_n$, $n,k\in\mathbb{N}$. Относительно нормирующих констант и векторов будем предполагать, что существуют m-мерные векторы $\boldsymbol{a}$, $\boldsymbol{a}_n$, $\boldsymbol{b}$, $\boldsymbol{b}_n$ и положительные числа σ_n такие, что

$$\boldsymbol{a}_n\to\boldsymbol{a},\quad \boldsymbol{b}_n\to\boldsymbol{b}\quad (n\to\infty) \tag{1.43}$$

и для всех $n,k\in\mathbb{N}$

$$b_{n,k}=(\sigma_n\sqrt{k})^{-1},\ d_n=(\sigma_n\sqrt{n})^{-1},\ \boldsymbol{a}_{n,k}=(\sigma_n k)^{-1}\sqrt{n}\,\boldsymbol{a}_n,\ \boldsymbol{c}_n=(\sigma_n\sqrt{n})^{-1}\boldsymbol{b}_n, \tag{1.44}$$

так что

$$\boldsymbol{Y}_{n,k}=\sigma_n\sqrt{k}(\boldsymbol{T}_{n,k}-\theta_n)-\sqrt{n/k}\,\boldsymbol{a}_n\ \text{ и }\ \boldsymbol{Z}_n=\sigma_n\sqrt{n}(\boldsymbol{T}_{n,N_n}-\theta_n)-\boldsymbol{b}_n.$$

При этом $\sigma_n^2\boldsymbol{I}$ можно считать асимптотической ковариацией $\boldsymbol{T}_{n,k}$ при $k\to\infty$, тогда как $\sqrt{n}(k\sigma_n)^{-1}\boldsymbol{a}_n$ можно считать смещением $\boldsymbol{T}_{n,k}$.

Как известно, характеристическая функция нормального распределения в $\mathbb{R}^m$ с нулевым вектором математических ожиданий и ковариационной матрицей Σ имеет вид $\varphi(\boldsymbol{t})=\exp\{-\frac{1}{2}\boldsymbol{t}^\top\Sigma\boldsymbol{t}\}$, $\boldsymbol{t}\in\mathbb{R}^m$. В дальнейшем будем предполагать, что статистика $\boldsymbol{T}_{n,k}$ асимптотически нормальна в том смысле, что существует положительно определенная симметричная матрица Σ такая, что для любого $T\in(0,\infty)$

$$\lim_{n\to\infty}\mathsf{E}\sup_{\|\boldsymbol{t}\|\leqslant T}\left|h_{n,N_n}(\boldsymbol{t})-\exp\{-\tfrac{1}{2}\boldsymbol{t}^\top\Sigma\boldsymbol{t}\}\right|=0, \tag{1.45}$$

где $h_{n,k}(\boldsymbol{t})$ – характеристическая функция случайного вектора $\boldsymbol{Y}_{n,k}$. Следующее утверждение фактически является частным случаем теоремы 1.5.

Теорема 1.9 ([47]). *Пусть семейство случайных величин $\{n/N_n\}_{n\in\mathbb{N}}$ слабо относительно компактно, нормирующие константы имеют вид* (1.44) *и удовлетворяют условию* (1.43). *Предположим, что статистика $\boldsymbol{T}_{n,k}$ асимптотически нормальна в смысле* (1.45). *Тогда сходимость*

$$\sigma_n\sqrt{n}(\boldsymbol{T}_{n,N_n}-\theta_n)-\boldsymbol{b}_n\Longrightarrow Z$$

неслучайно нормированных и центрированных статистик, построенных по выборке случайного объема, к некоторому случайному вектору $\mathbf{Z}$ *при* $n \to \infty$ *имеет место 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*

$$\mathsf{P}(n/N_n < x) \Longrightarrow G(x)$$

при $n \to \infty$.

Теперь рассмотрим условия сходимости распределений статистик, построенных по выборкам случайного объема, к многомерным обобщенным дисперсионным гамма-распределениям. Напомним определение последних.

Пусть $\boldsymbol{a}$ и $\boldsymbol{b}$ – m-мерные векторы, Σ – положительно определенная $(m \times m)$-матрица, $\nu \in \mathbb{R}$, $\kappa, \delta \in \mathbb{R}^+$. Распределение вероятностей $F_{GVG}(d\boldsymbol{x}; \boldsymbol{a}, \boldsymbol{b}, \Sigma, \nu, \kappa, \delta)$ на $(\mathbb{R}^m, \mathfrak{B}_m)$ задает многомерное обобщенное гиперболическое распределение, если $F_{GVG}(B; \boldsymbol{a}, \boldsymbol{b}, \Sigma, \nu, \kappa, \delta) = (\mathcal{N}_{\boldsymbol{b}+z\boldsymbol{a},\, z\Sigma} \circ P_{GG}(dz; \nu, \kappa, \delta)(B)$ для любого $B \in \mathfrak{B}_m$, то есть

$$F_{GVG}(d\boldsymbol{x}; \boldsymbol{a}, \boldsymbol{b}, \Sigma, \nu, \kappa, \delta) = \int_0^\infty \mathcal{N}_{\boldsymbol{b}+z\boldsymbol{a},\, z\Sigma}(d\boldsymbol{x}) P_{GG}(dz; \nu, \kappa, \delta).$$

Как уже неоднократно отмечалось выше, класс многомерных обобщенных дисперсионных гамма-распределений очень богат и, в частности, содержит (a) симметричные и несимметричные (skew) многомерные распределения Стьюдента, которым соответствует смешивающее обратное гамма-распределение; (b) многомерные дисперсионные гамма-распределения (variance gamma (VG) distributions), включая симметричные и несимметричные многомерные распределения Лапласа, которым соответствует смешивающее гамма-распределение; (c) многомерные нормальные\\обратные гауссовские распределения, которым соответствует смешивающее распределение Леви, и многие другие типы, в частности, не входящие в класс обобщенных гиперболических распределений.

Приводимая ниже теорема 1.10 жестко связывает условия сходимости распределений статистик, построенных по выборкам случайного объема, к многомерным обобщенным дисперсионным гамма-распределениям, со сходимостью распределения интенсивности потока информативных событий, в результате которых накапливается анализируемая выборка, к соответствующему обобщенному гамма-распределению и может служить удобным обоснованием высокой адекватности многомерных обобщенных дисперсионных гамма-распределений в качестве моделей статистических закономерностей поведения сложных стохастических систем.

Теорема 1.10. *Пусть семейство случайных величин* $\{n^{-1}N_n\}_{n\in\mathbb{N}}$ *слабо относительно компактно, нормирующие константы имеют вид* (5) *и удовлетворяют условию* (4) *с некоторыми* $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^m$. *Предположим, что статистика* $\boldsymbol{T}_{n,k}$ *асимптотически нормаль-*

на в смысле (6) *с некоторой симметричной положительно определенной матрицей* Σ. *Тогда распределение статистики* $\boldsymbol{T}_{n,N_n}$, *построенной по выборке случайного объема* N_n, *сходится при* $n \to \infty$ *к* m*-мерному обобщенному дисперсионному гамма-распределению*:

$$\boldsymbol{\mathcal{L}}\big(\sigma_n\sqrt{n}(\boldsymbol{T}_{n,N_n} - \theta_n) - \boldsymbol{b}_n\big) \Longrightarrow F_{GVG}(d\boldsymbol{x}; \boldsymbol{a}, \boldsymbol{b}, \Sigma, \nu, \kappa, \delta)$$

тогда и только тогда, когда

$$\mathsf{P}(N_n < nx) \Longrightarrow P_{GG}(x; -\nu, \kappa, 1/\delta).$$

Доказательство. Эта теорема является непосредственным следствием теоремы 1.9 с учетом легко проверяемого факта: если $\mathsf{P}(\xi < x) = P_{GG}(x; \nu, \kappa, \delta)$, то $\mathsf{P}(\xi^{-1} < x) = P_{GG}(x; -\nu, \kappa, 1/\delta)$, $x \in \mathbb{R}$.

Глава 2

Модифицированный сеточный метод разделения дисперсионно-сдвиговых смесей нормальных законов

2.1 Предварительные замечания. Цель исследования

При практическом решении задачи моделирования и исследования волатильности (изменчивости) хаотических стохастических процессов ключевым этапом является статистическое разделение смесей вероятностных распределений. Задача разделения смесей – статистического оценивания параметров смесей вероятностных распределений – в деталях разобрана, например, в книге [59].

Для решения задачи разделения смесей вероятностных распределений традиционно используются итерационные процедуры типа EM-алгоритма. 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, [59]. Вместе с тем, в указанной книге предложен и исследован принципиально новый «сеточный» метод приближенного решения задачи разделения смесей. В работе [60] 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 [59], в общем случае на практике приходится решать задачу разделения конечных смесей нормальных законов с произвольно большим числом неизвестных (параметров отдельно взятых 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. Как было отмечено в Главе 1, при решении многих задач, например, связанных с анализом процессов атмосферной или плазменной турбулентности, а также процессов, описывающих эволюцию различных финансовых индексов, высочайшую адекватность продемонстрировали модели, основанные на обширном классе дисперсионно-сдвиговых смесей нормальных законов. В частности, мы будем рассматривать хорошо зарекомендовавшие себя классы обобщенных гиперболических распределений (см. раздел 1.2.2 опр. (1.5)) и обобщенных дисперсионных гамма-распределений приведено (см. раздел 1.2.3, опр. (1.7)).

В указанных семействах смесей число неизвестных параметров равно пяти или шести, если учитывать неслучайный сдвиг. Вместе с тем, как было отмечено ранее, у подобных моделей имеются довольно серьезные теоретические обоснования: в работах [62, 63] показано, что указанные модели являются асимптотическими аппроксимациями в простой предельной схеме случайного суммирования и потому могут успешно применяться для анализа процессов по типу 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 [6].

В данной главе предлагается принципиально новый метод разделения дисперсионно-сдвиговых смесей нормальных законов, в частности, на примере обобщенных гиперболических и обобщенных дисперсионных гамма-распределений. Также в этой главе изучаются основные свойства данного метода и предложены практические рекомендации по его использованию, а также приводятся результаты применения как на искуственно сгенерированных выборках, так и на реальных данных.

2.2 Описание модифицированного сеточного метода разделения дисперсионно-сдвиговых смесей нормальных законов и его свойства

Оказывается, сеточные методы разделения смесей достаточно эффективны не только при разделении конечных смесей нормальных законов, но и при разделении произвольных дисперсионно-сдвиговых смесей нормальных законов. Поясним сказанное на примере задачи оценивания параметров обобщенных гиперболических распределений.

Для решения задачи оценивания параметров обобщенных гиперболических распределений традиционно используется метод, предложенный в статье [64], и по сути являющийся классическим EM-алгоритмом, приспособленным к конкретной задаче, и соответственно, наследующий присущие EM-алгоритмам недостатки.

Рассмотрим следующий альтернативный двухэтапный метод. На первом этапе на положительной полупрямой выделим основную часть носителя смешивающего распределения, то есть ограниченный интервал, вероятность которого, вычисленная в соответствии со смешивающим распределением, практически равна единице. На этот интервал накинем конечную сетку, содержащую, 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:

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

В смеси, стоящей в правой части соотношения (2.2), неизвестными являются только параметры $p_1, \ldots, p_{K-1}, \alpha$. Пусть $x_1, \ldots, x_n$ – анализируемая выборка значений случайной величины с оцениваемым обобщенным гиперболическим распределением. Итерационный процесс, определяющий сеточный EM-алгоритм для данной задачи, определяется следующим образом. 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

$$\varphi_{ij}^{(m)} = \frac{1}{\sqrt{u_i}} \varphi\Big(\frac{x_j - \alpha^{(m)} u_i}{\sqrt{u_i}}\Big), \quad g_{ij}^{(m)} = \frac{p_i^{(m)} \varphi_{ij}^{(m)}}{\sum_{r=1}^{K} p_r^{(m)} \varphi_{rj}^{(m)}}, \quad i = 1, \ldots, K;\ j = 1, \ldots, n.$$

Тогда, используя стандартные рассуждения, определяющие вычислительные формулы EM-алгоритма для параметров конечной смеси нормальных законов (см, например, [59], разделы 5.3.7-5.3.8), следует положить

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

Обозначим $\overline{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$. Используя соотношение (5.3.24) в [59], с учетом очевидного равенства $\sum_{i=1}^{K} g_{ij}^{(m)} = 1$ можно заметить, что уточненная оценка параметра α имеет вид

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

то есть равна отношению генерального выборочного среднего и текущего эмпирического среднего смешивающего распределения, что вполне согласуется с тем, что в соответствии с приводимым ниже соотношением (2.5) в данном случае $\mathsf{E}X = \alpha \mathsf{E}U$.

В силу монотонности классического EM-алгоритма справедливо следующее утверждение.

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*

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

Замечание 2.1. В разделе 5.7.4 книги [59] показано, что при каждом фиксированном значении параметра α сеточная функция правдоподобия $L(p_1, \ldots, p_{K-1}, \alpha;\, x_1, \ldots, x_n)$ вогнута по аргументам $p_1, \ldots, p_{K-1}$. Поэтому на каждом шаге итерационного процесса вместо соотношения (2.3) можно использовать любой более быстрый алгоритм максимизации функции $L(p_1, \ldots, p_{K-1}, \alpha^{(m)};\, x_1, \ldots, x_n)$ по переменным $p_1, \ldots, p_{K-1}$, Например, оценки весов $p_1, \ldots, p_K$ можно искать методом условного градиента [59, 65].

Таким образом, на первом этапе получаются оценки параметра α и весов всех узлов u_i конечной сетки, накинутой на носитель смешивающего обобщенного обратного гауссовского распределения $P_{GIG}(z;\, \nu, \mu, \lambda)$.

На втором этапе остается применить 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

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

где $u_0 = 0$, $u_{K+1} = \infty$.

На практике хорошие результаты показал подход с применением метода наименьших квадратов. Для поиска параметров подбираемых распределений использовалось несколько алгоритмов, в частности, алгоритм ns2sol, описанный в книге [66]. Данный алгоритм, доступный во многих статистических пакетах, отличается высоким быстродействием и

возможностью при желании задавать априорные разумные интервалы для поиска параметров.

2.3 О практическом выборе сетки на первом этапе модифицированного сеточного метода разделения дисперсионно-сдвиговых смесей нормальных законов

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. Рассмотрим этот вопрос подробнее.

Формально рассматриваемая задача выглядит так: по наблюдаемым значениям $x_1, \ldots, x_n$ из *смешанного* распределения требуется как можно точнее оценить носитель *смешивающего* распределения. Под носителем здесь предполагается не носитель в классическом смысле этого термина (в нашем случае это вся положительная полупрямая), а упомянутый ранее ограниченный интервал, вероятность которого, вычисленная в соответствии со смешивающим распределением, практически равна единице (то есть интервал, на который накидывается сетка $u_1, \ldots, u_K$).

В дальнейшем будем считать, что $x_1, \ldots, x_n$ – независимые реализации случайной величины $X = Y\sqrt{U} + \alpha U$, где Y – случайная величина со стандартным нормальным распределением, а U – независимая от нее случайная величина с обобщенным обратным гауссовским распределением. Тогда, очевидно, распределение случайной величины X имеет вид (2.1). Предположим, что у случайной величины U существуют моменты первых двух порядков. Тогда, как несложно видеть,

$$\mathsf{E}X = \mathsf{E}Y \cdot \mathsf{E}\sqrt{U} + \alpha \mathsf{E}U = \alpha \mathsf{E}U. \tag{2.5}$$

При этом по усиленному закону больших чисел с вероятностью единица $\overline{x} \longrightarrow \mathsf{E}X$ $(n \to \infty)$, и при больших n справедливо приближенное равенство $\mathsf{E}X \approx \overline{x}$, так что с учетом (2.5)

$$\mathsf{E}U \approx \frac{\overline{x}}{\alpha}. \tag{2.6}$$

Далее, очевидно,

$$\mathsf{E}X^2 = \mathsf{E}Y^2 \cdot \mathsf{E}U + 2\alpha \mathsf{E}X \cdot \mathsf{E}U^{3/2} + \alpha^2 \mathsf{E}U^2 = \mathsf{E}U + \alpha^2 \mathsf{E}U^2. \tag{2.7}$$

Поэтому, обозначив

$$m^2 = \frac{1}{n}\sum\nolimits_{i=1}^{n} x_i^2,$$

получаем приближенное равенство $\mathsf{E}X^2 \approx m^2$, так что с учетом (2.6) и (2.7) получаем

$$\mathsf{E}U^2 \approx \frac{1}{\alpha^2}\Big(m^2 - \frac{\overline{x}}{\alpha}\Big). \tag{2.8}$$

Если параметр α известен, то для определения верхней границы u^* сетки, накидываемой на носитель распределения случайной величины U, можно задать малое положительное число ε и воспользоваться требованием

$$\mathsf{P}(U \geqslant u^*) \leqslant \varepsilon. \tag{2.9}$$

Для гарантированного выполнения требования (2.9) можно использовать неравенство Маркова:

$$\mathsf{P}(U \geqslant u^*) \leqslant \frac{\mathsf{E}U^2}{(u^*)^2} \leqslant \varepsilon,$$

откуда с учетом (2.8)

$$(u^*)^2 \geqslant \frac{\mathsf{E}U^2}{\varepsilon} \approx \frac{1}{\alpha^2\varepsilon}\Big(m^2 - \frac{\overline{x}}{\alpha}\Big)$$

или

$$u^* \approx \frac{1}{\alpha\sqrt{\varepsilon}}\sqrt{m^2 - \frac{\overline{x}}{\alpha}}. \tag{2.10}$$

Если же параметр α, определяющий асимметрию распределения случайной величины X, неизвестен, то можно воспользоваться следующими рассуждениями. Обозначим

$$q_n = \frac{1}{n}\sum\nolimits_{i=1}^{n} \mathbf{1}(x_i < 0),$$

где $\mathbf{1}(A)$ – индикаторная функция множества (события) A. При этом по усиленному закону больших чисел с вероятностью единица $q_n \longrightarrow \mathsf{P}(X < 0)$ $(n \to \infty)$, так что при больших n справедливо приближенное равенство

$$q_n \approx \mathsf{P}(X < 0). \tag{2.11}$$

Но

$$\mathsf{P}(X < 0) = \int_0^\infty \Phi\big(-\alpha\sqrt{u}\big)p_{GIG}(u;\nu,\mu,\lambda)du = \mathsf{E}\Phi\big(-\alpha\sqrt{U}\big). \tag{2.12}$$

Предположим сначала, что $q_n < \frac{1}{2}$. Если n достаточно велико, то можно с большой степенью уверенности утверждать, что тогда $\overline{x} > 0$ и $-\alpha < 0$, то есть $\alpha > 0$ и, стало быть, на положительной полуоси значений аргумента u функция $\Phi(\alpha u)$ вогнута, то есть выпукла вверх. Тогда из (2.11) и (2.12), дважды применяя неравенство Йенсена, в силу монотон-

ности функции Φ получаем

$$1-q_n \approx 1-\mathsf{E}\Phi\big(-\alpha\sqrt{U}\big)=\mathsf{E}\Phi\big(\alpha\sqrt{U}\big)\leqslant\Phi\big(\alpha\mathsf{E}\sqrt{U}\big)\leqslant\Phi\big(\alpha\sqrt{\mathsf{E}U}\big). \tag{2.13}$$

Если теперь для $t\in(0,1)$ символом v_t обозначить t-квантиль стандартного нормального закона, то из (2.13) и (2.6) вытекает «приближенное неравенство» $v_{1-q_n}\leqslant\alpha\sqrt{\mathsf{E}U}$, то есть

$$\alpha\geqslant\frac{v_{1-q_n}}{\sqrt{\mathsf{E}U}}\approx\frac{v_{1-q_n}\sqrt{\alpha}}{\sqrt{\overline{x}}},$$

откуда получаем, что при достаточно больших n

$$\alpha\geqslant\frac{v_{1-q_n}^2}{\overline{x}}. \tag{2.14}$$

Если теперь задать малое положительное число ε, то то для определения верхней границы u^* сетки, накидываемой на носитель распределения случайной величины U, можно воспользоваться требованием (2.9), для гарантированного выполнения которого с учетом (2.6) и (2.14) можно использовать неравенство Маркова:

$$\mathsf{P}(U\geqslant u^*)\leqslant\frac{\mathsf{E}U}{u^*}\approx\frac{\overline{x}}{\alpha u^*}\leqslant\frac{(\overline{x})^2}{v_{1-q_n}^2u^*}\leqslant\varepsilon,$$

откуда окончательно вытекает оценка

$$u^*\approx\frac{(\overline{x})^2}{v_{1-q_n}^2\varepsilon}. \tag{2.15}$$

В случае $q_n\geqslant\frac{1}{2}$, если n достаточно велико, то можно с большой степенью уверенности утверждать, что $\overline{x}\leqslant 0$ и $-\alpha\geqslant 0$, то есть на положительной полуоси значений аргумента u функция $\Phi(-\alpha u)$ вогнута, то есть выпукла вверх. Тогда из (2.11) и (2.12), дважды применяя неравенство Йенсена, в силу монотонности функции Φ получаем

$$q_n\approx\mathsf{E}\Phi\big(-\alpha\sqrt{U}\big)\leqslant\Phi\big(-\alpha\sqrt{\mathsf{E}U}\big),$$

откуда вытекает «приближенное неравенство» $v_{q_n}\leqslant-\alpha\sqrt{\mathsf{E}U}$, то есть

$$-\alpha\geqslant\frac{v_{q_n}}{\sqrt{\mathsf{E}U}}\approx\frac{v_{q_n}\sqrt{|\alpha|}}{\sqrt{|\overline{x}|}},$$

и при достаточно больших n

$$|\alpha|\geqslant\frac{v_{q_n}^2}{|\overline{x}|}. \tag{2.16}$$

Для определения верхней границы u^* сетки, накидываемой на носитель распределения случайной величины U, снова зададим малое положительное число ε и потребуем, чтобы

было справедливо условие (2.9), для гарантированного выполнения которого с учетом (2.6) и (2.16) используем неравенство Маркова и тот факт, что $\operatorname{sign}\overline{x} = \operatorname{sign}\alpha$ при достаточно больших n:

$$\mathsf{P}(U \geqslant u^*) \leqslant \frac{\mathsf{E}U}{u^*} \approx \frac{\overline{x}}{\alpha u^*} = \frac{|\overline{x}|}{|\alpha| u^*} \leqslant \frac{(\overline{x})^2}{v_{q_n}^2 u^*} \leqslant \varepsilon. \tag{2.17}$$

В силу симметричности нормального распределения $v_t = -v_{1-t}$ для любого $t \in (0,1)$. Поэтому $v_{q_n}^2 = v_{1-q_n}^2$, и в случае $q_n \geqslant \frac{1}{2}$ соотношение (2.17) снова приводит к оценке (2.15).

Справедливости ради необходимо отметить, что оценки (2.10) и (2.15) являются завышенными, но они гарантируют, что $(1-\varepsilon)$-почти весь носитель распределения случайной величины U будет лежать внутри интервала $[0, u^*]$.

В разделе 2.7.2 данной главы приведены результаты анализа точности оценок (2.10) и (2.15) на примерах искуственно сгенерированных данных, и на их основе полученных результатов даны практические советы по использованию этих оценок при работе с реальными данными.

2.4 Результаты численных экспериментов на искуственно сгенерированных выборках

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

Тестирование применялось на задаче 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.

На рисунках 2.1 - 2.8 проиллюстрировано качество работы метода: изображены гистограммы и графики для истинной плотности (штрихованные линии), плотность промежуточной оценки, полученной сеточным EM-алгоритмом (пунктирные линии), а также плотность итоговой оценки (непрерывные линии).

Как видно из приводимых на рисунках значений, параметр распределения α всегда оценивается очень точно. Точность оценок остальных параметров удовлетворительная и может быть повышена за счет использования более частых сеток и более чувствительных критериев остановки EM-алгоритма на первом этапе. Следует отметить, что даже в тех случаях, в которых наблюдаются заметные расхождения оценок параметров и их точных значений, оценки самих плотностей довольно близки.

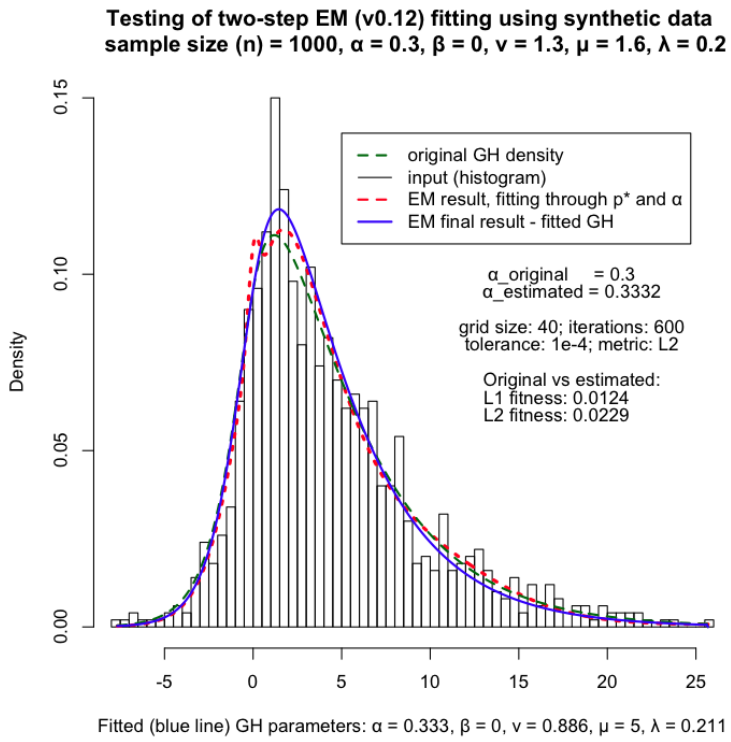

Рисунок 2.1: Тестирование метода на выборке размера 1000 для GH-распределения с параметрами $\alpha = 0.3$, $\beta = 0$, $\nu = 1.3$, $\mu = 1.6$, $\lambda = 0.2$

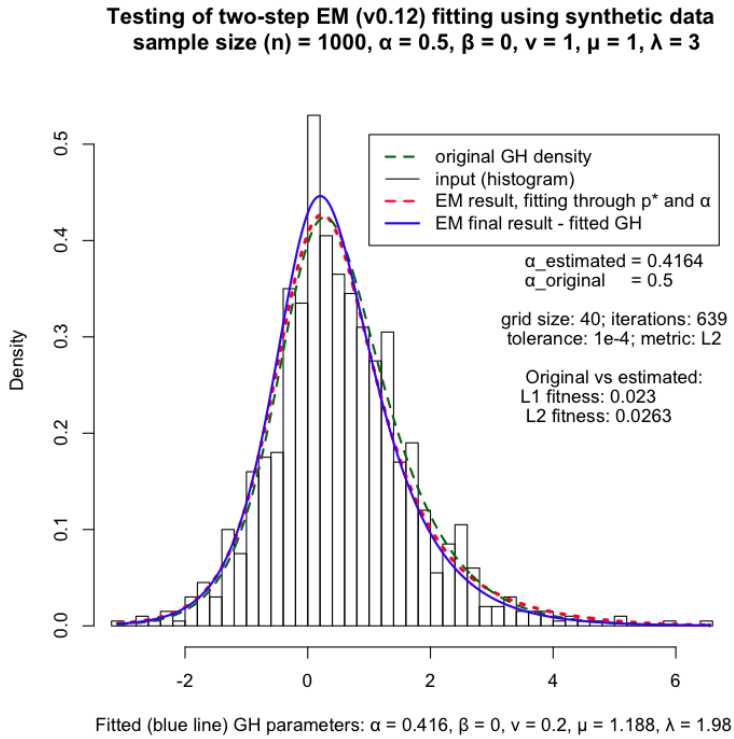

Рисунок 2.2: Тестирование метода на выборке размера 1000 для GH-распределения с параметрами $\alpha = 0.5$, $\beta = 0$, $\nu = 1$, $\mu = 1$, $\lambda = 3$

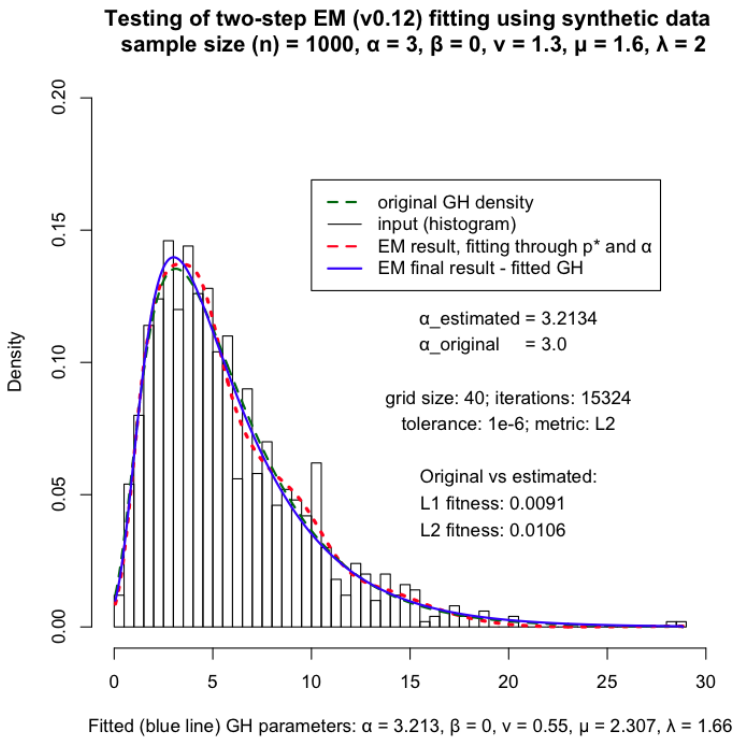

Рисунок 2.3: Тестирование метода на выборке размера 1000 для GH-распределения с параметрами $\alpha = 3$, $\beta = 0$, $\nu = 1.3$, $\mu = 1.6$, $\lambda = 2$

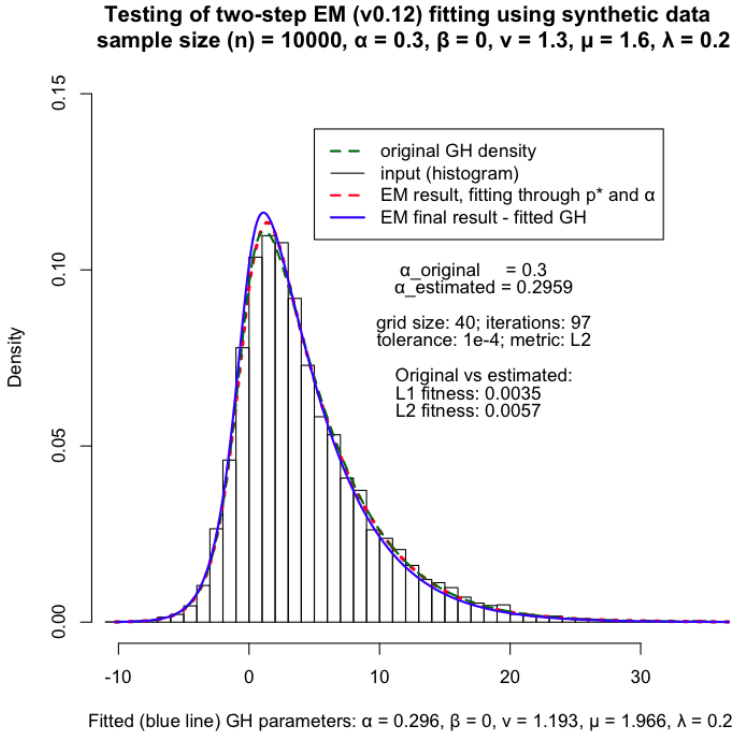

Рисунок 2.4: Тестирование метода на выборке размера 10000 для GH-распределения с параметрами $\alpha = 0.3$, $\beta = 0$, $\nu = 1.3$, $\mu = 1.6$, $\lambda = 0.2$

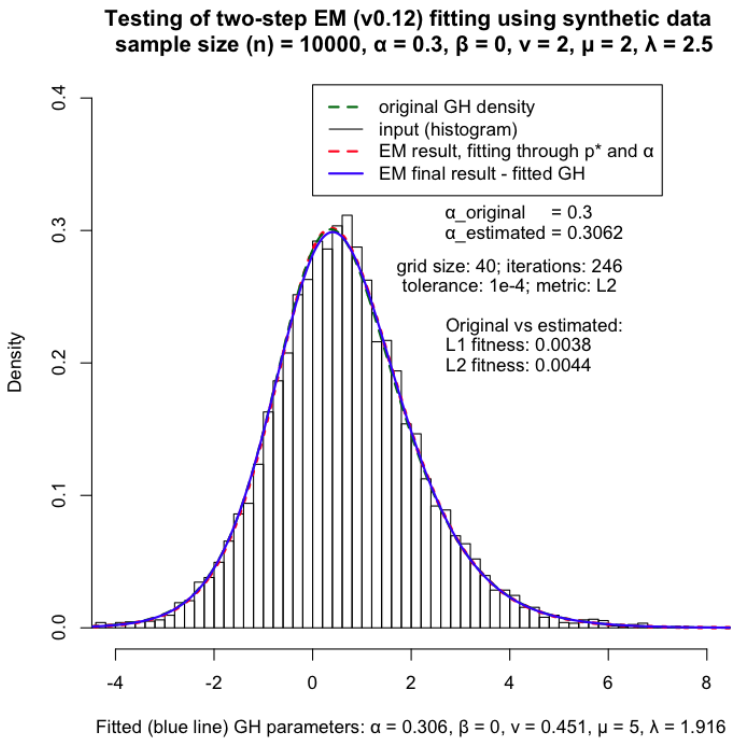

Рисунок 2.5: Тестирование метода на выборке размера 10000 для GH-распределения с параметрами $\alpha = 0.3$, $\beta = 0$, $\nu = 2$, $\mu = 2$, $\lambda = 2.5$

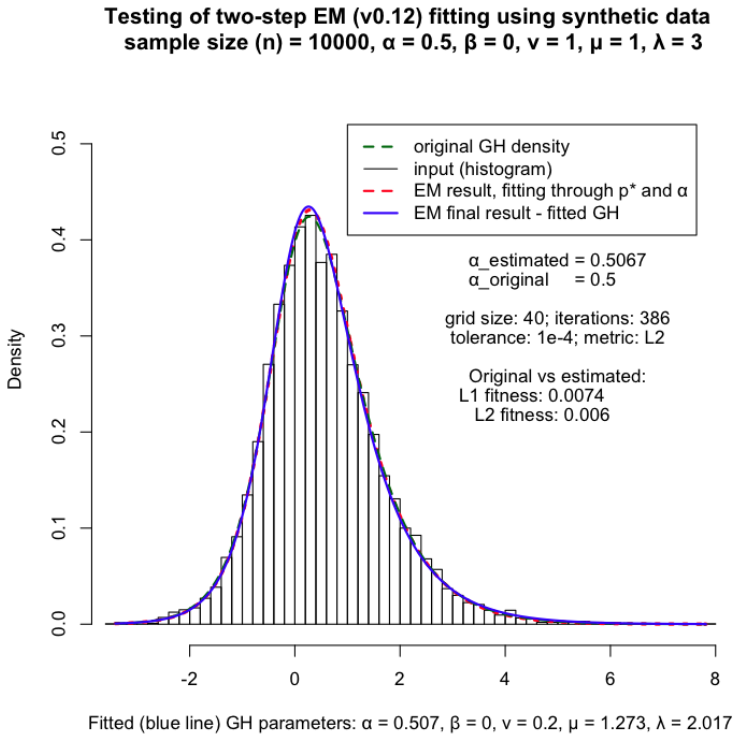

Рисунок 2.6: Тестирование метода на выборке размера 10000 для GH-распределения с параметрами $\alpha = 0.5$, $\beta = 0$, $\nu = 1$, $\mu = 1$, $\lambda = 3$

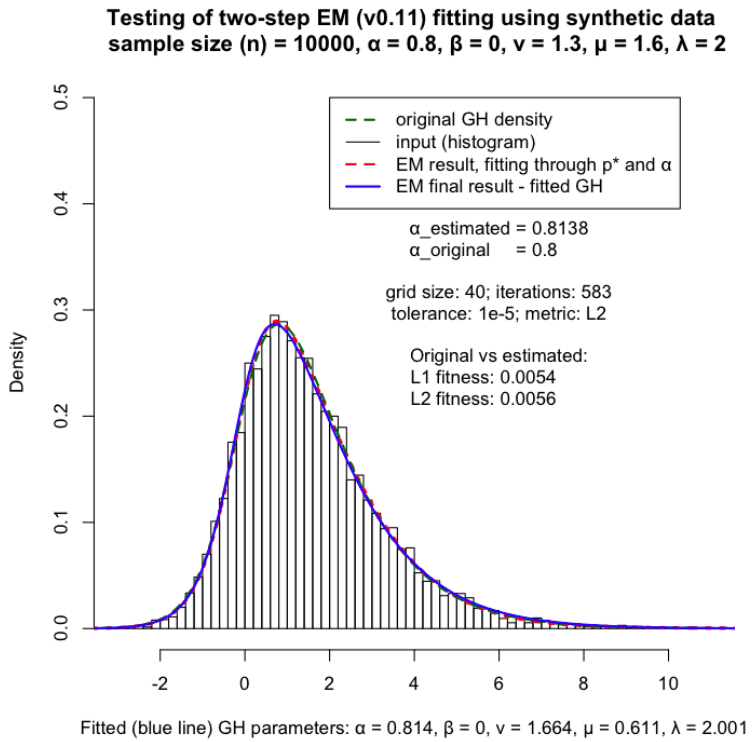

Рисунок 2.7: Тестирование метода на выборке размера 10000 для GH-распределения с параметрами $\alpha = 0.8$, $\beta = 0$, $\nu = 1.3$, $\mu = 1.6$, $\lambda = 2$

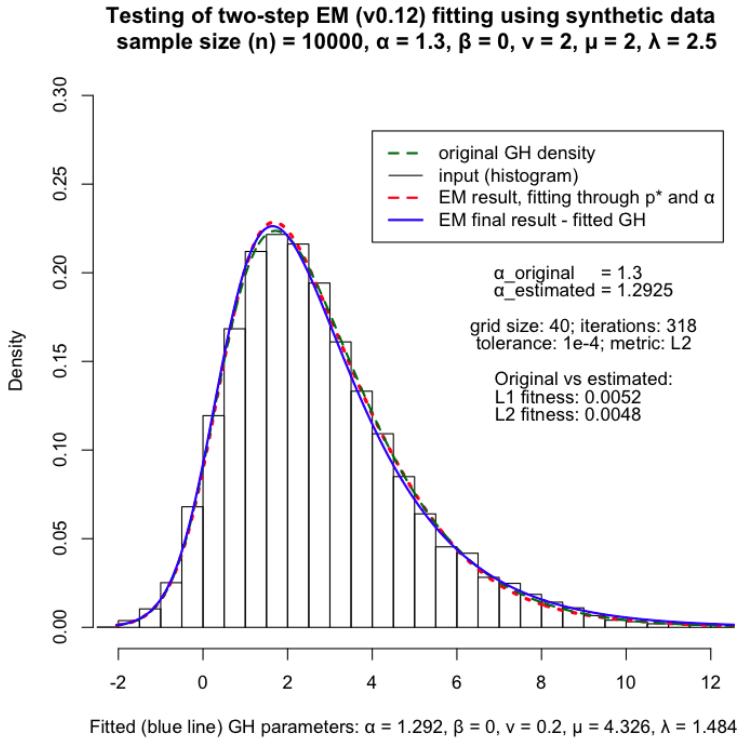

Рисунок 2.8: Тестирование метода на выборке размера 10000 для GH-распределения с параметрами $\alpha = 1.3$, $\beta = 0$, $\nu = 2$, $\mu = 2$, $\lambda = 2.5$

2.5 Результаты численных экспериментов на реальных данных

Помимо искуственно сгенерированных выборок, предложенный метод был применен к реальным финансовым и медицинским данным. В этом разделе приведены результаты 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- (обобщенные гиперболические) и GVG- (обобщенные дисперсионные гамма-) распределения.

Из результатов, приведенных ниже, 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 теста Хи-квадрат.

2.5.1 Основной индекс Корейской биржи - KOSPI

В качестве исходных данных возьмем значения индекса KOSPI (Korea Composite Stock Price Index) – основной индикатор корейской биржи, включающий в себя все торгующиеся на этой бирже компании. Индекс начал рассчитываться в 1983 году и до сих пор является ключевым для корейского рынка, поэтому он достаточно часто подвергается анализу и попыткам предсказания своей динамики, то есть является подходящим кандидатом для анализа. Подробную информацию про KOSPI можно найти на сайте агенства Блумберг: www.bloomberg.com/quote/KOSPI:IND.

Будем рассматривать приращения логарифмов значений индекса (цены) с частотой (тиком) равной 1 минуте, начиная с открытия биржи 8 декабря 2014 года на протяжении трех рабочих дней.

Размер окна установим равным 3 часа: $w = 180$ тиков, сдвиг окна минимальный – $s = 1$ тик.

Применяя предложенный метод к исходным данным, получим ряд параметров $p_i = (\alpha_i, \beta_i, \nu_i, \mu_i, \lambda_i)^T$ для случая GH-распределений и ряд параметров $p_i = (\alpha_i, \beta_i, \nu_i, \kappa_i, \delta_i)^T$ для случая GVG-распределений. В данном эксперименте без ущерба для точности положим все $\beta_i = 0$, оставив только четыре параметра в каждом из семейств рассматриваемых распределений. Как было отмечено ранее, метод устойчив к входным данным, при этом сдвиг окна минимальный, следовательно ряд p_i дает нам сглаженную картину. На Рис. 2.9 представлено изменение приближающего обобщенного гиперболического (GH-) распределения в динамике.

Зададимся следующим вопросом: какое из двух семейств распределений (GH или GVG) лучше аппроксимирует наблюдаемые данные.

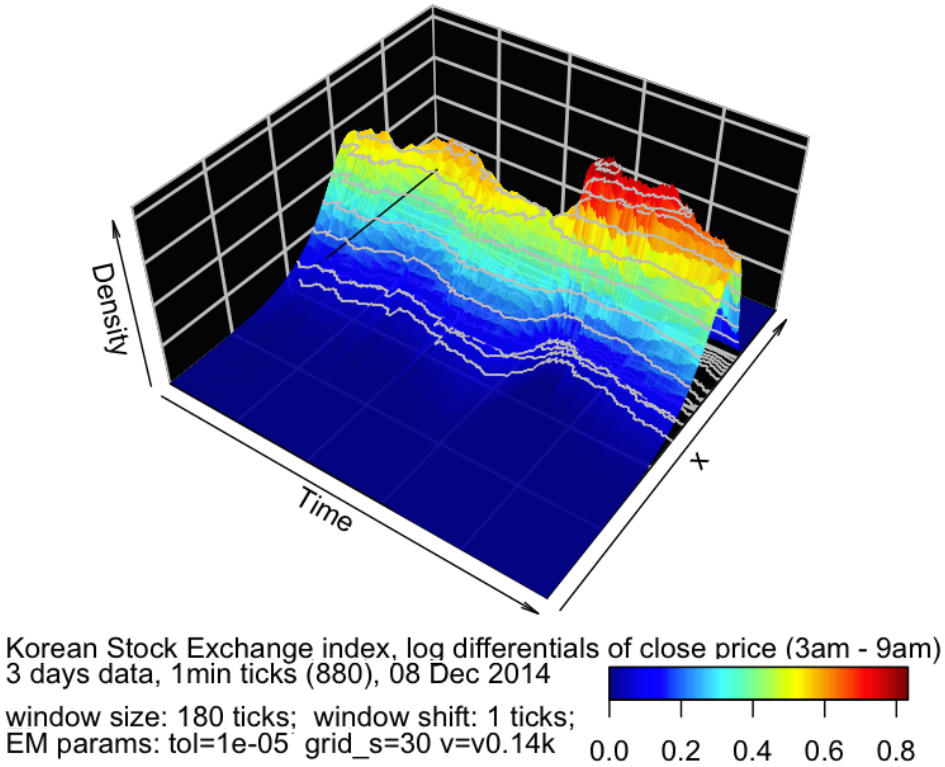

Рисунок 2.9: Изменение GH-приближения распределения индекса KOSPI во времени

Выберем несколько окон и посчитаем P-значение теста хи-квадрат. Как оказалось, GVG демонстрирует лучшую согласованность, см. рис 2.10, 2.11.

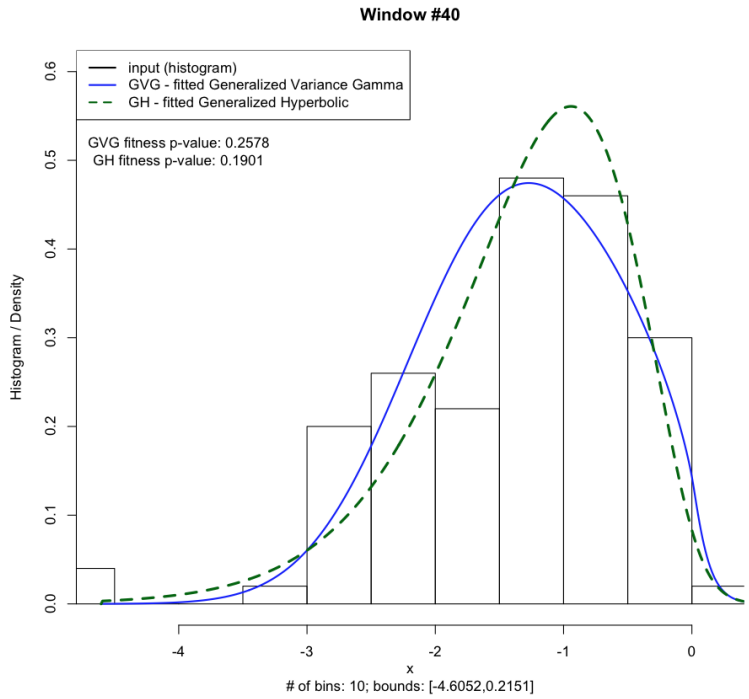

Рисунок 2.10: Сравнение приближений, GH- и GVG-распределения, KOSPI, окно 40

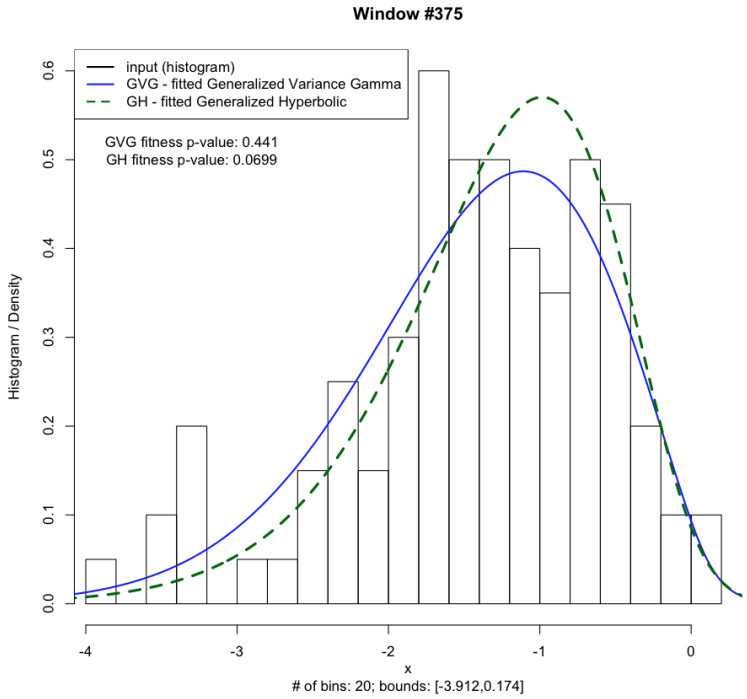

Рисунок 2.11: Сравнение приближений, GH- и GVG-распределения, KOSPI, окно 375

2.5.2 Индекс Dow Jones Industrial – DJI

В качестве другого набора исходных данных возьмем индекс Dow Jones Industrial (включает в себя 30 американский компаний), один из самых часто обсуждаемых и анализируемых производных инструментов. Как и в предыдущем случае, будем рассматривать минутные тики приращения логарифмов значений индекса. Начнем рассмотрение с открытия торгового дня 11 декабря 2014 года. Размер окна выберем равным двум часам (120 тиков). По сравнению с KOSPI, в данном случае мы выбрали немного меньший размер окна, а значит и размер сетки, накидываемой на носитель смешивающего распределения, нужно выбрать немного меньше (30).

На Рис. 2.12 показана GVG-аппроксимация данных на начало дня (первые $\approx$ 3 часа, 156 тиков). Полученный ряд представляет собой набор удобных данных для анализа будет использоваться в Главе 3 как входныые параметры для алгоритма прогнозирования.

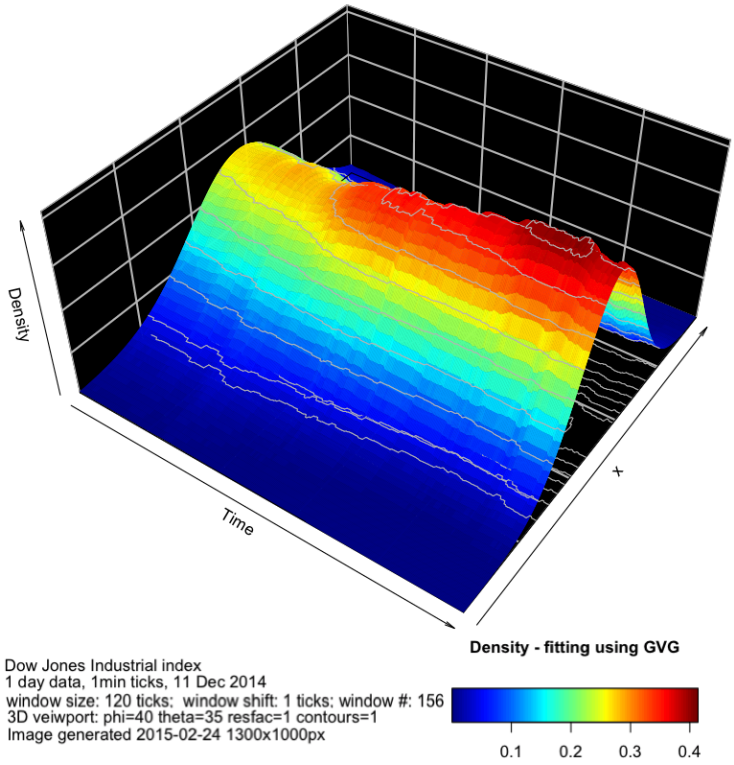

Рисунок 2.12: Изменение GVG-приближения распределения индекса DJI во времени

Подводя итог отметим, что в данном разделе к различным финансовым исходным данным был применен предложенный метод и показана адекватность предложенных моделей. Тем не менее, конечная практическая задача, где используется полученный результат, не была обозначена. Чтобы устранить этот пробел необходимо уточнить, что в Главе 3 полученные аппроксимации будут использованы в качестве выходных данных для алгоритма прогнозирования рисков.

2.6 Выявление двигательной активности в головном мозге человека с помощью предложенного метода

В данном разделе описывается успешное применение предложенного метода для решения одной из проблем, возникающих при исследовании активности головного мозга.

2.6.1 Постановка задачи и основные обозначения

В клинической практике нейрохирургического вмешательства есть очень интересная и крайне важная задача: точная *дооперационная* локализация невосполнимых зон мозга и их соотношение с пораженными областями. Решая эту задачу, можно минимизировать послеоперационные осложнения для пациента, или вовремя принять решение о непроведении операции, если есть уверенность в нарушении невосполнимых зон.

В частности, очень важна точна предоперационная локализация первичной моторной коры (M1). Эта задача особенно трудна, если головной мозг пациента уже страдает от различных поражений (например, эпилепсии) - тогда одной анатомической информации, полученной, например, из томограммы, оказывается недостаточно.

Одним из наиболее часто используемых экспериментальных методов исследования активности головного мозга в данной задаче является так называемый метод вызванных потенциалов (см., например, [75, 76]). Суть эксперимента заключается в следующем: субъект неоднократно делает некоторые движения пальцем в то время, как активность мозга и некоторые вспомогательные сигналы записываются для дальнейшего анализа. Ключевой проблемой является определение точек миограммы, которые соответствуют началу движений. Чем точнее обнаружены точки, тем более успешно при обработке магнитоэнцефалограмы можно идентифицировать датчики, которые находятся ближе всего к зонам мозговой активности. Более подробно об этом направлении исследований можно прочитать в работе [77].

Вкратце, предложенный в работе подход сводится к задаче анализа данных миограммы. Простейшая разумная математическая модель для миограммы – это циклический нестационарный случайный процесс, который может быть представлен в виде

$$\xi(t) = \sum_i ((s_i(t) + \varepsilon_i(t) + \theta_i(t))\mathbf{1}\{t_i \leqslant t < t_{i+1}\}),$$

где случайный процесс $s_i(t)$ соответствует компоненте сигнала, относящейся к движению пальцев ($s_i(t) = 0$ вне периода движения), $\varepsilon_i(t)$ - шум отдыха (равен нулю во время движения), $\theta_i(t)$ – шум движения (равен нулю во время отдыха); в нейрофизиологии полуинтервал $[t_i, t_{i+1})$ называется эпохой (каждая эпоха включает в себя интервал отдыха перед движением и само движение), i – номер эпохи.

Скользящая дисперсия, получаемая из миограммы, представляет собой также циклический нестационарный процесс, но с гораздо меньшим уровнем шума. Переход к скользящей дисперсии позволяет устранить тренды и подчеркнуть моменты перехода из отдыха в движение, которые затем определяются пороговой обработкой. Метод, предложенный в [77] продемонстрировал очень высокую точность. Тем не менее, из-за заметной ненормальности распределения шума в рамках метода, предложенного в [77] пороги были введены несколько искусственно.

Как развитие метода, описанного в [77], были предложены статистические методы, основанные на моделях смесей, предназначенные для точного определение точек, соответствующих началам движений.

2.6.2 Сглаживание сигнала с помощью скользящего разделения конечных смесей

Применяя так называемый метод скользящего разделения смесей (МСМ-метод), предложенный в [59], волатильность случайного процесса можно разбить на две компоненты: динамическую и диффузионную.

В рамках этого метода, одномерное распределение приращений основного процесса аппроксимируется конечными сдвиг-масштабными смесями нормальных распределений. Теоретические основы этих моделей подробно описаны в [59].

Для анализа динамики изменений в стохастическом процессе, проблема статистического оценивания неизвестных параметров распределения должна быть последовательно решена для части выборки, которая перемещается в направлении времени, т.е. двигающегося окна. Как правило, размер окна (подвыборки) фиксирован. После того, как анализируемые параметры получены для текущего положения, окно должно быть сдвинуто на один элемент исходной выборки, т.е. метод будет анализировать следующую подвыборку. Это позволяет обнаружить все возможные изменения в поведении компонент.

Положим, что функцию распределения для соответствующего момента времени (местоположения окна) можно представить в виде

$$F(x) = \sum_{i=1}^{k} \frac{p_i}{\sigma_i\sqrt{2\pi}} \int\limits_{-\infty}^{x} \exp\Big\{ - \frac{(t-a_i)^2}{2\sigma_i^2} \Big\}\, dt, \tag{2.18}$$

где

$$\sum_{i=1}^{k} p_i = 1, \quad p_i \geqslant 0. \tag{2.19}$$

(для всех $x \in \mathbb{R}$, $a_i \in \mathbb{R}$, $\sigma_i > 0$, $i = 1, \ldots, k$). Модель (2.18) называется конечной сдвиг-масштабной смесью нормальных законов. Параметры $p_1, \ldots, p_k$ – веса, удовлетворяющие (2.19). Параметр k – число компонент смеси. Параметры $a_1, \ldots, a_k$ ассоциированы с динамической компонентой волатильности (дисперсии) процесса, а параметры $\sigma_1, \ldots, \sigma_k$ ассоциированы с диффузионной, см. [59]. А именно, если Z – случайная величина с функцией распределения (2.18), то ее дисперсия может быть представлена как сумма двух компонент:

$$\mathsf{D}Z = \sum_{i=1}^{k} p_i (a_i - \overline{a})^2 + \sum_{i=1}^{k} p_i \sigma_i^2, \tag{2.20}$$

где

$$\overline{a} = \sum_{i=1}^{k} p_i a_i.$$

Первое слагаемое в правой части зависит только от весов p_i и ожидаемых значений a_i компонент смеси (2.18). Так как Z – приращение основного процесса, то a_i – ожидаемое значение приращения, т.е. трендовая составляющая. Таким образом, первая компо-

нента является частью итоговой дисперсии (изменчивости), что обусловлено существованием простейших трендов. Это называется *динамической* компонентой дисперсии. В то же время, второе слагаемое в правой части (2.20) зависит только от весов p_i дисперсий σ_i^2 компонент и представляет чисто стохастическую *диффузионную* компоненту итоговой дисперсии.

Оказывается, динамическая компонента итоговой дисперсии очень удобна для использования в решении исходной задачи. В дополнение к непосредственно исходному ряду значений миограммы, полученная динамическая компонента итоговой дисперсии также будет использоваться в качестве альтернативных исходных данных (временного ряда) для анализа.

Важно отметить, что распределение значений динамической компоненты, полученной за период отдыха, очень далеко от нормального, см. Рис. 2.13

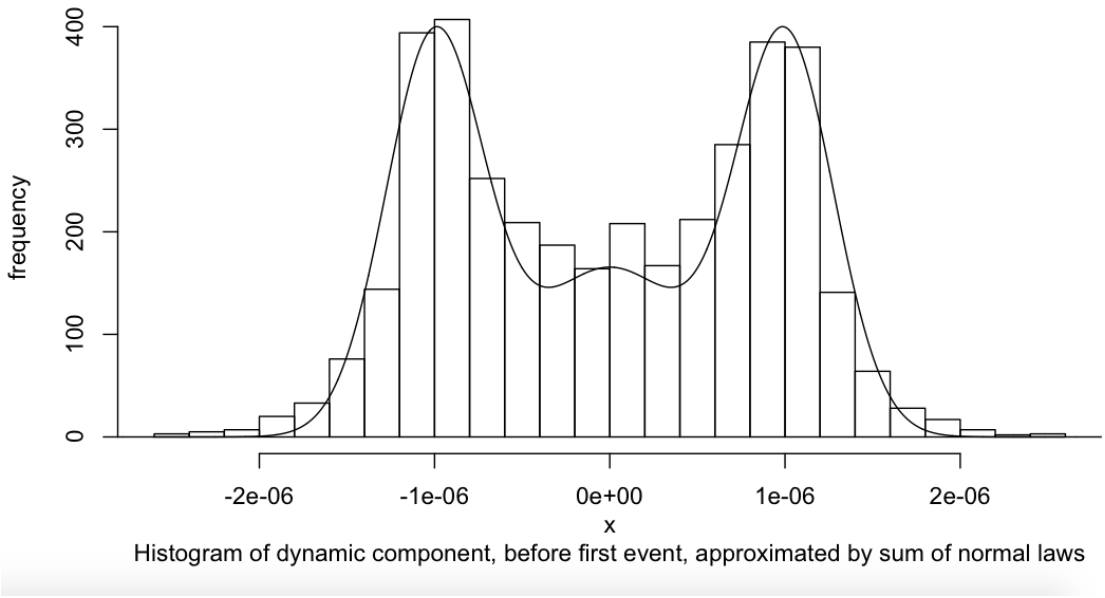

Рисунок 2.13: Гистограмма динамической компоненты волатильности в период отдыха перед первым событием и подогнанная конечная смесь 3 нормальных законов

2.6.3 Определение начальных точек с помощью модифицированного метода из динамической компоненты

Основная идея заключается в обработке динамической компоненты как случайного процесса (по сути, она таковым и является) и анализе ее распределений, предполагая, что они сами по себе являются нормальными смесями (см. Рис. 2.13).

Чтобы разделить эти смеси, используем только первый шаг предложенного метода. В результате получим значения весов p_i i-го узла сетки для каждого окна. Размер окна установим равным 100 (одна десятая секунды), сдвиг окна возьмем равным единице.

Для того, чтобы сравнить друг с другом векторы $\mathbf{p}_i$ вероятностей $(p_{i,1}, \ldots, p_{i,K})$, полученные на различных непересекающихся (для независимости) окнах, для каждого окна с номером $i > 100$ вычислим значение

$$z_i = \|\mathbf{p}_i - \mathbf{p}_{i-100}\| = \left[\sum_{j=1}^{K}(p_{i,j} - p_{i-100,j})^2\right]^{1/2}.$$

Установим пороговое значение для z_i равным $\theta = 0.97$, чтобы фиксировать лишь очень значительные изменения вектора $\mathbf{p}_i = (p_{i,1}, ..., p_{i,K})$ (K – число узлов сетки). Выделим все $z_i > \theta$. Данные события собраны в узкие группы, причем каждая группа *непрерывна* в том смысле, что она состоит из событий, следующих непосредственно друг за другом, т.е. на соседних окнах. Таким образом не возникает проблем с выделением целых групп вместо одиночных событий.

Из каждой группы возьмем первое значение и сдвинем его на 150 (размер окна, используемый методом (100) плюс размер окна, использованный ранее для получения динамической компоненты (50)). Примем во внимание тот факт, что каждая группа может иметь *отражение* после определяемого значения, связанное с тем, что параметры возвращаются к значениям, близким к исходным. Будем считать следующую группу отражением, если она отстоит не дальше, чем на 300 окон от первоначальной группы. Отражения исключим из процесса обнаружения.

Несложный алгоритм, описанный выше, дает следующие моменты времени (отмечены зеленым, см. Рис. 2.14): 4074, 5608, 6256, 8446, 11284, 12938, 15017, 17327, 19685, 21321, 23531. Фактические события (красные линии): 4032, 8443, 11298, 12917, 14976, 17326, 19688, 21337, 23539.

Как видно из иллюстрации, между первым реальным событием и вторым реальным событием было обнаружено несколько дополнительных событий. Все остальные события оценивается с точностью ≈ 7 мс, что является очень хорошим результатом для решаемой задачи.

Этот эксперимент приводит к выводу, что вектор $\mathbf{p}$ весов, как правило, обладает волатильностью. Сосредоточимся теперь на самих событиях. Если проанализировать поведение вектора $\mathbf{p}$ *внутри* обнаруженных событий, заметим, что он меняется медленнее, чем *вне* обнаруженных событий. Как результат этого факта, будем использовать критерий согласия Хи-квадрат, чтобы обнаружить периоды *стабильности* весов, которые намного легче определить.

Критерий согласия Хи-квадрат является одним из наиболее часто используемых критериев проверки простых гипотез. Кратко приведем его описание. Критерий проверяет гипотезу о принадлежности выборки $x_1, x_2, \ldots, x_n$ некоторому заданному распределению $F(x, \theta)$, то есть гипотезу вида $H_0 : F_n(x) = F(x, \theta)$.

Подготовка к процедуре проверки состоит в следующем:

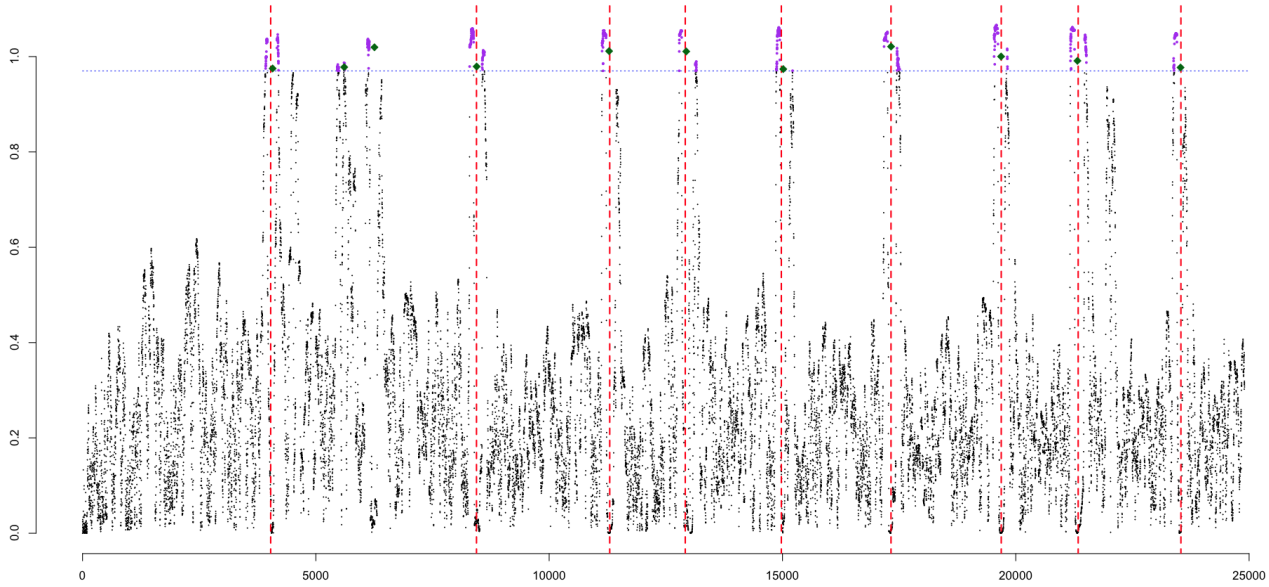

Рисунок 2.14: Определение движений с помощью динамической компоненты волатильности с помощью разницы векторов. Для наглядности, $z_i > \theta$ окрашены фиолетовым

1. область наблюдаемых значений разбивается на k непересекающихся интервалов точками $x_{(0)} \leqslant x_{(1)} \leqslant \ldots \leqslant x_{(k)}$;

2. для каждого интервала считается количество элементов выборки n_i попавших в этот интервал;

3. для каждого интервала считается теоретическая вероятность попадания в этот интервал $P_i(\theta) = F(x_{(i)}, \theta) - F(x_{(i-1)}, \theta)$.

Далее вычисляется статистика

$$X_n^2 = n \sum_{i=1}^{k} \frac{(n_i/n - P_i(\theta))^2}{P_i(\theta)},$$

где $n = \sum_{i=1}^{k} n_i$.

Если проверяемая гипотеза верна, то статистика X_n^2 в пределе имеет распределение χ_{k-1}^2. Этот факт позволяет принимать решение следующим образом: гипотеза H_0 отклоняется, когда заданный уровень значимости (P-значение или P-value)

$$p(X_n^2) = \frac{1}{2^{r/2}\Gamma(r/2)} \int_{X_n^2}^{\infty} s^{r/2-1} e^{-s/2} ds,$$

где $r = k-1$, меньше заранее заданного уровня значимости α (вероятность ошибки первого рода). Более подробное о критерии Хи-квадрат см., например, [14].

Вернемся к эксперименту. Для каждого окна вычислим P-значение критерия согласия Хи-квадрат с $k = 5$ корзинами. После этого выберем окна, где P-значение почти равно 1, то есть больше 0.9999. Запомним эти окна, см. Рис. 2.15 (отмечены фиолетовым).

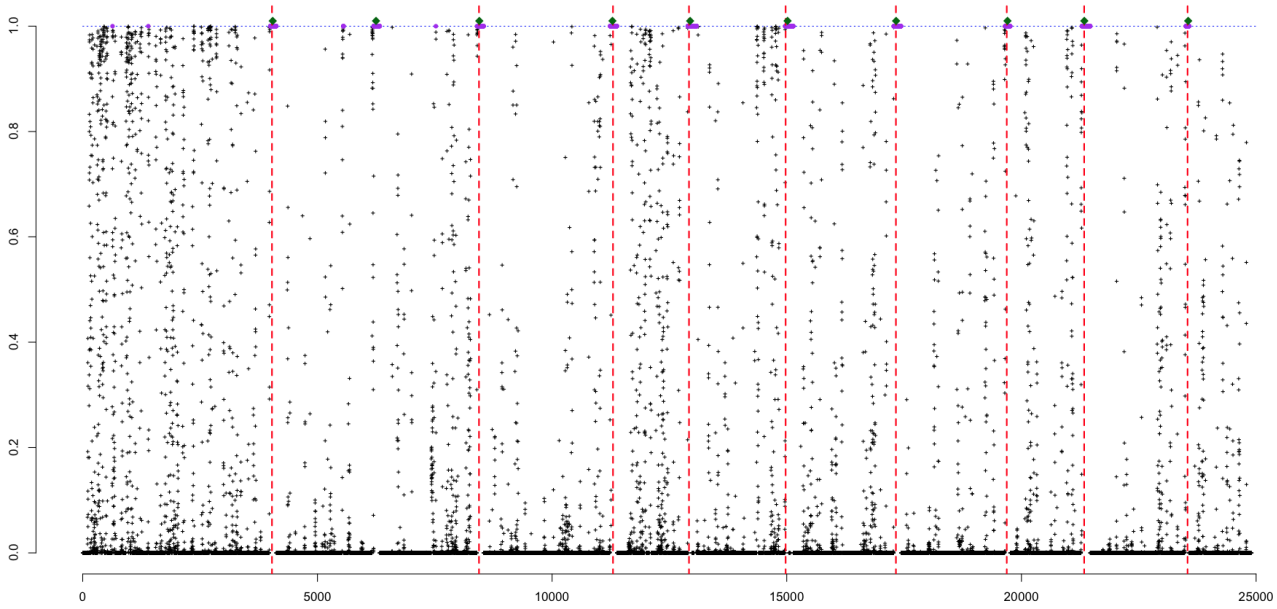

Рисунок 2.15: Определение движений с помощью динамической компоненты волатильности с помощью критерия согласия Хи-квадрат

Основной интерес представляют длинные периоды стабильности, а не случайные совпадения (следствиях шума), поэтому отбросим группы короче 50 миллисекунд. После того, как данная фильтрация сделана, возьмем первую точку в каждой группе. В предыдущем эксперименте было добавлено фиксированное значение, равное 150 к каждому событию. В этом же случае стабильный период обнаруживается только в одной группе (отражений нет), когда окно *находится полностью внутри* события. Поэтому к первому значению из группы нужно добавить только 50 миллисекунд, то есть размер окна, используемый для получения динамической составляющей.

Используя метод, описанный выше, были получены следущие точки (отмечены зеленым, сдвинуты выше фиолетовых интервалов для наглядности): 4048, 6249, 8450, 11286, 12941, 15017, 17330, 19703, 21340, 23553. Реальные события (красные линии): 4032, 8443, 11298, 12917, 14976, 17326, 19688, 21337, 23539.

Как и в предыдущем эксперименте, метод обнаруживает события между первым реальным событием и вторым, но в этом случае есть только одно ложное срабатывание, что означает, что этот метод работает лучше. Средняя точность оценки реальных событиях составляет ≈ 12 мс.

2.6.4 Определение начальных точек с помощью модифицированного метода непосредственно из миограммы

Как было упомянуто ранее, предложенный метод разделения дисперсионно-сдвиговых смесей был также применен непосредственно к исходным данным, а именно к временному ряду миограммы. В качестве семейств подбираемых распределений было использовано два класса, обсуждаемых ранее: обобщенное гиперболическое (GH-) и обобщенное дисперсионное гамма-распределения (GVG). Соответствующие функции распределения приведены в разделе 1.2.2 (см. (1.5)) и в разделе 1.2.3 (см. (1.7)).

Оба семейства распределений продемонстрировали хорошую аппроксимацию (fitness), но у GVG-распределений наблюдались лучшие P-значения при применении критерия согласия Хи-квадрат (см Рис. 2.17, 2.16). Цифры ниже показывают сравнение точности GVG-распределения и GH-распределения для двух случайно выбранных окон. На основании этого, для дальнейшего анализа было использовано только GVG-распределение.

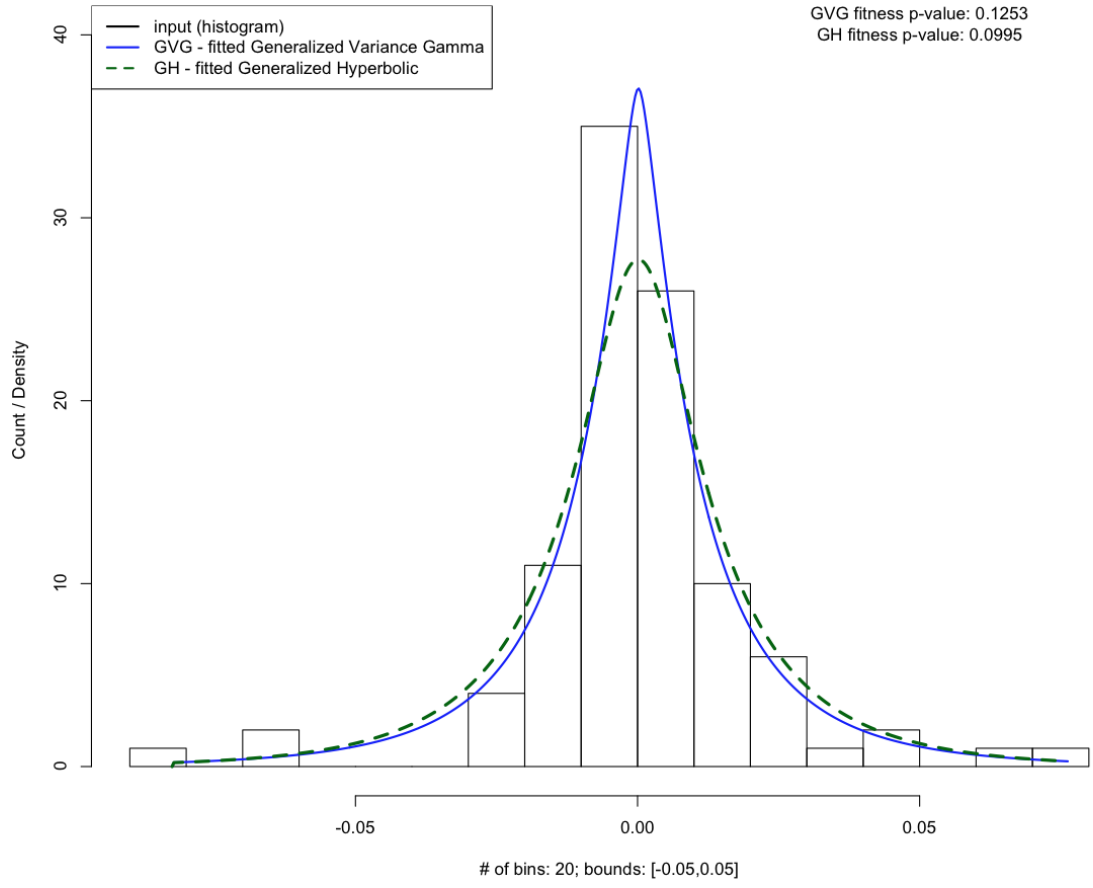

Рисунок 2.16: Сравнение аппроксимаций с помощью GH- и GVG-распределений, окно 19251

Параметр α представляет особый интерес. Он обладает волатильностью, но рядом с искомыми событиями имеет тенденцию сначала быстро уменьшаться, а затем быстро расти, постоянно демонстрируя большие абсолютные значения. Это значит, что можно определить события путем наблюдения за средними абсолютными значениями α.

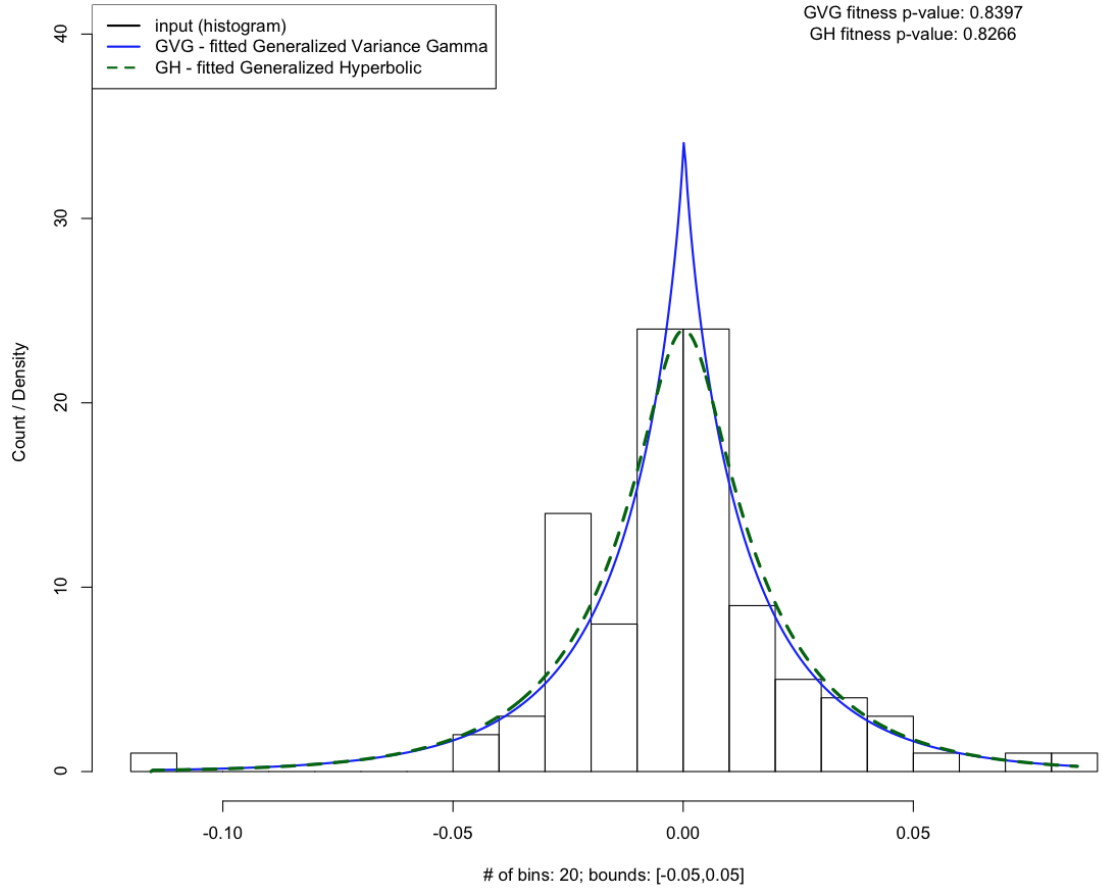

Рисунок 2.17: Сравнение аппроксимаций с помощью GH- и GVG-распределений, окно 10951

Используем порог 1.0 для того, чтобы фильтровать только большие значения (см. Рис. 2.18, отмечены фиолетовым). Как и в случаях выше, будем группировать данные и выбирать только первое значение в группе. Для финального значения надо вычесть 200 из выбранных первых значений (100 как исходный размер окна плюс 100 для вычисления среднего α).

В результате были получены следующие точки обнаружения: 3111, 4051, 4654, 7550, 7792, 8465, 11312, 12945, 15044, 17352, 19684, 21367. Реальные события (красные линии): 4032, 8443, 11298, 12917, 14976, 17326, 19688, 21337, 23539.

Как видно, «α-метод» обнаружил ложные события около первого и второго реальных событий. Также не было обнаружено последнее событие. Остальные события были найдены очень точно (≈ 25 мс).

Также могут быть использованы и другие метрики, рассчитанные для GVG-распределений. Любые числовые характеристики (моменты, квантили, асимметрия, эксцесс и т.д.) могут быть вычислены в явном виде по полученным параметрам и применены для обнаружения, если они демонстрируют определенное поведение до/после или в течение анализируемых событий.

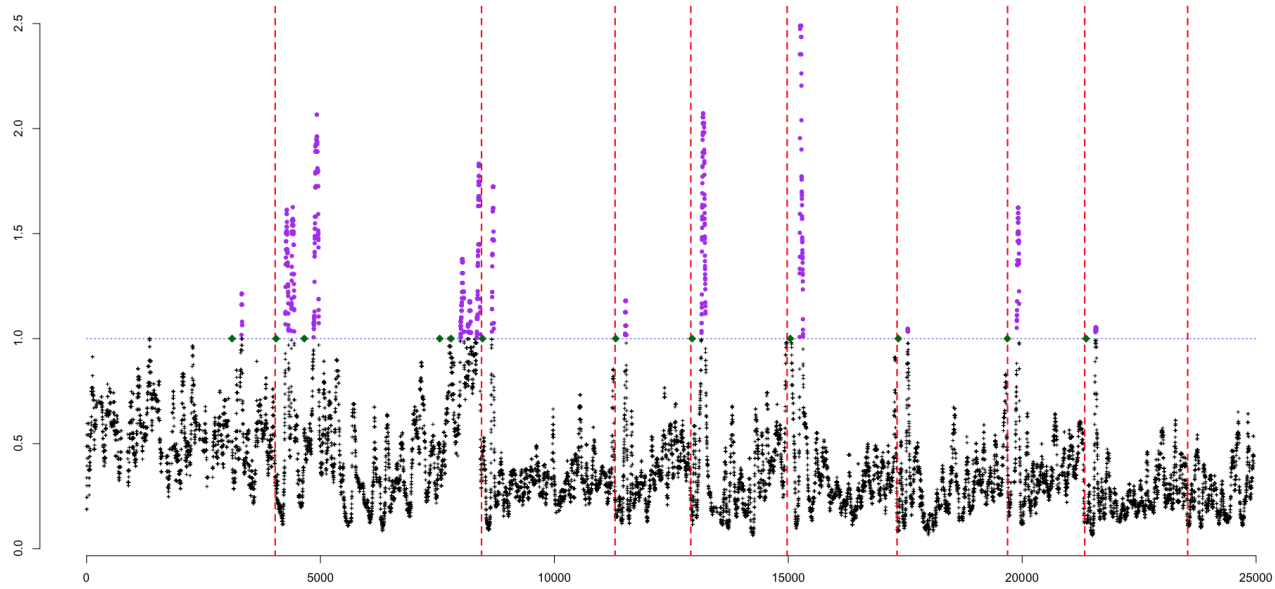

Рисунок 2.18: Определение движений с помощью параметра α GVG-распределения

2.7 Практические рекомендации при использовании метода

В данном разделе приведены советы и рекомендации по применению предложенного метода на основе опыта исследований его работы как на реальных, так и на искуственно сгенерированных данных.

2.7.1 Выбор оптимальных начальных параметров для запуска метода

Как было отмечено ранее, основными преимуществами предложенного метода по сравнению с традиционными EM-алгоритмами являются устойчивость к входным данным, высокая скорость работы и легкость интерпретации результатов.

Напомним, что перед запуском итерационного процесса первого шага алгоритма (2.3)-(2.4) требуется определить изначальные настройки, а именно:

- точность (критерий остановки);
- размерность K сетки оценки смешивающего распределения;
- верхняя граница $u*$ сетки оценки смешивающего распределения;
- начальные приближения коэффициента α и вектора p.

Выбор искомой точности обычно диктуется природой решаемой задачи и требованиями к скорости работы алгоритма. Тем не менее, не рекомендуется использовать точность

ниже, чем 1e-5, так как тестирование на искуственных выборках показало, что при определенных начальных приближениям искомые параметры не успевают «сойтись» к истинным значениям. Точность 1e-7 оказывалась достаточной для большинства исследованных практических задач, при этом время работы метода оставалось в пределах допустимых значений.

Выбор размерности сетки целиком зависит от количества доступной внешней информации, то есть от размера выборки. С одной стороны, нужно как можно точнее оценить форму смешивающего распределения (тем самым, чем больше в сетке узлов, тем лучше), но в то же время, имеется ограниченное количество входной информации, и число узлов в сетке должно быть значительно меньше размера выборки. К примеру, если исходная выборка содержит 1000 наблюдений, то выбирать сетку размерностью 250 бессмысленно. На практике для 1000 наблюдений хорошие результаты метод с сетками размерностью $K = 30$, $K = 40$; для выборок размерностью 10000 – $K = 70$, $K = 80$. Данные значения K показывают отличный баланс между точностью оценки и скоростью работы.

2.7.2 Выбор верхней границы сетки смешивающего распределения

В разделе 2.3 данной главы приведены оценки (2.10) и (2.15) для верхней границы $u*$ сетки, накидываемой на смешивающее распределение. Там же было отмечено, что данные оценки хоть и гарантируют, что почти весь носитель смешивающего распределения лежит внутри интервала $[0, u^*]$, тем не менее оценки являются завышенными.

С целью определения практических советов при использовании этих оценок в работе с реальными данными, проведем анализ их точности с использованием искуственно сгенерированных данных.

Для анализа адекватности оценок (2.10) и (2.15) был проведен анализ на 14 искуственно сгенерированных выборках обобщенных гиперболических (GH) распределений.

Зафиксируем $\varepsilon = 0.05$ и для каждой выборки вычислим оценки $u*$ по указанным формулам, далее сопоставим полученные оценки с истинным вычисленным значением $u*$. В таблице 2.1 приведены полученные результаты, в частности, отношения полученных оценок к истинным значениям в случаях, когда α известно, и когда α не известно.

Обратим особое внимание на то, во сколько раз полученные оценки отличаются от искомой. Результаты тестирования показывают, что для случая известного α полученную оценку можно делить на 2, а для неизвестного – на 8. При проверке на других искуственно сгенерированных распределениях (GVG) и при разделении реальных данных эффективность этого подхода подтвердилась.

Таблица 2.1: Результаты сравнения оценок $u*$ с истинными значениями для искуственно сгенерированных выборок

Выборка		$\varepsilon = 0.05$ A - оценка, α известно	B - оценка, α неизв.	C - истинное $u*$	A/C	B/C
$\alpha = 0.3$, β=0,	n=1000	85.1	446	37	2.3	12.06
ν =1.3, μ =1.6, λ =0.2	n=10000	81.4	379.8	37	2.2	10.27
$\alpha = 0.3$, β =0,	n=1000	5.4	37.2	4.4	1.22	8.49
ν =2, μ =2, λ =2.5	n=10000	10.4	45.7	2.36	4.4	10.41
$\alpha = 0.5$, β =0,	n=1000	7.2	25.7	2.6	2.78	9.91
ν =1, μ =1, λ =3	n=10000	5.8	23.7	2.6	2.25	9.11
$\alpha = 0.8$, β =0,	n=1000	10.6	51	4.3	2.45	11.79
ν =1.3, μ =1.6, λ =2	n=10000	10.3	47.2	4.3	2.37	10.9
$\alpha = 1.3$, β =0,	n=1000	10.6	58.9	4.4	2.41	13.41
ν =2, μ =2, λ =2.5	n=10000	10.7	55.9	4.4	2.43	12.73
$\alpha = 2$, β =0,	n=1000	10	59.7	4.3	2.32	13.79
ν =1.3, μ =1.6, λ =2	n=10000	10.2	65.3	4.3	2.35	15.09
$\alpha = 3$, β =0,	n=1000	83.1	3751.1	37	2.25	101.45
ν =1.3, μ =1.6, λ =0.2	n=10000	81.7	2899.3	37	2.21	78.41

2.7.3 Подход с использованием промежуточных результатов, многопроходность

Предложенная в предыдущем разделе схема выбора границ сетки является всего лишь рекомендацией, проверенной на практике. Бывают и исключительные ситуации, когда предложенная оценка не дает необходимой точности.

Тем не менее, при применении метода всегда можно легко понять, эффективны ли конкретные используемые настройки метода. Приведем несколько простых, но очень эффективных правил и советов по анализу *промежуточных результатов*, которые можно применять, не дожидаясь достижения заранее заданной точности:

- если несколько правых значений p_k одновременно близки к нулю (меньше заранее заданной значимости, скажем, 1e-9), их можно безболезненно убрать из рассмотрения, накинув новую сетку той же размерности от нуля до последнего значимого веса u_m, тем самым повысив точность при следующем проходе,

- если самое правое значение p_K значительно больше нуля (скажем, больше 0.01), это значит, что используемая верхняя граница недостаточно большая и ее необходимо расширить.

Данные советы помогают понять, насколько правильно был оценен носитель смешивающего распределения.

На практике оказывается исключительно полезно запустить второй, «уточняющий» проход алгоритма, использовав уже полученные данные первого прохода, в частности,

оценку коэффициента α как начальное приближение. Помимо этого, на втором проходе можно использовать полученные знания о сетке и о начальном приближение вектора p. При этом можно значительно повысить точность, не жертвуя временем выполнения.

Тем самым, за счет хороших начальных приближений и правильно выбранной сетки, накидываемой на носитель смешивающего распределения, можно за короткое время довольно сильно уточнить полученный ранее результат. Вышеизложенные идеи были реализованы для работы в автоматическом режиме и хорошо себя зарекомендовали на практике.

2.7.4 Адаптивный выбор сетки

Как было предложено в предыдущем разделе, эффективным оказывается многократный запуск метода на разных параметрах. В частности, к предложенному алгоритму можно применить техники, предложенные для общих сеточных методов в [5]. Основная идея заключается в следующем: мы не накладываем ограничений на расположение узлов сетки друг относительно друга. Другими словами, непересекающиеся отрезки, на которые мы разбиваем основную часть носителя, могут не быть одинаковой длины.

В работе [5] приводится пример двух хорошо работающих методов построения адаптивной сетки на втором шаге для сетки, накидываемой на математические ожидания и сетки, накидываемой на дисперсию, в случае сеточного аналога EM-алгоритма. В случае предложенного в этой главе метода имеется лишь одна сетка, накидываемая на носитель смешивающего распределения, и предложенные методы легко применить к построению лишь одной адаптивной сетки.

В частности, на практике хорошие результаты показал метод с "размещением"неиспользованных вершин. Подход состоит в следующем: сперва выделяются соседние участки носителя, где полученный вес компонент пренебрежимо мал - тем самым, данные участки можно объединить в один отрезок, высвободив какое-то количество дополнительных узлов. Первую процедуру назовем "чисткой". Вторая процедура называется "размещение и согласно названию, ее суть состоит в добавлении узлов на те участки, где на границах отрезка оба веса достаточно велики. Подобный подход позволяет очень точно описывать форму распределений, не жертвуя при этом скоростью работы т.к. количество узлов в сетке не увеличивается.

Глава 3

Метод прогнозирования финансовых рисков на основе разделения дисперсионно-сдвиговых смесей нормальных законов

3.1 Предварительные замечания. Основные определения

Многие игроки на финансовом рынке давно пришли к выводу, что анализировать и прогнозировать нужно не значения наблюдаемых процессов, а их *распределения*. При этом статистический анализ различных финансовых данных все чаще показывает, что очень многие исследуемые распределения отличаются наличием так называемых тяжелых хвостов. В частности, одной из важнейших практических задач является оценка этих хвостов, то есть задача оценки рисков.

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

Для упрощения задачи оценки и прогнозирования распределений часто используется подход снижения размерности путем априорного сужения классов допустимых смесей. В частности, модели, основанные на дисперсионно-сдвиговых смесях нормальных законов, показали высочайшую адекватность при решении практических задач, связанных с описанием эволюции различных финансовых индексов.

В данной главе описывается алгоритм прогнозирования параметров дисперсионно-сдвиговых смесей в общем виде (в частности, оценки рисков) на примере двух групп процессов: обобщенных гиперболических процессов (GH-processes) и обобщенных дисперсионных гамма-процессов (GVG-processes). Определение обобщенных гиперболических распределений было приведено в разделе 1.2.2 (см. (1.5)), обобщенных дисперсионных гамма-распределений приведено в разделе 1.2.3 (см. (1.7)).

3.2 Описание метода прогнозирования финансовых рисков и его свойства

Приведем описание метода на примере 5-параметрического семейства обобщенных гиперболических процессов. Необходимо отметить, что метод использует только сами значения параметров распределений, и никоим образом не учитывает вид самого семейства, поэтому данный метод без ограничения может быть применен к любому другому параметризованному семейству распределений.

Возьмем интересующий временной ряд и применим к нему стандартный подход разделения смесей нормальных законов с использованием скользящего окна. Для этого зафиксируем размер окна w и сдвиг окна $s <= w$. Далее на каждом окне применим модифицированный сеточный метод разделения дисперсионно-сдвиговых смесей нормальных законов, описанный в Главе 2.

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, $\theta_1, \theta_2, \ldots, \theta_N$, где каждое $\theta_i = (\alpha_i, \beta_i, \nu_i, \mu_i, \lambda_i)^T$.

Конечная задача прогнозирования заключается в получении оценок $\theta_{n+1}, \theta_{n+2}, \ldots$ для окон, которые будут частично или полностью состоять 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,
$F_j \in R^{5\times5}$ – матрицы-регрессоры.

Процедура поиска параметров модели имеет вид:

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

По сути это типичная регрессионная модель, где поиск матриц F_j проводится путем обучения модели с использованием минимизации суммарного RSS (остаточной суммы

квадратов) на $\hat{N} - R$ предсказаниях модели на известных данных. Для простоты обозначений, примем $N = \hat{N} - R$.

Соотношение (3.4) представляет из себя разновидность линейной регрессии и ее программная реализация не представляет труда. Более того, во многих статистических пакетах есть встроенные функции расчета матриц F_j для случаев $R = 1, 2$.

3.3 Подход к определению точности получаемых прогнозов

Остаточная сумма квадратов (RSS), полученная в результате поиска матриц F_j может служить оценкой качества модели, при условии ее деления на N, но эта метрика не учитывает самого вида распределений. При этом некоторые параметры распределений имеют гораздо больший вклад в вид итогового распределения, нежели оставшиеся.

С целью более точной оценки успеха построенного прогноза, рассмотрим вопрос близости непосредственно самого GH-распределения с полученными параметрами $\tilde{\theta}_{i+1}$ по отношению к истинному GH-распределению с параметрами θ_{i+1} для конкретных данных. Для этого введем несколько дополнительных метрик.

Обозначим для краткости $\tilde{F}(x)$ и $F(x)$ – функции распределения P_{GH} обобщенного гиперболического распределения с параметрами $\tilde{\theta}_{i+1}$ и θ_{i+1} соответственно, аналогично обозначим за $\tilde{f}(x)$ и $f(x)$ соответствующие плотности распределений.

3.3.1 Метрики C, L_1, L_2

Эти часто применяемые метрики для определения близости распределений имеют вид:

$$C(\tilde{f}, f) = sup|\tilde{f}(x) - f(x)|, x \in \mathbb{R}$$

$$L_1(\tilde{f}, f) = \int_{-\infty}^{+\infty} |\tilde{f}(x) - f(x)|dx$$

$$L_2^2(\tilde{f}, f) = \int_{-\infty}^{+\infty} (\tilde{f}(x) - f(x))^2 dx$$

Чем ниже абсолютные значения метрик C, L_1 или L_2, тем ближе распределения друг к другу.

3.3.2 Метрика «пересечения» плотностей (Intersect)

Эта метрика имеет смысл графического пересечения плотностей:

$$I(\tilde{f}, f) = 1 - \int_{-\infty}^{+\infty} \min(\tilde{f}(x), f(x))dx$$

Чем ближе Intersect к нулю, тем ближе распределения друг к другу.

3.3.3 Метрики, связанные с точностью оценки хвостов

Так как одним из самых важных в данном случае вопросов является оценка хвостов распределений, то есть рисков, целесообразно ввести метрики, которые позволят оценить именно качество прогнозирования хвостов. Вводить метрики будем исходя из необходимости найти два самых часто встречающихся при практическом анализе рисков интервала – область, где сосредоточено 90% распределения, а также область, где сосредоточено 95% распределения.

Рассмотрим метрики:

$$W_{0.05} = \tilde{F}(x_{0.05}), \quad W_{0.025} = \tilde{F}(x_{0.025}),$$

$$W_{0.95} = \tilde{F}(x_{0.95}), \quad W_{0.975} = \tilde{F}(x_{0.975}),$$

где x_α – квантиль уровня α для распределения $F(x)$. Из-за того, что GH-распределения является абсолютно непрерывным, можно переписать эти метрики в более удобной форме:

$$W_{0.05} = \tilde{F}(F^{-1}(0.05)), \quad W_{0.025} = \tilde{F}(F^{-1}(0.025)),$$

$$W_{0.95} = \tilde{F}(F^{-1}(0.95)), \quad W_{0.975} = \tilde{F}(F^{-1}(0.975)).$$

Чем ближе $T_{0.05}$ к значению 0.05, тем точнее оценен левый хвост. Аналогично для правого хвоста распределения – чем ближе $T_{0.95}$ к значению 0.95, тем точнее оценка.
Помимо оценки веса левого или правого хвоста имеет смысл рассмотреть, насколько точно соотносятся абсолютные значения самих квантилей распределений, то есть насколько точен прогноз границ хвостов:

$$S_{0.05} = |\tilde{x}_{0.05} - x_{0.05}|, \quad S_{0.025} = |\tilde{x}_{0.025} - x_{0.025}|,$$

$$S_{0.95} = |\tilde{x}_{0.95} - x_{0.95}|, \quad S_{0.975} = |\tilde{x}_{0.95} - x_{0.975}|,$$

или если использовать запись через существующую обратную функцию распределения:

$$S_{0.05} = |\tilde{F}^{-1}(0.05) - F^{-1}(0.05)|, \quad S_{0.025} = |\tilde{F}^{-1}(0.025) - F^{-1}(0.025)|,$$

$$S_{0.95} = |\tilde{F}^{-1}(0.95) - F^{-1}(0.95)|, \quad S_{0.975} = |\tilde{F}^{-1}(0.975) - F^{-1}(0.975)|.$$

Чем меньше значение метрик S_α, тем точнее было оценено положение «границ» хвостов.

Применяя описанные метрики при разных параметрах прогнозирования (порядок регрессии, количество исторических наблюдений для обучения модели), можно выбрать оптимальную для каждой конкретной задачи модель прогнозирования.

3.4 Результаты практического применения метода прогнозирования на реальных данных

В качестве исходных данных возьмем использованный ранее индекс KOSPI (Korea Composite Stock Price Index) - основной индикатор корейской биржи, см. раздел 2.5.1. Рассматриваются изменения логарифмов значений индекса с частотой (тиком) равной 1 минуте, начиная с открытия биржи 8 декабря 2014 года на протяжении трех рабочих дней. Размер окна установлен как 3 часа: $w = 180$, сдвиг окна минимальный, одно наблюдение: $s = 1$ тик.

Применив метод, предложенный в Главе 2, был получен входной ряд параметров $\theta_i = (\alpha_i, \beta_i, \nu_i, \mu_i, \lambda_i)^T$. В конкретной модели без ущерба для точности положено $\beta_i = 0$.

3.4.1 Описание процедуры прогнозирования, исходные выбранные модели

Среди всех окон (864) выделим точку во времени ($T = 500$), которую будем считать за текущий момент времени, границу известных данных. Данные до этой точки будут использоваться для обучения модели, данные после – для получения прогнозов и сравнения полученных результатов с историческими значениями.

Для анализа качества прогнозирования будем использовать метрики, посчитанные для следующих окон в будущем:

- $T + 1$ – окно, сдвинутое на 1 тик, минимальный прогноз в 1 минуту, 99% данных известны,
- $T + 10$ – прогноз в 10 тиков (10 минут), 94% наблюдений известны в момент T,
- $T + 60$ – прогноз в 60 тиков (1 час), 66% наблюдений известны в момент T,
- $T + 120$ – прогноз в 120 тиков (2 часа), 33% наблюдений известны в момент T,
- $T + 180$ – прогноз в 180 тиков (3 часа), 0% наблюдений известны в момент T.

Рис. 3.1 наглядно иллюстрирует рассматриваемую временную линию в данном эксперименте. На линии отмечен момент T, который будем считать текущим моментом времени.

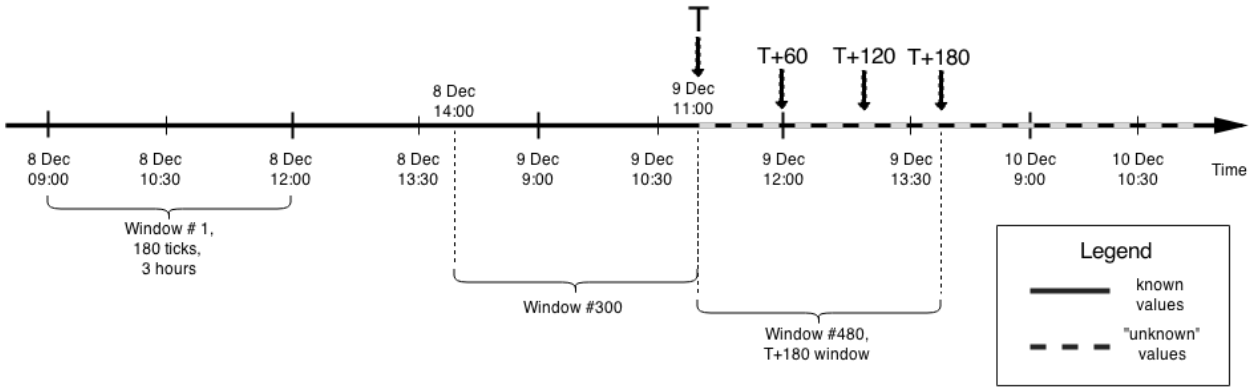

Рисунок 3.1: Временная линия в эксперименте KOSPI, местное время Korea Exchange

Таблица 3.1: Рассматриваемые модели, параметры R и N

	$N = 10$	$N = 20$	$N = 50$	$N = 100$	$N = 200$
$R = 1$ (16 коэфф.)	✓	✓	✓	✓	✓
$R = 2$ (32 коэфф.)			✓	✓	✓
$R = 3$ (48 коэфф.)				✓	✓

Все события левее от точки T являются доступными историческими данными, события правее будут прогнозироваться (а затем сравниваться с истинными значениями).

В качестве моделей будем использовать разные модели с $R = 1$, 2 и 3, то есть матричную регрессию первого, второго и третьего порядков.

Помимо использования регрессий разного порядка также будем изменять количество последних известных исторических наблюдений N, используемых для поиска параметров регрессии. Модели, использующие меньше исторических данных с большей вероятностью отразят недавние тренды в прогнозе, при этом они менее точны т.к. используют меньше входных данных.

Задача поиска матриц-регрессоров в данном случае является задачей поиска 16, 32 и 48 численных параметров для случаев $R = 1$, 2 и 3 соответственно. Учитывая количество неизвестных параметров, зададим количество используемых известных исторических наблюдений для обучения модели, см. 3.1.

3.4.2 Выбор лучшей модели с использованием фиксированного горизонта прогнозирования

После того, как были расчитаны необходимые коэффициенты для выбранного семейства моделей, зафиксируем наиболее актуальный горизонт прогнозирования: допустим, важными являются прогнозы на 1 час вперед, или используя введенное ранее обозначение, окно $T + 60$.

Таблица 3.2: Анализ качества прогноза в зависимости от модели, +1 час ($T + 60$). Метрики C, L_1, L_2, I

Модель			Прогноз $T + 60$			
R	N	RSS	C	L_1	L_2	I
1	10	0.583	0.080	0.106	0.070	5.68
1	20	0.377	0.091	0.118	0.079	6.23
1	50	0.350	**0.062**	**0.082**	**0.054**	**4.40**
1	100	**0.315**	0.123	0.159	0.106	8.36
1	200	0.920	0.286	0.355	0.235	18.05
2	50	**0.275**	0.130	0.169	0.112	8.81
2	100	0.291	**0.078**	**0.116**	**0.073**	**6.22**
2	200	0.903	0.272	0.340	0.225	17.33
3	100	**0.271**	**0.064**	**0.102**	**0.063**	**5.55**
3	200	0.758	0.269	0.332	0.221	16.94

Таблица 3.3: Анализ качества прогноза в зависимости от модели, +1 час ($T + 60$). Оценка хвостов распределений для интервалов содержащих 90%, 95%

			Прогноз $T + 60$							
Модель			90% интервал				95% интервал			
R	N	RSS	$W_{0.05}$	$W_{0.95}$	$S_{0.05}$	$S_{0.95}$	$W_{0.025}$	$W_{0.975}$	$S_{0.025}$	$S_{0.975}$
1	10	0.5838	0.0597	0.9350	0.1122	0.0512	0.0332	0.9688	0.1730	0.0335
1	20	0.3778	0.0556	0.9311	0.0658	0.0632	0.0305	0.9663	0.1195	0.0459
1	50	0.3508	**0.0550**	**0.9383**	**0.0570**	**0.0419**	0.0296	**0.9699**	0.0968	**0.0291**
1	100	0.3152	0.0584	0.9203	0.1021	0.0929	0.0331	0.9600	0.1776	0.0728
1	200	0.9201	0.0415	0.8662	0.1215	0.1938	**0.0235**	0.9275	**0.0406**	0.1616
2	50	0.2756	0.0521	0.9163	0.0263	0.1052	0.0288	0.9568	0.0853	0.0879
2	100	0.2918	0.0676	0.9340	0.2018	0.0549	0.0389	0.9680	0.2837	0.0375
2	200	0.9038	0.0417	0.8707	0.1171	0.1880	0.0235	0.9301	0.0401	0.1573
3	100	0.2717	0.0683	**0.9371**	0.2065	**0.0457**	0.0390	**0.9695**	0.2835	**0.0307**
3	200	0.7581	**0.0437**	0.8734	**0.0890**	0.1817	**0.0248**	0.9324	**0.0060**	0.1493

В таблицах 3.2, 3.3 приведен сравнительный анализ расчетов основных метрик для различных моделей.

Для удобства анализа полученных результатов жирным шрифтом выделены лучшие значения метрик в каждой из групп $R = 1$, 2 и 3.

Обратим внимание, что по полученным значениям можно сделать вывод, что в группах $R = 1$ и $R = 2$ модель с минимальным покомпонентным RSS не является наилучшей с точки зрения оценки интересующих показателей, что подтверждает разумность ввода дополнительных метрик. Среди всех моделей выбор падет на две: $R = 1, N = 50$; и $R = 1, N = 100$ – соответствующие ряды в таблице подсвечены серым цветом. Исходя из расчитанных метрик наилучшей моделью в данном случае является регрессия первого порядка: $R = 1, N = 50$.

Таблица 3.4: Анализ качества прогноза в зависимости от горизонта прогнозирования, $R = 1, N = 50$. Метрики C, L_1, L_2 и I

R	N	Горизонт	C	L_1	L_2	I
1	50	T + 1	0.0066	0.0071	0.0050	0.62%
1	50	T + 10	0.0495	0.0570	0.0397	3.13%
1	50	T + 20	0.0308	0.0462	0.0277	2.54%
1	50	T + 30	0.0288	0.0340	0.0234	1.97%
1	50	T + 40	0.1052	0.1416	0.0947	7.44%
1	50	T + 50	0.0803	0.1233	0.0790	6.59%
1	50	T + 60	0.0626	0.0821	0.0548	4.40%
1	50	T + 70	0.0863	0.1197	0.0788	6.32%
1	50	T + 80	0.0249	0.0413	0.0237	2.29%
1	50	T + 90	0.0414	0.0551	0.0362	3.01%
1	50	T + 100	0.0445	0.0593	0.0384	3.21%
1	50	T + 110	0.0307	0.0494	0.0287	2.74%
1	50	T + 120	0.0515	0.0765	0.0463	4.08%
1	50	T + 130	0.0743	0.1029	0.0656	5.37%
1	50	T + 140	0.1328	0.1810	0.1157	9.27%
1	50	T + 150	0.0997	0.1511	0.0910	7.82%
1	50	T + 160	0.1429	0.2063	0.1280	10.55%
1	50	T + 170	0.1779	0.2520	0.1581	12.80%
1	50	T + 180	0.1639	0.2385	0.1483	12.13%

3.4.3 Анализ точности прогнозирования и особенностей полученной модели

Займемся изучением вопроса точности прогнозирования для выбранной модели $R = 1, N = 50$ в зависимости от горизонта прогнозирования. Для выбранных точек в будущем построим соответствующий прогноз и исследуем соответствующие метрики, см. таблицы 3.4, 3.5.

Как и ожидалось, наблюдается рост несоответствия между прогнозом и истинными значениями с расширением горизонта прогнозирования. На Рис. 3.2 – 3.5 показано изменение истинного и прогнозируемого распределений с расширением горизонта прогнозирования, а так же приведены значения параметров указанных распределений.

Отдельный интерес представляет сравнение динамики самих параметров распределения прогноза и истинных исторических значений, без использования самого вида распределения. На Рис. 3.6 показано соотношение между прогнозом параметра λ (после $T = 300$) и истинными значениями.

На основании полученных данных можно сделать вывод, что найденная описанным методом модель хорошо подходит для прогнозирования как минимум на 2 часа вперед или меньшие интервалы. Важно отметить, что найденная модель достаточно точно оценивает хвосты распределений, позволяя нам полагаться на полученные прогнозы при оценке рисков.

Таблица 3.5: Анализ качества прогноза в зависимости от горизонта прогнозирования, $R = 1, N = 50$. Метрики оценки хвостов

			90% Интервал				95% Интервал			
R	N	Forecast	$W_{0.05}$	$W_{0.95}$	$S_{0.05}$	$S_{0.95}$	$W_{0.025}$	$W_{0.975}$	$S_{0.025}$	$S_{0.975}$
1	50	T + 1	0.0499	0.9494	0.0011	0.0021	0.0250	0.9749	0.0010	0.0001
1	50	T + 10	0.0515	0.9434	0.0174	0.0224	0.0268	0.9727	0.0406	0.0129
1	50	T + 20	0.0431	0.9469	0.0878	0.0110	0.0210	0.9739	0.0984	0.0066
1	50	T + 30	0.0512	0.9467	0.0143	0.0118	0.0263	0.9741	0.0293	0.0051
1	50	T + 40	0.0616	0.9276	0.1244	0.0746	0.0352	0.9646	0.1959	0.0553
1	50	T + 50	0.0709	0.9380	0.2111	0.0423	0.0416	0.9715	0.2955	0.0203
1	50	T + 60	0.0550	0.9383	0.0570	0.0419	0.0296	0.9699	0.0968	0.0291
1	50	T + 70	0.0595	0.9301	0.1056	0.0694	0.0333	0.9652	0.1664	0.0536
1	50	T + 80	0.0429	0.9490	0.0928	0.0038	0.0208	0.9752	0.1062	0.0013
1	50	T + 90	0.0505	0.9415	0.0062	0.0320	0.0261	0.9709	0.0256	0.0243
1	50	T + 100	0.0502	0.9598	0.0021	0.0422	0.0243	0.9804	0.0159	0.0373
1	50	T + 110	0.0556	0.9575	0.0662	0.0319	0.0281	0.9794	0.0695	0.0301
1	50	T + 120	0.0553	0.9620	0.0624	0.0534	0.0275	0.9818	0.0556	0.0491
1	50	T + 130	0.0489	0.9671	0.0141	0.0809	0.0230	0.9845	0.0496	0.0733
1	50	T + 140	0.0486	0.9758	0.0183	0.1349	0.0218	0.9887	0.0821	0.1184
1	50	T + 150	0.0595	0.9723	0.1103	0.1137	0.0292	0.9874	0.0927	0.1055
1	50	T + 160	0.0552	0.9789	0.0625	0.1611	0.0255	0.9906	0.0121	0.1462
1	50	T + 170	0.0458	0.9833	0.0546	0.2009	0.0193	0.9925	0.1529	0.1786
1	50	T + 180	0.0468	0.9827	0.0412	0.1984	0.0201	0.9924	0.1305	0.1790

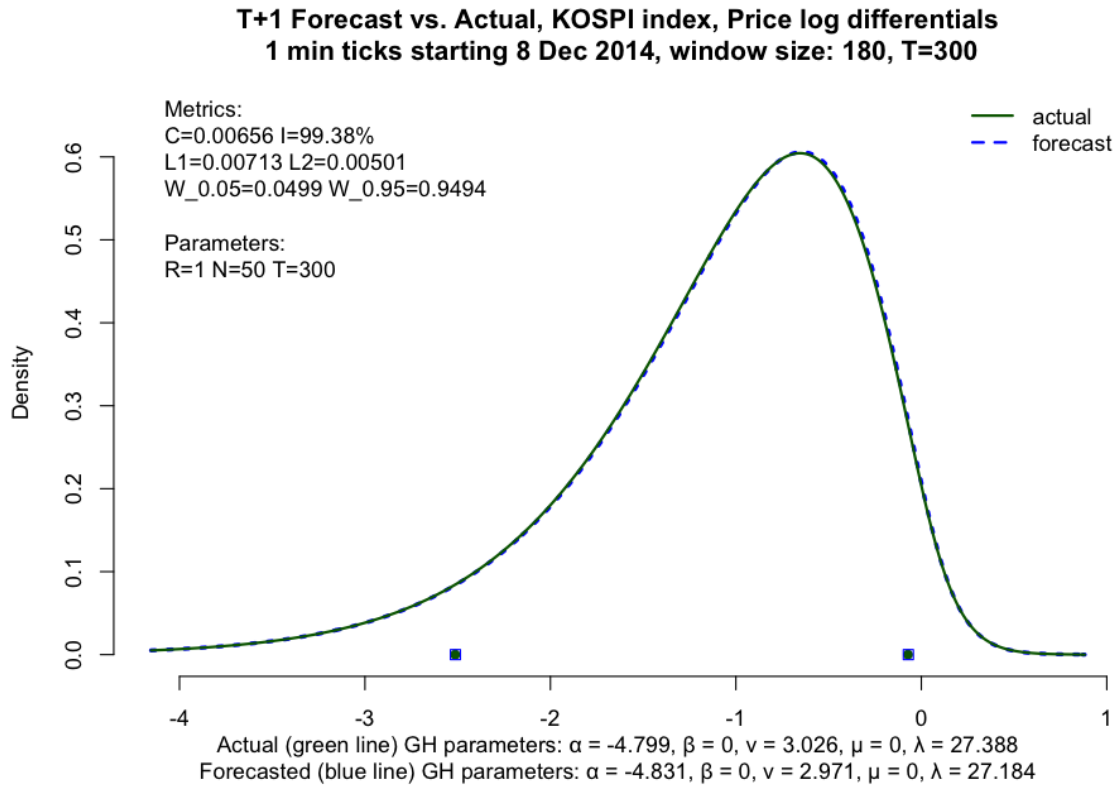

Рисунок 3.2: Прогнозируемое и истинные распределения, горизонт 1 минута, $T + 1$

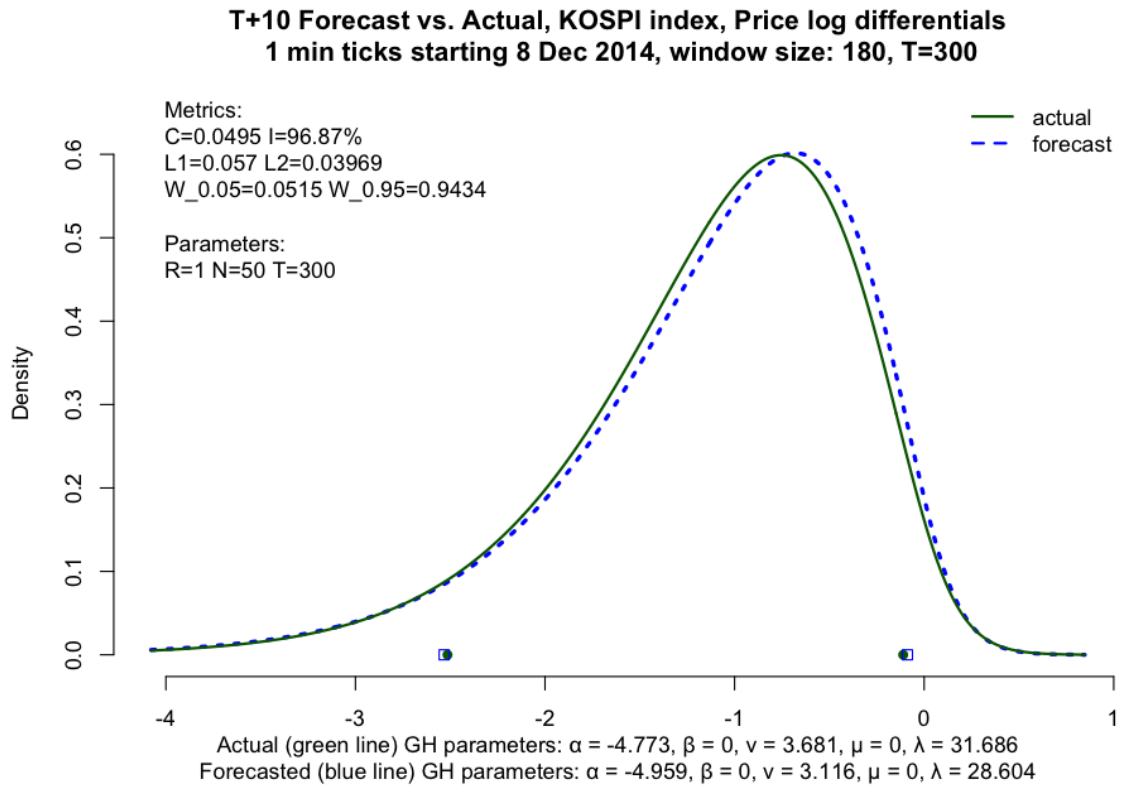

Рисунок 3.3: Прогнозируемое и истинные распределения, горизонт 10 минут, $T + 10$

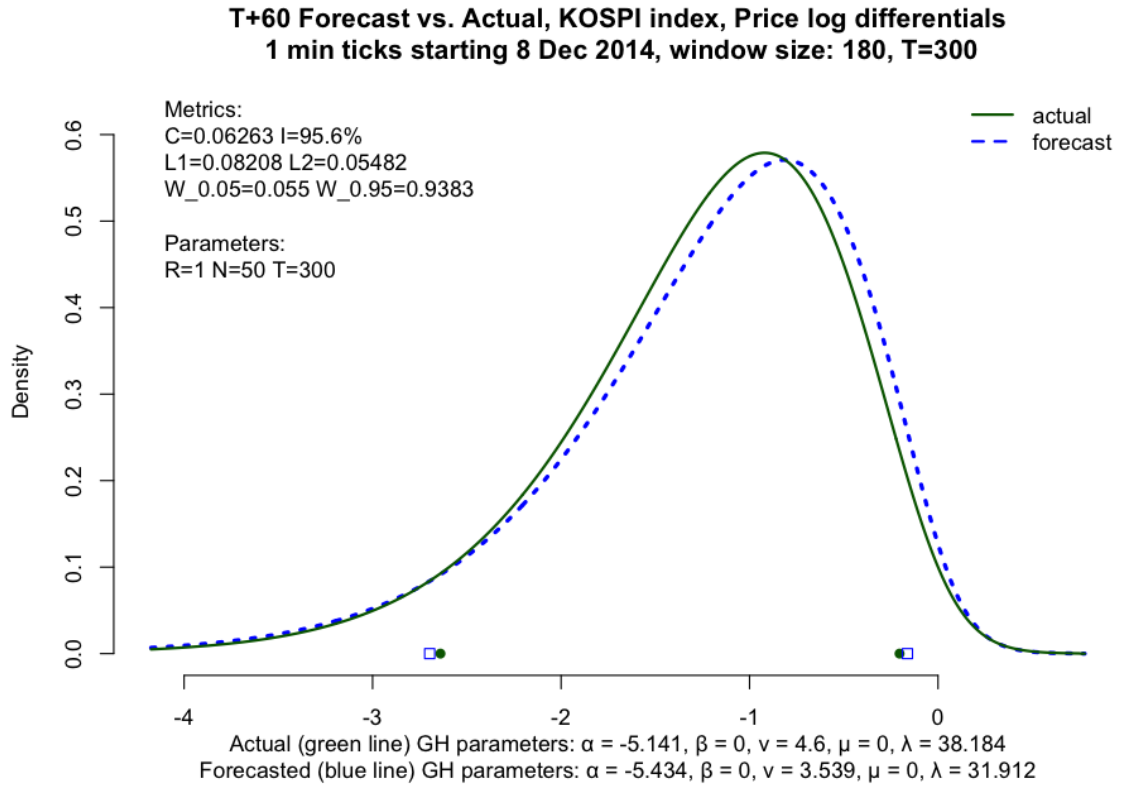

Рисунок 3.4: Прогнозируемое и истинные распределения, горизонт 1 час, $T + 60$

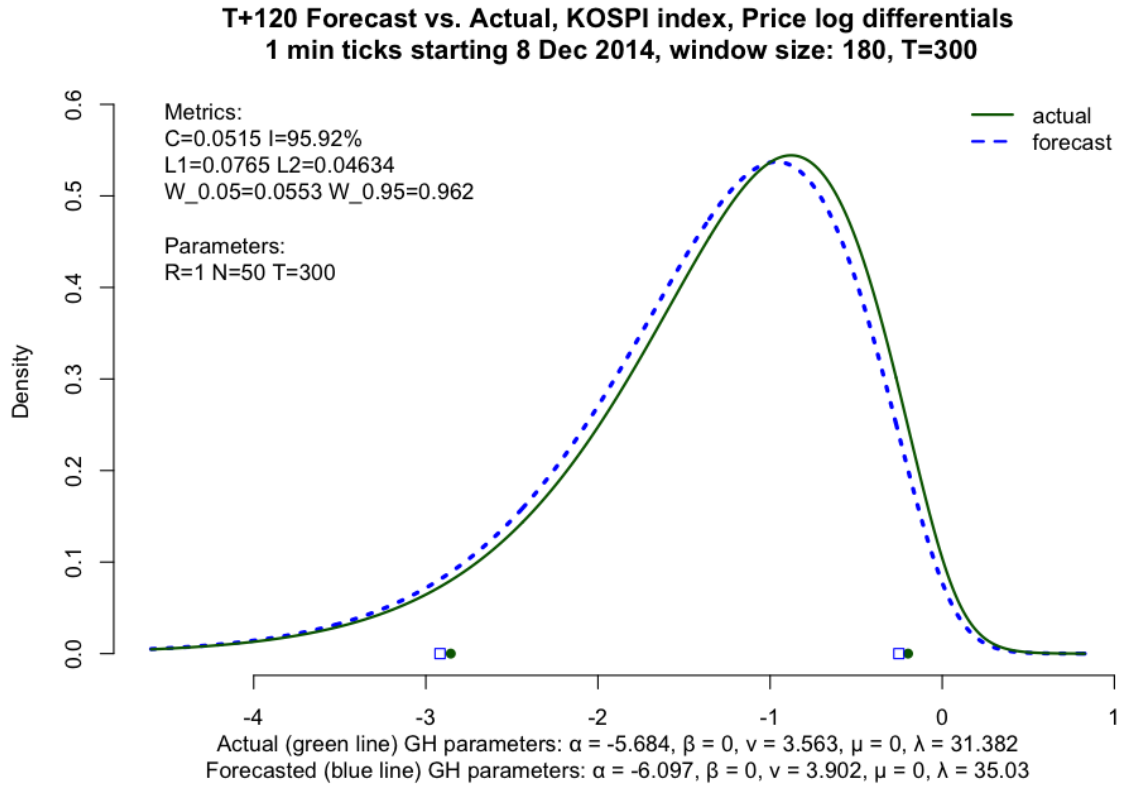

Рисунок 3.5: Прогнозируемое и истинные распределения, горизонт 2 часа, $T + 120$

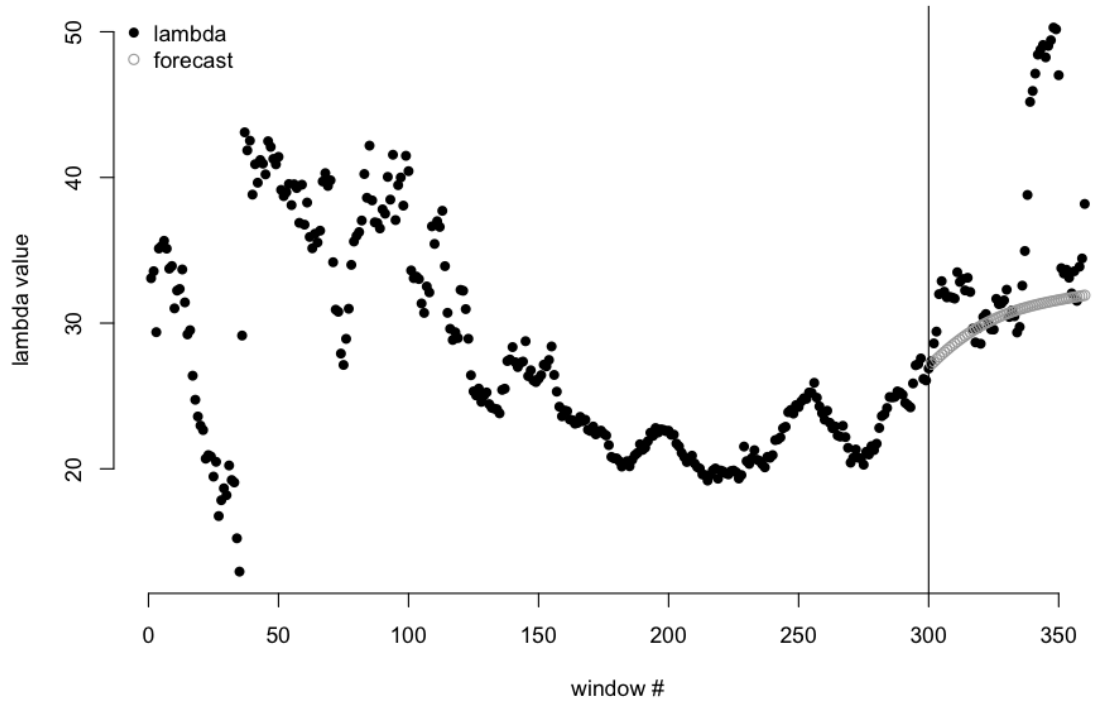

Рисунок 3.6: Прогнозируемое и истинные значения параметра λ

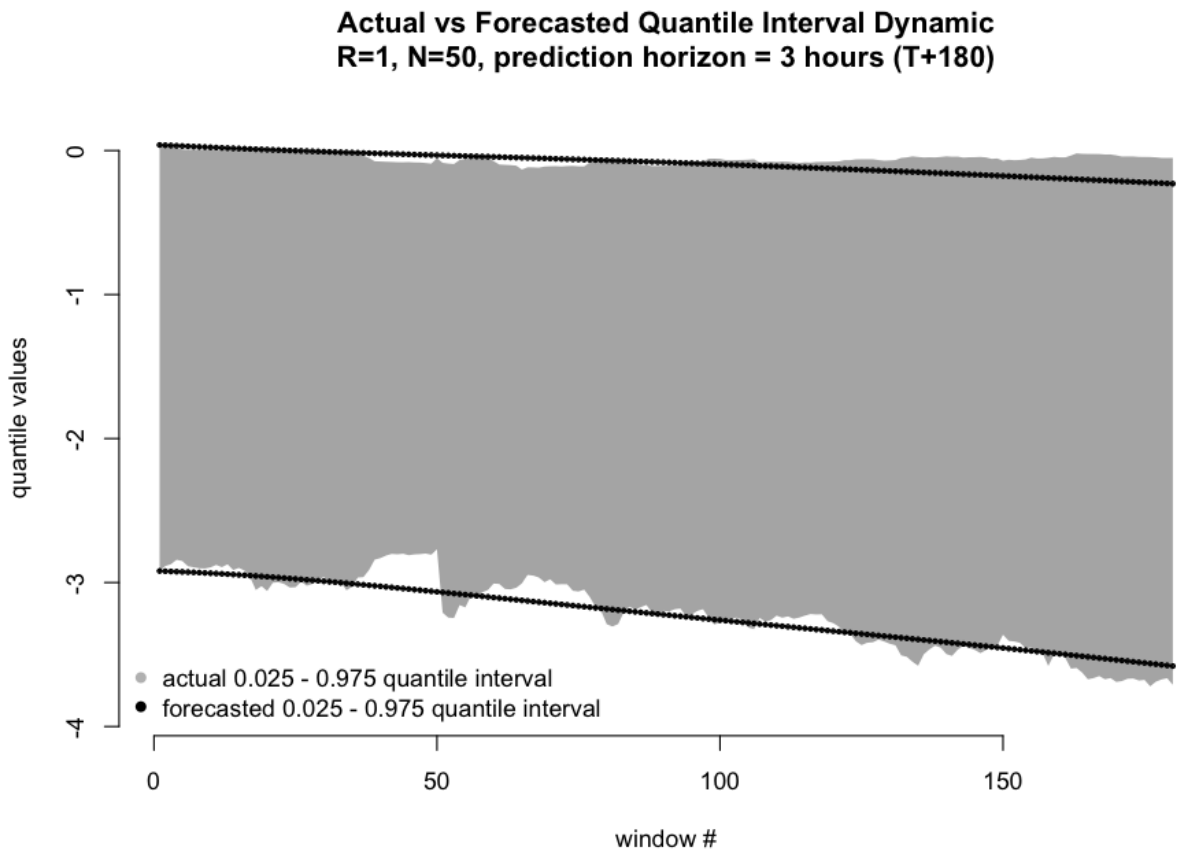

Рисунок 3.7: Прогнозируемое и истинные значения квантильного интервала $(x_{0.025}, x_{0.975})$ (интервал содержит 95% распределения)

При использовании модели на поступающих в реальном времени данных, постепенно становятся доступными новые исходные параметры, что позволяет проводить повторное обучение модели пересчитывая регрессионные матрицы. Это особенно удобно, потому что поиск необходимых матриц происходит очень быстро и не представляет большой вычислительной сложности.

3.4.4 Прогнозирование интерквантильных интервалов

Особый интерес в задаче оценивания и прогнозирования рисков, как было отмечено ранее, представляет оценка и прогнозирование интерквантильных интервалов.

На Рис. 3.7 отдельно представлена динамика квантильного интервала $(x_{0.025}, x_{0.975})$ (95% распределения) и соответствующего прогноза во времени после момента $T = 300$ на протяжении трех часов.

Данный рисунок иллюстрирует высокую точность прогнозирования хвостов при использовании предложенного метода.

Можно получить интересную картину, если непрерывно следить за получаемым прогнозом начиная с одной минуты и вплоть до трех часов, постоянно сравнивая его с историческими распределениями. Удобно наблюдать за этой эволюцией в формате видео или ускоренного слайд-шоу. Пример для модели выше доступен по ссылке: git.io/b4jA

3.4.5 Прогнозирование значений наблюдаемого процесса

Как было отмечено ранее, предложенный метод прогнозирует *распределение* случайного процесса, а не непосредственно будущие наблюдения. Если необходимо прогнозировать сами значения исследуемого процесса, в качестве прогноза можно использовать либо медиану (в случае квадратичной функции потерь), либо математическое ожидание (в случае линейной функции потерь) полученного распределения.

Рис. 3.8, 3.9 иллюстрируют предложенный подход в разных масштабах. Как видно на рисунках, только 14 из 180 истинных значений (меньше 8%) находятся за пределами *прогнозируемого* 95% интерквантильного интервала для трех часов прогнозирования.

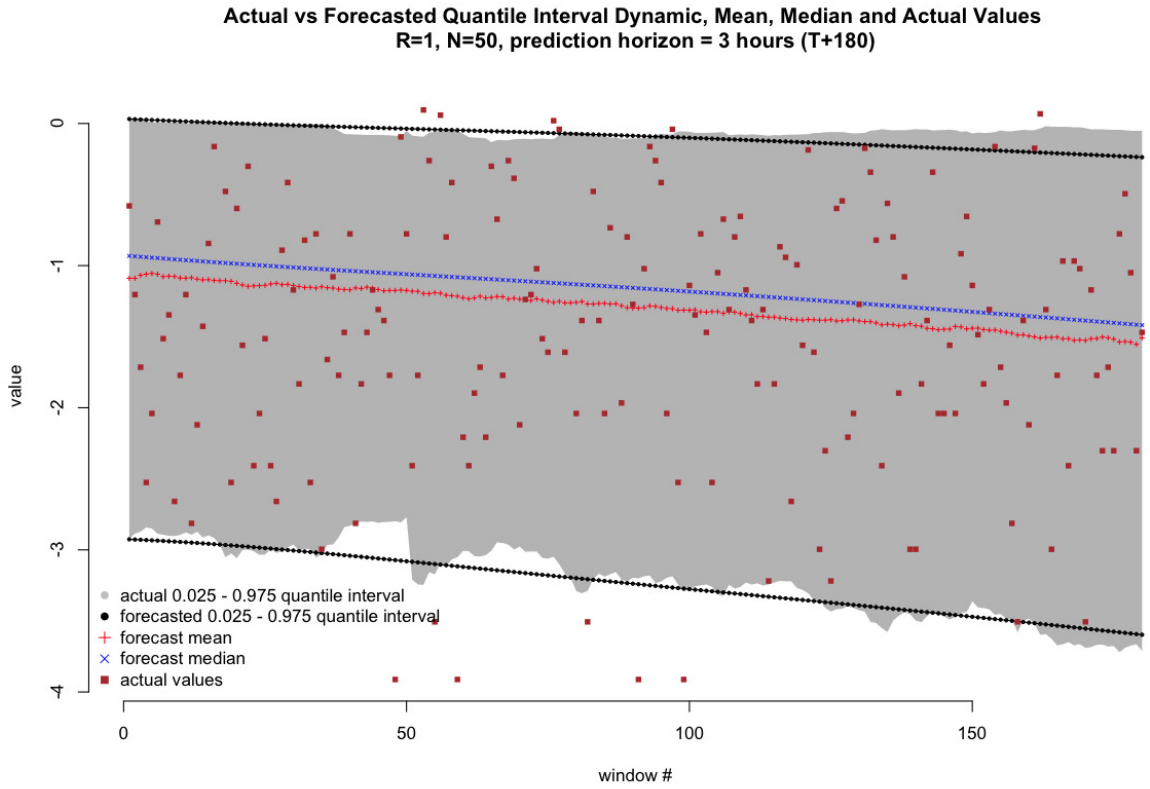

Рисунок 3.8: Истинные и предсказанные значения начиная с момента $T = 300$

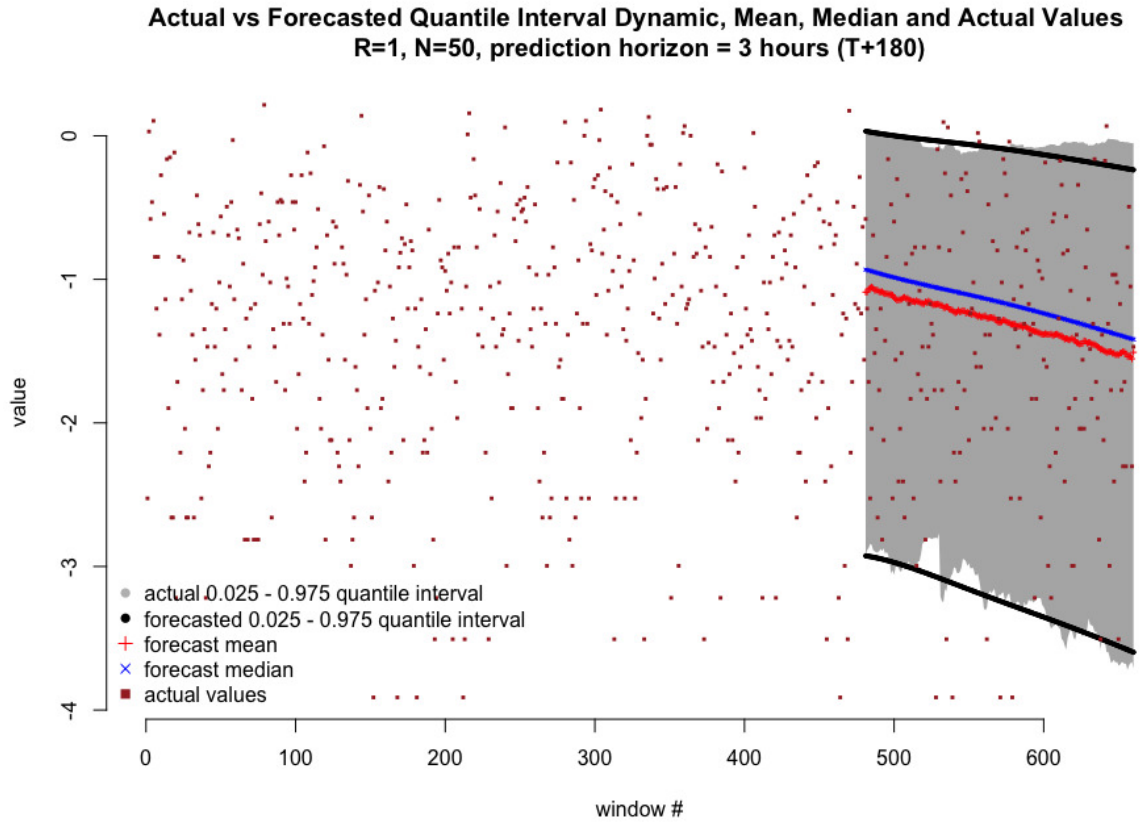

Рисунок 3.9: Истинные и предсказанные значения начиная с момента $T = 1$

3.4.6 Дальнейшие шаги по улучшению предложенного метода

Предложенный метод прогнозирования можно продолжать улучшать во многих направлениях. В частности, если заранее известно, какая информация о прогнозируемом распределении важнее всего для исследователя, то можно сразу заниматься минимизацией нужной метрики для поиска матриц F_j вместо минимизации RSS, то есть использовать другое соотношение (2). При этом надо отметить, что задача поиска минимума вероятно перестанет быть задачей линейного программирования, тем самым требуя существенно больше вычислительного времени и ресурсов.

Одним из потенциальных методов решения проблемы недостатка вычислительного времени является увеличение сдвига окна $s = 1$ при расчетах входных параметров. При увеличении сдвига уменьшается частота, с которой нужно расчитывать все параметры, но, к сожалению, снижается гладкость входного ряда. Важно обратить внимание, что при увеличении сдвига окна информация о самих входных данных практически не теряется, так как все наблюдения все же попадают в какие-то окна, но при этом тестировать алгоритмы разложения и прогнозирования рекомендуется на гладких окнах, чтобы убедиться в их устойчивости к незначительным колебаниям исходных данных.

Помимо этого, вполне разумно предложить использовать более сложные модели, чем матричная регрессия порядка R, следуя тем же самым принципам обучения через исто-

рические данные, описанным в данной работе, и выбирать подходящую модель исходя из интересующих метрик.

3.5 Дополнительная валидация результатов

3.5.1 Выбор альтернативной начальной точки

Приведенный в предыдущем разделе анализ показывает, что полученную модель можно было использовать для прогнозирования распределений до двух часов вперед, оставаясь в пределах нужной точности прогнозирования (≈ 0.05 для оценки границ хвостов).

Убедимся, что ту же самую модель можно применить и в другой точке отсчета, допустим, не $T = 300$, а $T = 100$, $T = 500$, $T = 700$ и другие.

Так как матрицы-регрессоры вычисляются очень быстро, произведя подобную проверку могут быть получены дополнительные гарантии качества модели при минимальных приложенных усилиях.

Продемонстрируем эффективность выбранной модели $R = 1, N = 50$ на примере $T = 100$. В таблице 3.5 приведены соответствующие метрики.

Таблица 3.6: Анализ качества прогноза в зависимости от горизонта прогнозирования, $R = 1, N = 50$, $T = 100$

			90% интервал				95% интервал			
R	N	Прогноз	$W_{0.05}$	$W_{0.95}$	$S_{0.05}$	$S_{0.95}$	$W_{0.025}$	$W_{0.975}$	$S_{0.025}$	$S_{0.975}$
1	50	T + 1	0.0439	0.9588	0.0752	0.0383	0.0205	0.9797	0.1083	0.0330
1	50	T + 10	0.0427	0.9478	0.0938	0.0087	0.0207	0.9740	0.1064	0.0061
1	50	T + 20	0.0358	0.9490	0.1956	0.0038	0.0166	0.9748	0.2309	0.0014
1	50	T + 30	0.0326	0.9671	0.2500	0.0756	0.0136	0.9837	0.3409	0.0625
1	50	T + 40	0.0432	0.9665	0.0866	0.0721	0.0196	0.9834	0.1379	0.0601
1	50	T + 50	0.0344	0.9661	0.2177	0.0699	0.0146	0.9830	0.2995	0.0566
1	50	T + 60	0.0394	0.9730	0.1390	0.1070	0.0169	0.9867	0.2200	0.0906
1	50	T + 70	0.0355	0.9750	0.2005	0.1188	0.0143	0.9871	0.3109	0.0945
1	50	T + 80	0.0412	0.9800	0.1128	0.1524	0.0168	0.9898	0.2219	0.1247
1	50	T + 90	0.0472	0.9809	0.0340	0.1585	0.0198	0.9901	0.1309	0.1286
1	50	T + 100	0.0502	0.9797	0.0018	0.1492	0.0217	0.9895	0.0807	0.1207
1	50	T + 110	0.0426	0.9798	0.0940	0.1499	0.0172	0.9895	0.2098	0.1204
1	50	T + 120	0.0337	0.9778	0.2292	0.1363	0.0132	0.9888	0.3540	0.1124

Важно отметить, что для расчета метрик в таблице выше были использованы абсолютно другие матрицы-регрессоры, посчитанные на данных, абсолютно не пересекающихся с

Таблица 3.7: Рассматриваемые модели для DJI, параметры R и N

	$N = 20$	$N = 30$	$N = 40$
$R = 1$ (16 коэфф.)	✓	✓	✓
$R = 2$ (32 коэфф.)		✓	✓

Таблица 3.8: Анализ качества прогноза в зависимости от модели, +10 минут ($T + 10$)
Метрики C, L_1, L_2, I.

Модель			Прогноз $T + 10$			
R	N	RSS	C	L_1	L_2	I
1	20	12.26	0.0126	0.0363	0.0161	2.06%
1	30	10.35	0.0116	0.0351	0.0157	1.99%
1	40	9.48	0.0105	0.0293	0.0130	1.69%
2	30	7.45	0.0131	0.0367	0.0162	2.07%
2	40	8.59	**0.0098**	**0.0274**	**0.0123**	**1.59%**

данными для случая $T = 300$. Тем не менее, точность оценки хвостов остается на хорошем уровне.

3.5.2 Применение метода прогнозирования на данных Dow Jones Industrial

Применим предложенный метод прогнозирования на другом финансовом показателе – ранее исследованном индексе Dow Jones Industrial.

На Рис. 2.12 в Главе 2 показано, как выглядит GVG-приближение начала дня (первые ≈ 3 часа, 156 тиков) в динамике.

Исходя из размера выбранного окна будем рассматривать модели, приведенные в 3.7.

Зафиксируем начальный момент времени $T = 90$. Возьмем три горизонта прогнозирования: $T+10$ (10 минут), $T+30$ (полчаса), $T+60$ (один час). Для каждого из них построим прогнозы с использованием выбранных моделей и выберем лучшую модель. Результаты приведены в таблицах 3.8 – 3.12.

Анализ таблицы 3.12 показывает, что прогноз на час вперед не дает необходимой точности оценки хвостов. Обратим внимание на столбцы метрик $S_{0.025}$, $S_{0.975}$ – неточность довольно высока.

Вследствие этого сконцентрируемся на более краткосрочном прогнозе в пределах получаса. Для выбора лучшей модели проанализируем таблицы 3.8 – 3.10. Заметим, что модель с параметрами $R = 2, N = 40$ дает абсолютно лучший результат в прогнозировании на 10 минут вперед и демонстрирует неплохие результаты при прогнозировании на полчаса вперед, а именно лучше всех оценивает левый хвост и вторая по точности оценивает правый. Исходя из этого, модель $R = 2, N = 40$ и будет финальным выбором. Данная модель является регрессией второго порядка.

Таблица 3.9: Анализ качества прогноза в зависимости от модели, +10 минут (T+10). Оценка хвостов распределений для интервалов содержащих 90%, 95%

			Прогноз $T+10$							
Модель			90% интервал				95% интервал			
R	N	RSS	$W_{0.05}$	$W_{0.95}$	$S_{0.05}$	$S_{0.95}$	$W_{0.025}$	$W_{0.975}$	$S_{0.025}$	$S_{0.975}$
1	20	12.26	0.0582	0.9444	0.0928	0.0686	0.0307	0.9707	0.1136	0.0951
1	30	10.35	0.0585	**0.9465**	0.0948	**0.0436**	0.0308	**0.9719**	0.1152	**0.0703**
1	40	9.48	0.0549	0.9437	0.0562	0.0749	0.0285	0.9706	0.0709	0.0939
2	30	7.45	0.0566	0.9423	0.0750	0.0921	0.0297	0.9695	0.0942	0.1170
2	40	8.59	**0.0535**	0.9435	**0.0402**	0.0772	**0.0276**	0.9706	**0.0528**	0.0942

Таблица 3.10: Анализ качества прогноза в зависимости от модели, +30 минут ($T+30$) Метрики C, L_1, L_2, I.

Модель			Прогноз $T+30$			
R	N	RSS	C	L_1	L_2	I
1	20	12.26	0.0361	0.0943	0.0435	5.03%
1	30	10.35	**0.0349**	0.0947	0.0442	5.06%
1	40	9.48	0.0346	**0.0889**	**0.0413**	**4.72%**
2	30	7.45	0.0435	0.1201	0.0576	6.29%
2	40	8.59	0.0405	0.1143	0.0552	5.97%

Таблица 3.11: Анализ качества прогноза в зависимости от модели, +10 минут (T+10). Оценка хвостов распределений для интервалов содержащих 90%, 95%

			Прогноз $T+30$							
Модель			90% интервал				95% интервал			
R	N	RSS	$W_{0.05}$	$W_{0.95}$	$S_{0.05}$	$S_{0.95}$	$W_{0.025}$	$W_{0.975}$	$S_{0.025}$	$S_{0.975}$
1	20	12.26	0.0712	0.9357	0.2175	0.1626	0.0397	0.9649	0.2580	0.2016
1	30	10.35	0.0743	**0.9400**	0.2455	**0.1176**	0.0417	**0.9673**	0.2886	**0.1603**
1	40	9.48	0.0628	0.9316	0.1354	0.1994	0.0341	0.9630	0.1663	0.2269
2	30	7.45	0.0551	0.9206	0.0565	0.2975	0.0294	0.9564	0.0853	0.3233
2	40	8.59	**0.0522**	0.9229	**0.0245**	0.2739	**0.0274**	0.9582	**0.0473**	0.2928

Таблица 3.12: Анализ качества прогноза в зависимости от модели, +1 час ($T+60$) Метрики C, L_1, L_2, I.

Модель			Прогноз $T+60$			
R	N	RSS	C	L_1	L_2	I
1	20	12.26	0.0363	0.0984	0.0467	5.20%
1	30	10.35	**0.0298**	**0.0776**	**0.0354**	**4.19%**
1	40	9.48	0.0522	0.1715	0.0793	8.84%
2	30	7.45	0.1156	0.4164	0.1861	21.20%
2	40	8.59	0.1146	0.4133	0.1849	21.02%

Таблица 3.13: Анализ качества прогноза в зависимости от модели, +10 минут (T+10). Оценка хвостов распределений для интервалов содержащих 90%, 95%

			Прогноз $T+60$							
Модель			90% интервал				95% интервал			
R	N	RSS	$W_{0.05}$	$W_{0.95}$	$S_{0.05}$	$S_{0.95}$	$W_{0.025}$	$W_{0.975}$	$S_{0.025}$	$S_{0.975}$
1	20	12.26	**0.0564**	0.9260	**0.0715**	0.2533	0.0300	0.9596	0.0984	0.2825
1	30	10.35	0.0640	**0.9345**	0.1505	**0.1749**	0.0349	**0.9643**	0.1851	**0.2135**
1	40	9.48	0.0425	0.9116	0.0920	0.3648	**0.0215**	0.9516	**0.0769**	0.3801
2	30	7.45	0.0212	0.8515	0.4515	0.7377	0.0098	0.9128	0.4414	0.7480
2	40	8.59	0.0202	0.8562	0.4697	0.7028	0.0092	0.9167	0.4646	0.7071

Рис. 3.10, 3.11 показывает, как выбранная модель прогнозирует распределения в моменты T+10, T+30.

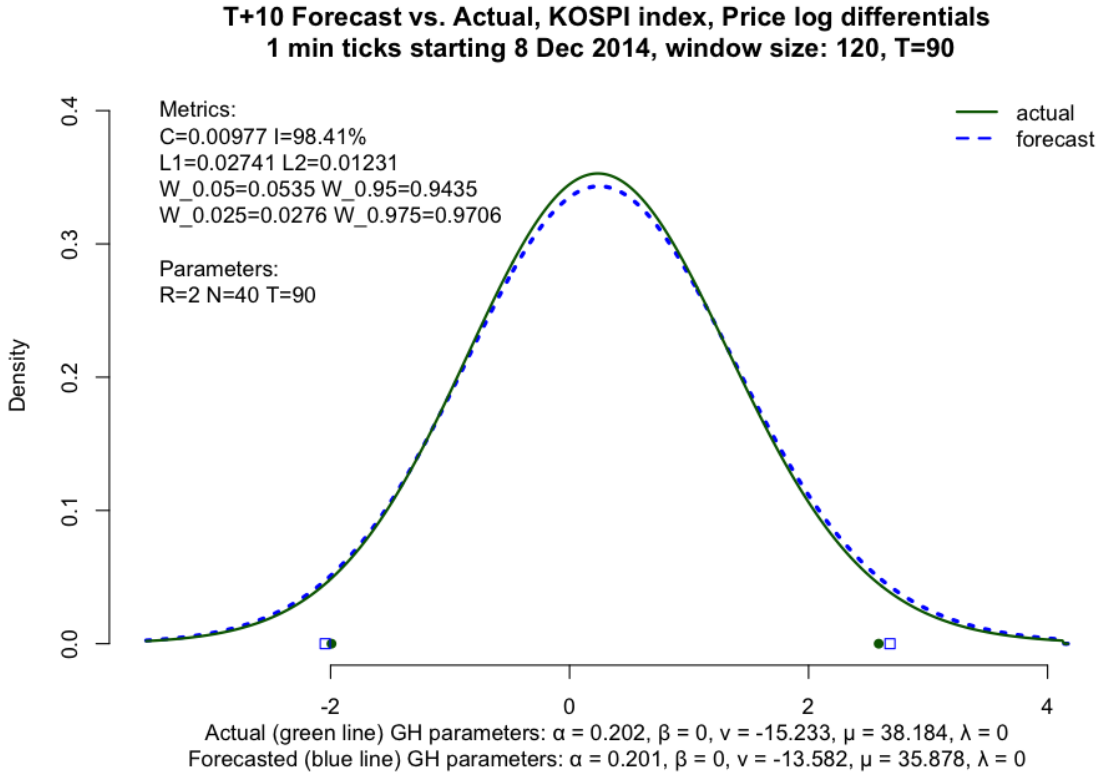

Рисунок 3.10: DJI: сравнение прогноза и истинного распределения для $T+10$

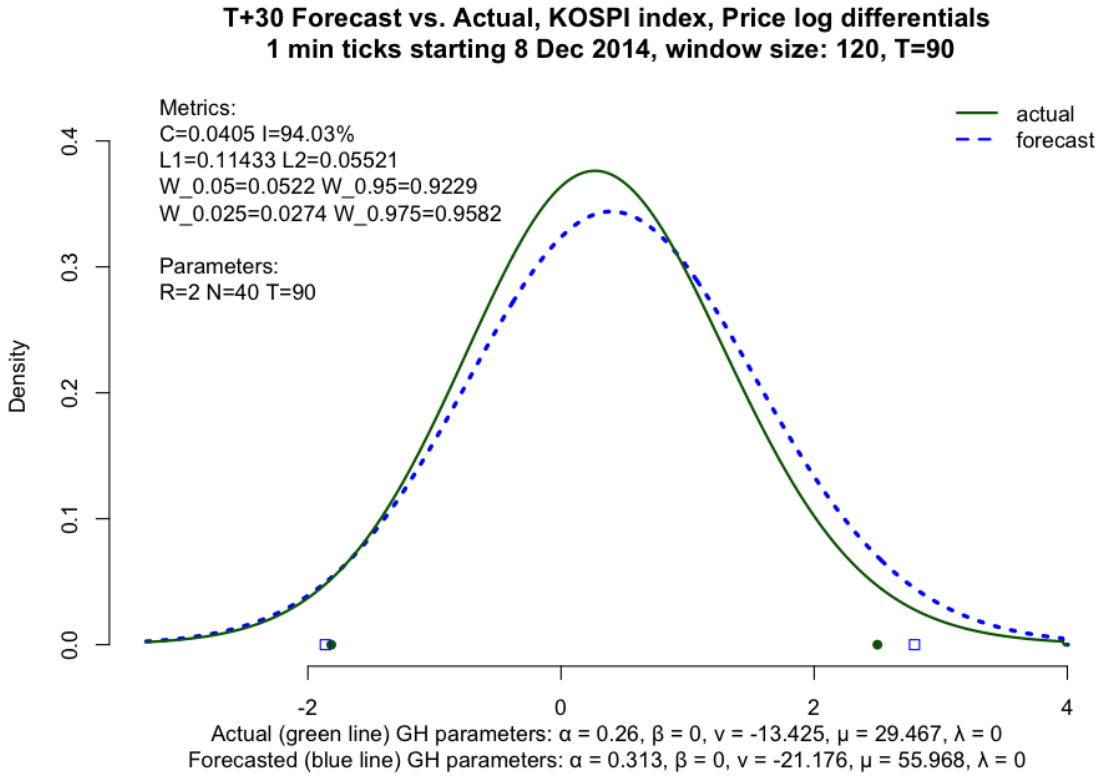

Рисунок 3.11: DJI: сравнение прогноза и истинного распределения для $T + 30$

Важно отметить, что алгоритм прогнозирования не учитывает никаких априорных ограничений рассматриваемых параметров. В случае обобщенных гиперболических распределений имеются внешние условия:

$$\nu \in \mathbb{R},$$

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

Дополнительной входной проверкой модели на адекватность является проверка данных граничных условий на рассматриваемом горизонте прогнозирования. К примеру, если модель на десятом прогнозируемом окне выдает отрицательное значение λ, ее можно смело исключать из рассмотрения.

3.6 Применение метода прогнозирования в задаче анализа текстовой информации для предотвращения утечек данных

3.6.1 Описание задачи

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).

Одним из возможных инструментов для обнаружения подобных утечек является анализ работы отдельно взятого пользователя с текстовой информацией. В статье [82] описывается новый подход к решению этой задачи. Коротко изложим суть подхода. С помощью тематического моделирования выделяются основные тематики информации, с которой работает конкретный пользователь. В режиме скользящего окна формируются временные ряды изменяющихся весов для каждой из тематик. Учитывая некоторое количество накопленных данных строится прогноз изменения временного ряда для следующих точек во времени. Сильное несовпадение прогноза и рассчитанного вектора будет говорить о том, что модель поведения пользователя сильно изменилась, и скорее всего, доступом к его компьютеру завладело третье лицо.

Описанный подход можно разделить на две последовательных задачи. Первая заключается в непосредственном анализе информации, выделению основных тематик и разложению (факторизации) анализируемого контента по выбранным тематикам. На выходе получается набор тематик и соответствующих весов, меняющийся во времени, и это служит исходными данными для второй задачи, а именно прогнозирования весов тематик в будущем и сравнение полученных прогнозов с реальными данными с целью принятия решения о нарушении правил безопасности.

Оказывается, описанный в этой главе метод прогнозирования показывает хорошие результаты при применении его для решения второй задачи. В этом разделе описывается опыт успешного применения метода к исходным данным из статьи [82], а также приводится сравнение предложенного метода с другими использованными методами. Помимо этого, предлагается простой и легко интерпретируемый подход к сравнению прогнозов с реальными данными на основе критерия согласия Хи-квадрат, который позволяет принять решение о возможной утечке данных.

3.6.2 Метод прогнозирования и метод принятия решения об утечке данных

Метод прогнозирования финансовых рисков, описанный в разделе 3.2, прогнозирует непосредственно параметры распределений. При этом, как было отмечено ранее, никаких дополнительных условий на сами параметры не накладывается, за исключением требований к гладкости входных рядов, с целью увеличения точности прогнозирования. Допускается использование любого количества параметров с любыми диапазонами значений.

Данный метод был применен для альтернативного решения второй задачи (прогнозирования распределений тематик и принятия решений) на основе данных из работы [82]. В качестве входной анализируемой информации в этой работе использовался набор Enron – электронная почта 150 сотрудников, в основном топ-менеджмента, обанкротившейся в следствие крупного скандала американской энергетической корпорации Enron. Данный набор широко используется в разных работах, посвященных тематическому анализу данных.

Рассматривался случай $K = 3$ тематик и $K = 6$ тематик. Для каждого из двух случаев был подготовлен набор из 21 исходных рядов, где каждый ряд состоял из $N \approx 40$ векторов-весов длины K.

В качестве моделей для прогноза использовалась регрессия первого порядка ($R = 1$) и регрессия второго порядка ($R = 2$), для обучения модели использовались все доступные наблюдения.

В каждом случае прогноз строился на 7 окон вперед. После построения прогноза, определение момента, когда нужно блокировать доступ пользователя к система (т.е. определение несанкционированного доступа), предлагается проводить по следующему **правилу:** среднее P-значение критерия согласия Хи-квадрат (см. краткое описание в разделе 2.6.3) для наблюдаемых параметров и прогнозируемых параметров, построенных по следующим $N_p = 4$ окнам должно опуститься ниже порога $q = 0.8$.

Для каждого исходного ряда для анализа точности прогнозирования дополнительно было доступно 3 ряда: реальный результат деятельности текущего пользователя, а также результат деятельности двух отличных от него пользователей.

3.6.3 Результаты прогнозирования

Анализ остаточной суммы квадратов (RSS), получаемый при поиске параметров модели прогнозирования, показал, что для случая трех тематик ($K = 3$) в большинстве случаев лучше работает регрессия второго порядка, а для шести ($K = 6$) – первого порядка.

Рассмотрим пример прогноза для $K = 3$ тематик. Для каждого эксперимента построим прогноз и сравним *первый* полученный прогноз с тремя векторами – истинным и двумя альтернативами.

На Рис. 3.12 - 3.14 изображены исторические значения каждого из $K = 3$ параметров, а также проверочные значения (пользователь и 2 альтернативы) для эксперимента N. 14.

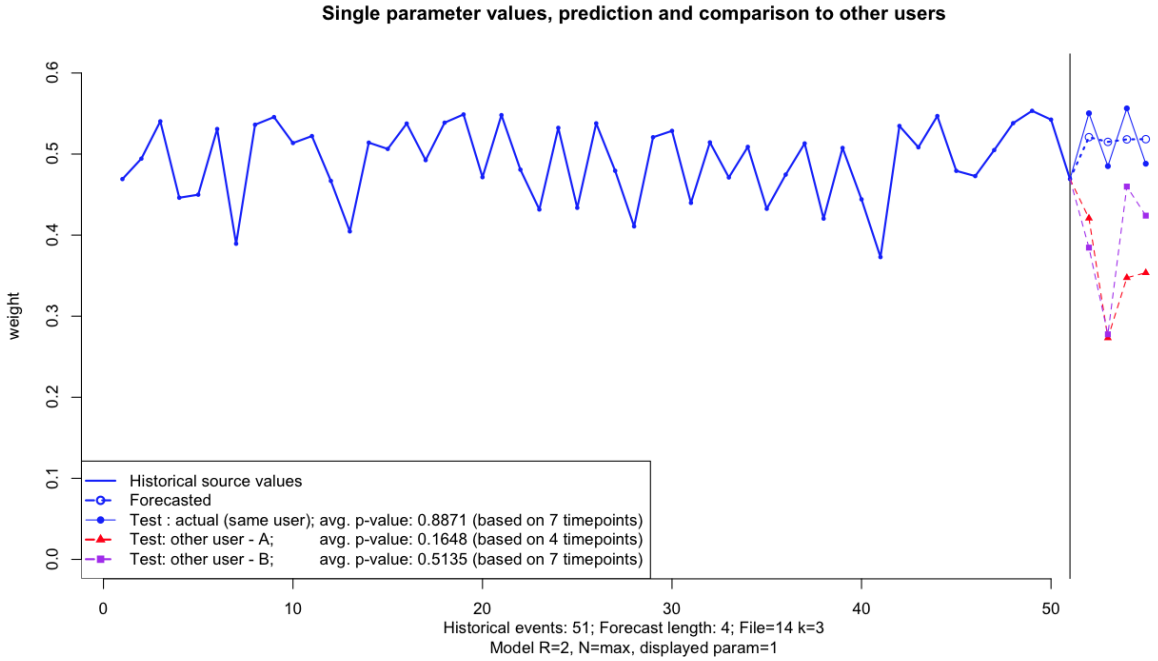

Рисунок 3.12: Значения веса 1; эксперимент 14. $N_p \geq 4$, $K = 3$, $R = 2$

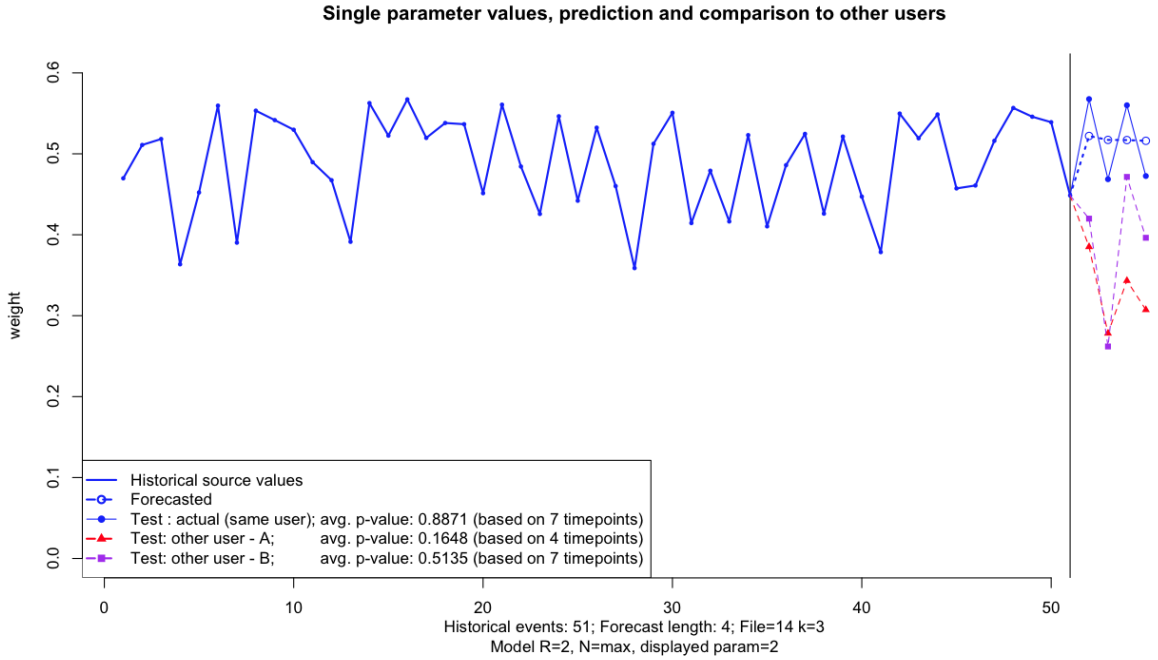

Рисунок 3.13: Значения веса 2; эксперимент 14. $N_p \geq 4$, $K = 3$, $R = 2$

Для сравнения близости прогноза к исходным данным, как было предложено ранее, применим критерий согласия Хи-квадрат. Таблица 3.14 содержит соответствующие значе-

ния. Для удобства анализа, максимальный результат в каждой строке выделен жирным шрифтом.

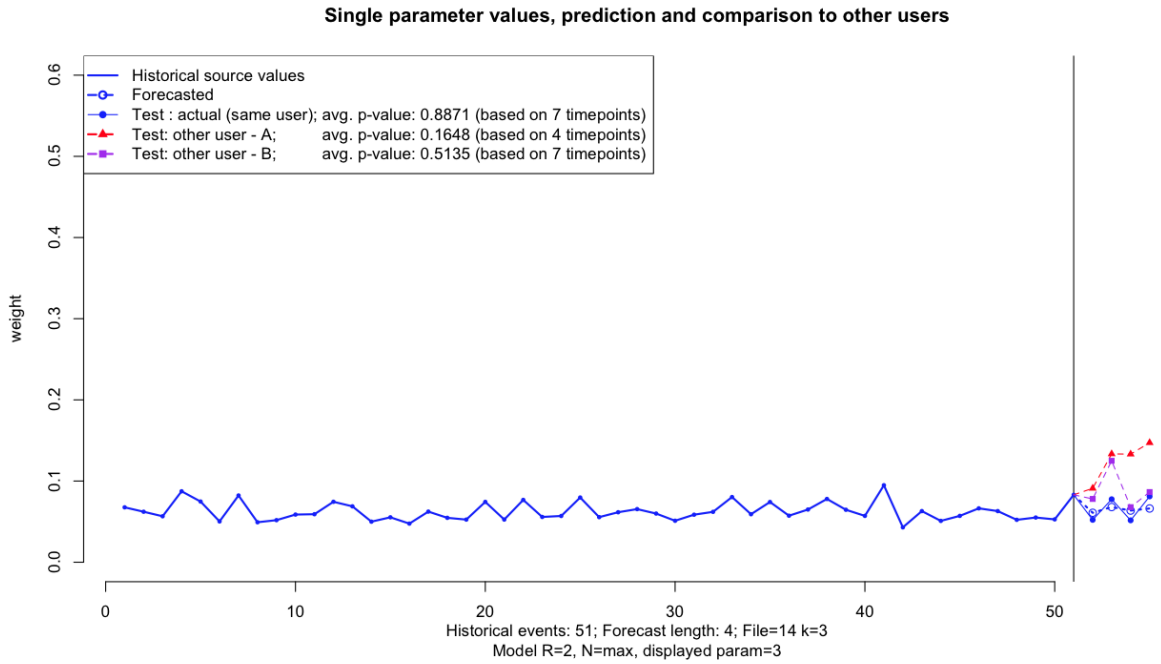

Рисунок 3.14: Значения веса 3; эксперимент 14. $N_p \geq 4$, $K = 3$, $R = 2$

Как видно из полученных данных, при использовании только одного (первого) значения прогноза в целом результаты хороши, но существуют случаи (эксперименты 9, 10, 12, 17, 18), в которых альтернативный пользователь имеет согласованность выше, чем истинный пользователь.

Для устранения этих случайностей вспомним, что прогноз был построен на несколько, а именно 7, наблюдений вперед, и воспользуемся *средним* P-значением для нескольких доступных окон: для каждого эксперимента зафиксируем число N_i, равное минимуму из длин а) доступных рядов для истинного значения, б) горизонта прогнозирования (7), в) доступных данных для пользователей А и Б. Далее в каждом эксперименте посчитаем N_i * P-значений для каждого случая и усредним. Чтобы исключить элемент случайности, надежными будем считать только те проверки, которые были сделаны на данных с $N_i \geq 4$.

Как видно из таблицы 3.15, все средние P-значения для прогноза и истинных данных выше, чем альтернативы. Для иллюстрации правильности выбора порогового значения для детекции, на Рис. 3.15 изображены средние P-значения из таблицы 3.15.

У предложенного подхода к определению возможной утечки данных есть 5 срабатываний «false positive» (ошибок первого рода) и нет срабатываний «false negative» (ошибок второго рода). Это было достигнуто намеренно путем выбора достаточно большого порогового значения, т.к. по умолчанию чаще всего принято считать, что пропуск угрозы несет в себе гораздо больше ущерба, чем ложная детекция угрозы.

Таблица 3.14: P-значения при использовании первого прогноза для текущего пользователя и для двух отличных от него пользователей. $K = 3$, $R = 2$, $N = \max$

	P-значение для пользователей		
Эксперимент	Истинный	Альтернатива A	Альтернатива B
0	**0.9232**	0.0001	0.4064
1	**0.9957**	0.0128	0.0060
2	**0.5860**	0.0193	0.4259
3	**0.1999**	0.0012	0.1409
4	**0.9319**	0.3607	0.7513
5	**0.9800**	0.0126	0.5579
6	**0.9169**	0.0001	0.8545
7	**0.8173**	0.0237	0.6613
8	**0.8903**	0.1172	0.8791
9	0.8939	0.1178	**0.9216**
10	0.8984	0.0467	**0.9559**
11	**0.8471**	0.0726	0.8023
12	**0.8193**	0.7771	0.9152
13	0.7173	0.0100	**0.7177**
14	**0.9895**	0.2466	0.5571
15	**0.9913**	0.0792	0.5757
16	**0.7203**	0.3426	0.2900
17	0.8119	0.3185	**0.8627**
18	0.7842	0.0615	**0.8373**
19	**0.9995**	-	0.0025
20	**0.9966**	-	0.2921

Таблица 3.15: Средние P-значения при использовании $N_i \geq 4$ прогнозов для текущего пользователя и для двух отличных от него пользователей. $K = 3$, $R = 2$, $N = \max$

		P-значение		
Эксперимент	N_i	Истинный пользователь	Альтернативный пользователь A	Альтернативный пользователь B
0	5	**0.9475**	0.0058	0.4297
2	7	**0.8740**	0.0526	0.4173
4	6	**0.7773**	0.1931	0.6977
5	6	**0.9129**	0.0437	0.7022
6	6	**0.8388**	0.2325	0.2830
7	7	**0.7932**	0.1689	0.4705
8	4	**0.9462**	0.2099	0.6478
9	7	**0.9718**	0.1755	0.5798
10	7	**0.7456**	0.0449	0.6423
11	5	**0.8959**	0.1940	0.6008
12	7	**0.7696**	0.3960	0.7490
13	6	**0.8609**	0.0183	0.7779
14	4	**0.9288**	0.0666	0.5681
18	5	**0.8870**	0.4639	0.4840
19	4	**0.9766**	-	0.0969
20	5	**0.7365**	-	0.3334

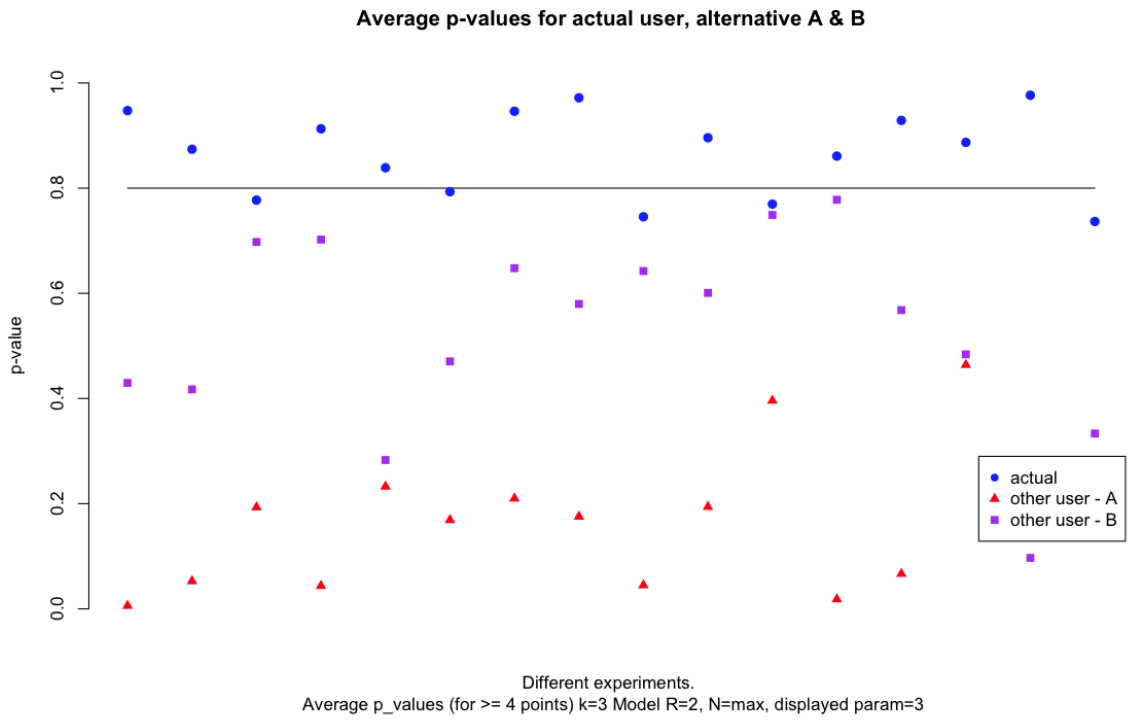

Рисунок 3.15: Средние P-значения; $N_p \geq 4$, $K = 3$, $R = 2$

Результаты выше доказывают состоятельность предложенного подхода и позволяют использовать метод прогнозирования в «боевых» условиях. Более того, как мы отмечали ранее, подбор параметров регрессии представляет из себя очень быструю процедуру, что может быть важно при обработке больших объемов данных и сохранении низкого времени реагирования на угрозу.

В разделе 3.6.5 также приводятся дополнительные советы и соображения, которые позволят еще больше улучшить точность работы алгоритма и его применимость на практике.

3.6.4 Сравнение полученных результатов с результатами других алгоритмов

Описанный подход к анализу средних P-значений для оценки качества прогнозирования был также применен к методам, используемым в работе [82]: двум алгоритмам с использованием технологий Microsoft (будем обозначать их как «ms.0.0.3t» и «ms.0.0») и двум других алгоритмам (обозначаим их «nmf1», «nmf4»).

В таблицах 3.16, 3.17 приведены таблицы для P-значений (только первого предсказания и среднее P-значение соответственно) для алгоритма «ms.0.0.3t».

В таблице 3.18 приведены сравнения P-значений критерия согласия Хи-квадрат для первого прогноза, полученного разными алгоритмами, с истинными значениями пользователя.

Таблица 3.16: P-значения при использовании первого прогноза для текущего пользователя и для двух отличных от него пользователей, алгоритм ms.0.0.3t

	P-значение		
Эксперимент	Истинный	Альтернатива A	Альтернатива B
0	0.9528	0.0011	0.5025
1	0.954	0.0191	0.0084
2	0.664	0.0298	0.4928
3	0.355	0.0068	0.2615
4	0.8585	0.2833	0.684
5	0.8491	0.0044	0.4197
6	0.9435	2e-04	0.9325
7	0.8526	0.0459	0.7915
8	0.9405	0.0798	0.8103
9	0.9536	0.0652	0.9695
10	0.9152	0.0515	0.9732
11	0.8413	0.0261	0.774
12	0.7964	0.736	0.8883
13	0.6397	0.0063	0.624
14	0.9625	0.3216	0.6465
15	0.9544	0.0488	0.4684
16	0.8123	0.2324	0.188
17	0.446	0.648	0.5052
18	0.1797	0.1004	0.213
19	0.9987	–	0.0023
20	0.9983	–	0.2261

Таблица 3.17: Средние P-значения при использовании $N_i \geq 4$ прогнозов для текущего пользователя и для двух отличных от него пользователей, алгоритм ms.0.0.3t

		P-значение		
Эксперимент	N_i	Истинный пользователь	Альтернативный пользователь A	Альтернативный пользователь B
0	5	0.8904	0.018	0.5583
2	7	0.888	0.0535	0.4359
4	6	0.7547	0.1799	0.6888
5	6	0.9037	0.0498	0.7243
6	6	0.8285	0.2425	0.3029
7	7	0.8528	0.234	0.5329
8	4	0.9038	0.2768	0.704
9	7	0.9654	0.1533	0.5746
10	7	0.7456	0.0439	0.6335
11	5	0.8662	0.1607	0.6194
12	7	0.7492	0.3718	0.7209
13	6	0.5805	0.0522	0.5611
14	4	0.8555	0.088	0.5518
18	5	0.5332	0.3266	0.3197
19	4	0.9958	0.1057	0.1057
20	5	0.7409	0.3213	0.3213

Таблица 3.18: Сравнение P-значений первого прогноза для текущего пользователя

	P-значение				
Эксперимент	Корчагин	ms.0.0.3t	ms.0.0	nmf1	nmf4
0	0.9232	0.9528	0.9528	**0.9539**	**0.9359**
1	**0.9957**	0.954	0.9524	0.983	0.9626
2	0.5860	0.664	**0.69**	0.4332	0.4273
3	0.1999	0.355	0.355	0.3638	**0.3974**
4	0.9319	0.8585	0.8585	0.9307	**0.9399**
5	**0.9800**	0.8491	0.7419	0.9322	0.9289
6	0.9169	0.9435	**0.979**	0.9453	0.956
7	0.8173	0.8526	0.8526	**0.9615**	0.9579
8	0.8903	**0.9405**	**0.9405**	0.8799	0.8792
9	0.8939	0.9536	**0.9537**	0.9406	0.9406
10	0.8984	**0.9152**	**0.9152**	0.8763	0.8978
11	0.8471	0.8413	**0.9588**	0.8592	0.8405
12	0.8193	0.7964	0.7964	**0.9249**	0.8996
13	0.7173	0.6397	0.6397	**0.7618**	0.6906
14	**0.9895**	0.9625	0.9	0.9805	0.9777
15	**0.9913**	0.9544	0.6262	0.9812	0.9592
16	0.7203	**0.8123**	**0.8123**	0.7525	0.8037
17	**0.8119**	0.446	0.6239	0.7872	0.8037
18	**0.7842**	0.1797	0.1797	0.7804	0.7811
19	**0.9995**	0.9987	0.9987	0.9928	0.9945
20	0.9966	**0.9983**	0.9926	0.9928	0.9886

Таблица 3.19: Сравнение средних P-значений при использовании $N_i \geq 4$ прогнозов для текущего пользователя с использованием разных алгоритмов

		среднее P-значение				
Эксперимент	N_i	Корчагин	ms.0.0.3t	ms.0.0	nmf1	nmf4
0	5	0.9475	0.8904	0.8904	**0.9515**	0.9467
2	7	0.8740	**0.888**	0.8874	0.869	0.8653
4	6	0.7773	0.7547	0.7547	0.8007	**0.8179**
5	6	0.9129	0.9037	0.5922	**0.923**	0.9214
6	6	0.8388	0.8285	**0.8548**	0.8526	0.847
7	7	0.7932	**0.8528**	**0.8528**	0.8336	0.8297
8	4	**0.9462**	0.9038	0.9038	0.9368	0.936
9	7	0.9718	0.9654	0.9665	**0.9743**	**0.9743**
10	7	**0.7456**	**0.7456**	**0.7456**	0.6726	0.6811
11	5	0.8959	0.8662	**0.932**	0.8708	0.8657
12	7	0.7696	0.7492	0.7492	**0.7985**	0.7884
13	6	0.8609	0.5805	0.5805	**0.8912**	0.8841
14	4	**0.9288**	0.8555	0.8425	0.9055	0.9101
18	5	**0.8870**	0.5332	0.5332	0.8639	0.871
19	4	0.9766	**0.9958**	**0.9958**	0.9899	0.9901
20	5	0.7365	**0.7409**	0.7345	0.7267	0.723
$\sum$		**13.8626**	13.0542	12.8159	13.8606	13.8518

На основании данных, приведенных в Таб. 3.18 посчитаем, какой из алгоритмов наибольшее число раз показал самое высокое значение согласованности первого прогноза (из 21 эксперимента):

- Корчагин – **7**,
- ms.0.0.3t – 4,
- ms.0.0 – **7**,
- nmf1 – 4,
- nmf4 – 3.

Эти данные позволяют нам сделать следующий вывод: **предложенный в этой главе метод отлично справляется с задачей прогнозирования первого прогноза и наибольшее количество раз показал лучший первый прогноз, наравне с алгоритмом «ms.0.0»**.

Как было отмечено ранее, более надежной оценкой является не одно P-значение для первого прогноза, а среднее значение для нескольких прогнозов. В таблице 3.19 приведены сравнения *средних* P-значений критерия Хи-квадрат для полученных разными алгоритмами прогнозов с истинными значениями пользователя.

На основе полученных данных посчитаем, каково *суммарное среднее значение согласованности* в 16 экспериментах. Очевидно, что чем выше это значение, тем лучше работает алгоритм прогнозирования. Результаты приведены в посл. строчке таблицы 3.19, где по сути просто просуммированы значения в каждом из столбцов.

Отметим, что *каждый* из остальных четырех методов хотя бы в одном из экспериментов дает прогноз с низким (< 0.7) средним p-значением, в то время как предложенный метод во всех без исключения экспериментах продемонстрировал результаты с высокими p-значениями (≥ 0.7).

На основании этого можно сделать окончательный вывод: **предложенный метод как минимум не уступает по точности прогнозирования ближайших значений алгоритмам, используемым в работе [82]**, и может быть успешно использован при практическом решении данной задачи на больших массивах данных. При этом предложенный метод особенно удачно определяет непосредственно первое прогнозируемое значение в сравнении с другими алгоритмами.

3.6.5 Дальнейшие шаги по улучшению используемого метода

Эксперименты выше были проведены на ограниченном наборе данных и доказали состоятельность предложенного подхода. Для реального (production) использования метода можно предпринять несколько шагов, которые позволят добиться большей точности прогнозирования и уменьшения количества ложных срабатываний:

- увеличить минимальное требование к N_i, то есть увеличить период, на котором прогноз сравнивается с истинными значениями. При этом увеличится время работы злоумышленника с системой в случае срабатывания тревоги и количество потерянны данных может быть больше;

- пересмотреть важность ошибок первого и второго родов между собой и сдвинуть порог вверх или вниз с целью оптимизации соотношения ошибок первого и второго рода;

- увеличить K, при этом, некоторые компоненты могут иметь большую важность, чем другие (например, финансовые или клиентские данные);

- увеличить частоту поступающих данных;

- более тщательно подобрать модель (R, N) на основе метрик, отличных от RSS (по аналогии с разделом 3.3).

Заключение

Данная работа посвящена изучению специальных вероятностных моделей стохастических процессов и явлений, имеющих вид дисперсионно-сдвиговых смесей нормальных законов, в частности, обобщенных гиперболических и обобщенных дисперсионных гамма-распределений. Основные результаты работы заключаются в следующем.

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, 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 МЭГ 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.

Дальнейшие перспективы развития исследований могут быть связаны с использованием предложенных моделей и методов применительно не только к анализу финансовых данных (в том числе высокочастотных), но и к анализу данных турбулентной плазмы, данных, получаемых при океанологических исследованиях, а также многих других актуальных практических задачах. Помимо этого, в разделах 3.4.6 и 3.6.5 предложены конкретные шаги для улучшения предложенных методов декомпозиции и прогнозирования.

Список литературы

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.

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, том 1559, с. 2394–2397.

11. *A. K. Gorshenin, V. Yu. Korolev, A. Yu. Korchagin, T. V. Zakharova, A. I. Zeifman* Statistical detection of movement activities in a human brain by separation of mixture distributions // Available at: arXiv:1503.00299 [stat.AP], 2015.

12. *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.

13. *Биллингсли П.* Сходимость вероятностных мер. – М.: Наука, 1977.

14. *Кендалл М., Стьюарт А.* Статистические выводы и связи. – М.: Наука, 1973.

15. *Гихман И. И., Скороход А. В.* Теория случайных процессов. Т. 1. – М.: Наука, 1971.

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

17. *Гнеденко Б. В., Фахим Х.* Об одной теореме переноса // Доклады АН СССР, 1969. Т. 187. Вып. 1. С. 15–17.

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

19. *Гумбель Э.* Статистика экстремальных значений. – М.: Мир, 1965.

20. *Королев В. Ю., Соколов И. А.* Математические модели неоднородных потоков экстремальных событий. – М.: Торус Пресс, 2008.

21. *Королев В. Ю.* О распределении размеров частиц при дроблении // Информатика и ее применения, 2009. Т. 3. Вып. 3. С. 60–68.

22. *Королев В. Ю., Бенинг В. Е., Шоргин С. Я.* Математические основы теории риска. 2-е издание, переработанное и дополненное // М.: ФИЗМАТЛИТ, 2011. 620 С.

23. *Королев В. Ю., Шоргин С. Я.* Математические методы анализа стохастической структуры информационных потоков. – М.: ИПИ РАН, 2011. 130 с.

24. *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.

25. *Королев В. Ю.* О взаимосвязи обобщенного распределения Стьюдента и дисперсионного гамма-распределения при статистическом анализе выборок случайного объема // ДАН, 2012. Т. 445. Вып. 6. С. 622–627.

26. *Королев В. Ю., Соколов И. А.* Об условиях сходимости распределений экстремальных порядковых статистик к распределению Вейбулла // Информатика и ее применения, 2014. Т. 8. Вып. 3. С. 2–10.

27. *Круглов В. М.* Сходимость числовых характеристик сумм независимых случайных величин со значениями в гильбертовом пространстве // Теория вероятностей и ее применения, 1973. Т. 18. Вып. 4. С. 734–752.

28. *Лоэв М.* Теория вероятностей. – М.: Мир, 1962.

29. *Ширяев А. Н.* Основы стохастической финансовой математики. Том 1. Факты. Модели. М.: «Фазис», 1998.

30. *Barndorff-Nielsen O. E.* Exponentially decreasing distributions for the logarithm of particle size // Proc. Roy. Soc. Lond., Ser. A, 1977. Vol. 353. P. 401–419.

31. *Barndorff-Nielsen O. E.* Hyperbolic distributions and distributions of hyperbolae // Scand. J. Statist., 1978. Vol. 5. P. 151–157.

32. *Barndorff-Nielsen O. E.* Models for non-Gaussian variation, with applications to turbulence // Proc. Roy. Soc. London, Ser. A, 1979. Vol. A(368). P. 501–520.

33. *Barndorff-Nielsen O. E., Kent J., Sørensen M.* Normal variance-mean mixtures and z-distributions // Int. Statist. Rev., 1982. Vol. 50. No. 2. P. 145–159.

34. *Barndorff-Nielsen O. E., Blæsild P., Schmiegel J.* A parsimonious and universal description of turbulent velocity increments // European Physical Journal, 2004. Vol. B 41. P. 345–363.

35. *Carr P. P., Madan D. B., Chang E. C.* The Variance Gamma process and option pricing // European Finance Review, 1998. Vol. 2. P. 79–105.

36. *Qian Chen, Gerlach R. H.* The two-sided Weibull distribution and forecasting financial tail risk. OME Working Paper No. 01/2011 – Sydney: Business School, The University of Sydney, 2011.

37. *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.

38. *Eberlein E., Keller U.* Hyperbolic Distributions in Finance // Bernoulli, 1995. Vol. 1, No. 3. P. 281–299.

39. *Eberlein E., Keller U., Prause K.* New insights into smile, mispricing and value at risk: the hyperbolic model // Journal of Business, 1998. Vol. 71. P. 371–405.

40. *Eberlein E., Prause K.* The Generalized Hyperbolic Model: Financial Derivatives and Risk Measures // Freiburg: Universität Freiburg, Institut für Mathematische Stochastic, Preprint N 56, 1998.

41. *Eberlein E.* Application of Generalized Hyperbolic Lévy Motions to Finance // Freiburg: Universität Freiburg, Institut für Mathematische Stochastic, Preprint No. 64, 1999.

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

43. *Goldie C. M.* A class of infinitely divisible distributions // Math. Proc. Cambridge Philos. Soc., 1967. Vol. 63. P. 1141-1143.

44. *Jacod J., Shiryaev A. N.* Limit theorems for stochastic processes. 2nd edition. Volume 288 of Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. – Berlin: Springer-Verlag, Berlin, 2003.

45. *Kalashnikov V. V.* Geometric Sums: Bounds for Rare Events with Applications. – Dordrecht: Kluwer Academic Publishers, 1997.

46. *Королев В. Ю., Закс Л. М., Зейфман А. И.* О сходимости случайных блужданий, порожденных обобщенными процессами Кокса к процессам Леви // Информатика и ее применения, 2013. Т. 7. Вып. 2. С. 84–91.

47. *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 [math.PR]. 4 October 2014.

48. *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 // Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2526053

49. *LeCam L.* Maximum likelihood; an introduction // International Statistical Review, 1990. Vol. 58. P. 153–171.

50. *Madan D. B., Seneta E.* The variance gamma (V.G.) model for share market return // Journal of Business, 1990. Vol. 63. P. 511–524.

51. *Prause K.* Modeling Financial Data Using Generalized Hyperbolic Distributions // Freiburg: Universität Freiburg, Institut für Mathematische Stochastic, Preprint N 48, 1997.

52. *Rényi A.* On the central limit theorem for the sum of a random number of independent random variables // Acta Math. Acad. Sci. Hung., 1960. Vol. 11. P. 97–102.

53. *Seshadri V.* Halphen's laws // Kotz, S., Read, C. B., Banks, D. L. (Eds.). Encyclopedia of Statistical Sciences, Update Volume 1. New York: Wiley, 1997. P. 302–306.

54. *Sichel H. S.* Statistical evaluation of diamondiferous deposits // Journal of South Afr. Inst. Min. Metall., 1973. Vol. 76. P. 235–243.

55. *Sornette D., Simonetti P., Andersen J. V.* Φ^q-field theory for portfolio optimization: fat-tails and non-linear correlations // Physics Reports, 2000. Vol. 335(2). P. 19–92.

56. *Stacy E. W.* A generalization of the gamma distribution // Annals of Mathematical Statistics, 1962. Vol. 33. P. 1187–1192.

57. *Teicher H.* Identifiability of mixtures // Ann. Math. Stat., 1961. Vol. 32. P. 244–248.

58. *VZolotarev . M.* Modern Theory of Summation of Random Variables. – Utrecht: VSP, 1997.

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

60. *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.

61. *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.

62. *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.

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

64. *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.

65. *Королев В. Ю., Назаров А. Л.* Разделение смесей вероятностных распределений при помощи сеточных методов моментов и максимального правдоподобия // Автоматика и телемеханика, 2010. Вып. 3. С. 98–116.

66. *Dennis J. E., Schnabel R. B.* Numerical Methods for Unconstrained Optimization and Nonlinear Equations. – Englewood Cliffs: Prentice-Hall, 1983.

67. *Barndorff-Nielsen O. E.* Processes of normal inverse Gaussian type // Finance and Stochastics, 1998. Vol. 2. P. 41–18.

68. *Bening V. E., Korolev V. Yu.* Generalized Poisson Models and their Applications in Insurance and Finance. – Utrecht: VSP, 2002.

69. *Бенинг В. Е., Королев В. Ю.* Об использовании распределения Стьюдента в задачах теории вероятностей и математической статистики // Теория вероятностей и ее применения, 2004, 49:3, С. 417–435.

70. *Blæsild P.* The two-dimensional hyperbolic distribution and related distributions, with an application to Johannsen's bean data // Biometrika, 1981. Vol. 68, P. 251–263.

71. *Blæsild P., Jensen J. L.* Multivariate distributions of hyperbolic type // *C. Taillie, G. P. Patil, B. Baldessari* (*Eds.*). Statistical Distributions in Scientific Work. Vol. 4. – Dordrecht: Reidel, 1981. P. 45–66.

72. *Григорьева М. Е., Королев В. Ю.* О сходимости распре-делений случайных сумм к скошенным экспоненциально-степенным законам // Информатика и её применения, 2013. Т. 7. Вып. 4. С. 66–74.

73. *Королев В. Ю.* О предельных распределениях случайно индексированных случайных последовательностей // ТВП, 37:3 (1992), С. 564-–570.

74. *Korolev V. Yu., Skvortsova N. N.* Stochastic Models of Structural Plasma Turbulence. – Utrecht: VSP, 2006.

75. *McGillem C.D., Aunon J.I.* Analysis of Event-Related Potentials. // Methods of Analysis of Brain Electrical and Magnetic Signals: EEG Handbook. 1987, A.S. Gevins, A. Remond (Eds.). Amsterdam: Elsevier Science Publishers, P. 131–169.

76. *Fabiani M., Gratton G., Federmeier K.* Event-Related Brain Potentials: Methods, Theory and Application. // Handbook of Psychophysiology, 2007. Cambridge: Cambridge University Press, pp. 85–119.

77. *Захарова Т. В., Никифоров С. Ю., Гончаренко М. Б., Драницына М. А., Климов Г. А., Хазиахметов М. Ш., Чаянов Н. В.* Методы обработки сигналов для локализации невосполнимых областей головного мозга // Системы и средства информатики, 22:2 (2012), сс. 157-–175.

78. *Горшенин А. К., Королёв В. Ю., Турсунбаев А. М.* Медианные модификации EM- и SEM-алгоритмов для разделения смесей вероятностных распределений и их применение к декомпозиции волатильности финансовых временных рядов. // Информатика и ее применения, 2:4 (2008), сс. 12-–47.

79. *Gorshenin A. K., Korolev V. Yu.* Modeling 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). Digitaldruck Pirrot GmbHP, Dudweiler, Germany, 2013, pp. 569–572.

80. *Горшенин А. К.* Информационная технология исследования тонкой структуры хаотических процессов в плазме с помощью анализа спектров // Информатика и ее применения, Т. 24, Ч. 1, сс. 116-125.

81. *Jørgensen B.* Statistical properties of the generalized inverse Gaussian distribution. // Lecture Notes in Statistics, 1982, vol. 9. Springer, Berlin.

82. *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.

Список рисунков

Список таблиц