НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ
ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

На правах рукописи

Ковалевский Артем Павлович

СТАТИСТИЧЕСКИЕ КРИТЕРИИ АПОСТЕРИОРНОГО ОБНАРУЖЕНИЯ РАЗЛАДКИ ВРЕМЕННЫХ РЯДОВ И ИХ ПРИМЕНЕНИЯ

Специальность 05.13.17 — Теоретические основы информатики

Диссертация на соискание ученой степени
доктора физико-математических наук

Новосибирск — 2018

Содержание

ВВЕДЕНИЕ

В диссертации решены задачи построения и сравнения статистических критериев апостериорного обнаружения разладки временных рядов для широкого класса вероятностных моделей.

Постановка задачи состоит в следующем. У исследователя есть конечное число наблюдений и вероятностная модель, предложенная для объяснения этих наблюдений. Проверяется основная гипотеза о том, что наблюдения соответствуют предложенной модели. В качестве альтернативной гипотезы предлагается гипотеза о том, что происходит разладка — изменение параметров модели в некоторый неизвестный исследователю момент времени.

В простейшей постановке наблюдается временной ряд, для которого рассматриваются две вероятностные модели: согласно основной гипотезе, элементы временного ряда являются независимыми одинаково распределенными случайными величинами; согласно альтернативной гипотезе, в некоторый момент внутри интервала наблюдения происходит разладка: случайные величины по-прежнему предполагаются независимыми, но до момента разладки имеют одно распределение, а после момента разладки — другое. Задача оценивания параметров разладки решалась А. А. Боровковым и Ю.Ю. Линке, Вальдом (A. Wald), И. В. Никифоровым, Н. Клигене и Л. Телькснисом, И. Ш. Торговицким, А. Н. Ширяевым, В. И. Лотовым, Карлстейном (E. Carlstein), Ксорго и Хорвафом (M. Csorgo, L. Horvath), Дембгеном (L. Dümbgen), Шабаном (S. A. Shaban).

Различают две задачи обнаружения разладки: последовательную процедуру и апостериорную. При последовательной процедуре значения появляются одно за другим, и акцент делается на наискорейшем обнаружении разладки при фиксированных вероятностях

ошибок. Для обнаружения разладки используется последовательный критерий отношения правдоподобия Вальда. А. Н. Ширяевым изучены последовательные решающие правила для марковских процессов. В работе С. Э. Воробейчикова и Ю. С. Пономаревой предлагается применение последовательного метода наименьших квадратов для обнаружения разладки авторегрессионных моделей и их обобщений.

Задача апостериорного обнаружения разладки состоит в том, что временной ряд известен полностью, и надо сделать выбор между моделью выборки и моделью разладки. Эта задача рассмотрена в книге Б. Е. Бродского и Б. С. Дарховского. Отметим, что как в этой книге, так и в монографии Ксорго и Хорвафа отсутствует сравнение критериев обнаружения разладки на основании относительной асимптотической эффективности по Питмену. Между тем такое сравнение, как будет показано ниже, приводит к выбору единственного критерия, относительно наиболее асимптотически эффективного по Питмену в широком классе, и потому наиболее подходящего для практического различения близких гипотез.

Математические подходы к анализу текстов развиты в работах Н. А. Морозова, А. А. Маркова, Г. Хетсо, Д. В. Хмелева, А. А. Поликарпова, В. В. Поддубного и О. Г. Шевелева. Применение разработанных в диссертации критериев к анализу однородности художественных текстов на русском языке осуществляется на основании метода, предложенного Н. А. Морозовым и развитого рядом авторов, в том числе В. П. Фоменко и Т. Г. Фоменко. Метод состоит в том, что по тексту строится временной ряд индикаторов служебных слов (предлогов, союзов, частиц): выбирается значение 1, если слово является служебным, и значение 0 иначе. Обобщением этого метода является метод построения нескольких связанных временных рядов

по тексту. Временные ряды могут отражать, в частности, число букв или слогов в слове, число слов в предложении (Г. Хетсо). Однако все исследованные нами временные ряды, за исключением ряда индикаторов служебных слов, не обладают в нужной степени селективностью автора: методами дисперсионного анализа можно убедиться в том, что внутригрупповые дисперсии для текстов каждого автора обеспечивают основной вклад в суммарную дисперсию, и различия в средних значениях характеристик для разных авторов незначимы. Таким образом, эти характеристики (число букв в слове и т.п.) менее полезны для анализа однородности текста.

Результат применения статистического критерия к анализу однородности каждого конкретного текста — достигнутый уровень значимости гипотезы об однородности. Чем меньше (ближе к нулю) достигнутый уровень значимости для данного текста, тем более это говорит против гипотезы об однородности. В целом, для более длинных текстов одного автора достигнутые уровни значимости оказываются ниже, чем для коротких. Но значительно ниже достигнутые уровни значимости для текстов, полученных склейкой (конкатенацией) двух произведений разных авторов. Разработанный метод позволяет не только диагностировать наличие разладки, то есть неприемлемость модели выборки, но и указывать момент разладки — склейка текстов, содержащих многие сотни страниц, отыскивается с точностью порядка одной страницы.

Однако отмеченное выше свойство уменьшения достигаемого уровня значимости с ростом объема текста говорит о том, что модель выборки не является удовлетворительной для временного ряда, полученного по произведению или собранию сочинений автора. Действительно, статистический анализ показывает наличие корреляций, медленно убывающих с ростом лага. Адекватной моделью

для таких временных рядов является модель фрактального шума.

Фрактальный гауссовский шум — это стационарная гауссовская последовательность с нулевым математическим ожиданием и корреляционной функцией, убывающей по степенному закону таким образом, что для частичных сумм S_n выполнено равенство $\mathbf{D}S_n = \sigma^2 n^{2H}$, где $H \in (0,\ 1]$ называется показателем Херста. В случае склейки текстов двух авторов получается модель разладки фрактального гауссовского шума.

Модель фрактального гауссовского шума введена независимо друг от друга А. Н. Колмогоровым и Винером в 1940 году, а свое название и широкую известность получила благодаря статье Мандельброта и ван Несса. Она оказывается полезной для описания экономических и естественно-научных временных рядов. Использование модели осложнено трудоемкой процедурой оценивания ее параметров.

Для решения этой проблемы автором разработан бинарный знаковый метод оценивания параметра H, имеющий низкую вычислительную сложность. Дисперсия оценки, полученной этим методом, не намного больше дисперсии оптимальной оценки. Асимптотика дисперсии оценки вычислена аналитически. Разработаны и реализованы алгоритмы моделирования фрактального гауссовского шума и алгоритм оценивания параметра Херста бинарным знаковым методом. Кроме того, разработаны анлитические вероятностные методы вычисления дисперсии оценки, согласующиеся . Результаты моделирования и теоретических согласуются с результатами, полученными аналитическими методами теории вероятностей.

Бинарный знаковый метод оценивания параметра H применяется к временным рядам, построенным по текстам на естественном языке с помощью авторского инварианта. На его основании строится

статистический критерий проверки гипотезы об отсутствии фрактальности. Критерий основан на разности между оценкой параметра Херста и значением 1/2, соответствующим белому гауссовскому шуму. Построенный критерий позволяет при изучении текстов обосновать выбор модели фрактального шума против альтернативной модели выборки. Этот вывод делается на основании того, что коэффициент Херста значимо отличается от 1/2.

Для исследования разладки в текстах, то есть для выбора между моделью фрактального гауссовского шума и разладки фрактального гауссовского шума, разработан статистический критерий, основанный на разности оценок разными методами: методом дисперсии и бинарным знаковым методом. Этот критерий дает хорошие результаты при анализе однородности текста: достигнутые уровни значимости далеки от нуля для текстов одного автора и близки к нулю для склейки текстов разных авторов.

Обнаружение разладки процессов гармонических колебаний со случайным шумом основано на изучении асимптотического поведения сумм остатков соответствующей регрессионной модели. Это изучение было осуществлено МакНилом в 1978 году. В диссертации доказана теорема, распространяющая результат МакНила на эмпирический мост. Получены следствия о предельных распределениях ряда функционалов от эмпирического моста. Сравнение статистических критериев, основанных на этих функционалах, проведено на численном примере в случае, когда альтернатива состоит в однократном изменении математического ожидания в случайный момент времени. Сравнение проводилось путем моделирования и выявило преимущество функционала супремума отклонения эмпирического моста.

Цель работы

Целью работы является разработка статистических критериев проверки гипотез для класса вероятностных моделей временных рядов. Этот класс включает в себя модели случайной выборки и ее разладки, фрактального гауссовского шума и его разладки, циклического тренда со случайным шумом и его разладки. Разработанные критерии апостериорного обнаружения разладки применяются для анализа однородности текстов и для выбора адекватной вероятностной модели в ряде задач анализа медицинских и экономических данных.

В рамках указанной цели были поставлены следующие задачи.

1. Построить класс статистических критериев проверки однородности временного ряда, основанных на функционалах от эмпирического моста. Сравнить критерии из этого класса в смысле асимптотической относительной эффективности по Питмену и выбрать наилучший критерий.
2. Построить класс статистических критериев проверки однородности временного ряда, основанных на предположениях нормальности.
3. Разработать знаковый метод оценивания параметра Херста и его модификации для моделей фрактального гауссовского шума и фрактального броуновского моста.
4. Построить класс статистических критериев для различения модели выборки и фрактального шума; модели фрактального шума и его разладки.
5. Применить разработанные критерии обнаружения разладки к анализу однородности текста. Формализовать модели временных рядов, построенных по тексту одного автора и по склейке

текстов разных авторов. Исследовать результативность применения разработанных статистических критериев проверки однородности к текстам на естественном языке и обосновать алгоритм выявления склейки текстов.

6. Применить разработанные критерии для выбора адекватной вероятностной модели к анализу медицинских и экономических данных.

7. Построить статистические критерии обнаружения разладки процесса гармонических колебаний со случайным шумом, пригодные для использования при компьютерном анализе изменений прочностных характеристик конструкции на основании записи ее колебаний.

8. На численном примере сравнить предложенные критерии в случае, когда альтернатива состоит в однократном изменении математического ожидания в случайный момент времени, равномерно распределенный на интервале наблюдения.

Методы исследования. Исследования, проведенные в работе, основаны на применении и развитии методов статистического анализа временных рядов, в частности, методов апостериорного обнаружения разладки, изложенных в монографиях Б. Е. Бродского и Б. С. Дарховского, Ксорго и Хорвафа; методов сравнения критериев, изложенных в книге Я. Ю. Никитина; знакового метода оценивания корреляционной функции А. М. Яглома; алгоритмов оценивания параметров и проверки гипотез для фрактального гауссовского шума, развитых в ряде работ последних десятилетий; математических методов классификации текстов, развитых в работах Н. А. Морозова, А. А. Маркова, Г. Хетсо, Д. В. Хмелева, В. В. Поддубного и О. Г. Шевелева; классических методов доказательства предельных

теорем теории вероятностей в функциональных пространствах.

Апробация работы

Основные результаты диссертации докладывались и обсуждались на следующих всероссийских и международных конференциях:

- Workshop on Mathematics of Stochastic Networks, EURANDOM, The Netherlands, November 2001.
- 6th International Symposium on science and technology (KORUS–2002), Novosibirsk, June 24–26, 2002.
- Квантитативная лингвистика: исследования и модели (КЛИМ–2005), Новосибирский государственный педагогический университет, Новосибирск, 6–10 июня 2005 г.
- IV International Conference on Limit Theorems in Probability Theory and Their Applications, Novosibirsk, 21–25 August 2006.
- XIV Всероссийская школа-коллоквиум по стохастическим методам и VIII Всероссийский симпозиум по прикладной и промышленной математике, Сочи-Адлер, 29 сентября – 7 октября 2007 г.
- IX Международная научно-техническая конференция «Актуальные проблемы электронного приборостроения АПЭП-2008», Новосибирский Государственный Технический Университет, г. Новосибирск, 24–26 сентября 2008 г.
- Programme «Stochastic Processes in Communication Sciences», Isaac Newton Institute for Mathematical Sciences, Cambridge, 25 May – 15 June 2010.
- X Международная научно-техническая конференция «Актуальные проблемы электронного приборостроения АПЭП-2010», Но-

восибирский государственный технический университет, г. Новосибирск, 22–24 сентября 2010 г.

- III Всероссийский семинар «Фундаментальные основы МЭМС- и нанотехнологий». Новосибирск, 25–27 мая 2011 г.
- V International Conference «Limit Theorems in Probability Theory and Their Applications». Novosibirsk, August 15–21, 2011.
- 22nd Annual Conference of The International Environmetrics Society. Hyderabad, India, January 1–6, 2012.
- Applied methods of statistical analysis. Applications in survival analysis, reliability and quality control. Novosibirsk, September 26–30, 2013.
- 11 International conference on ordered statistical data. Bedlewo, Poland, June 2–6, 2014.
- XXI Всероссийская школа-коллоквиум по стохастическим методам. Кисловодск, 11–17 июня 2017 г.
- 13 International conference on ordered statistical data. Cadiz, Spain, May 22–25, 2018.

Кроме того, основные результаты диссертации докладывались неоднократно на:

- объединенном семинаре кафедры теории вероятностей и математической статистики НГУ и лаборатории теории вероятностей и математической статистики ИМ СО РАН под руководством академика А. А. Боровкова;
- научном семинаре кафедры высшей математики Новосибирского государственного технического университета под руководством профессора В. А. Селезнева;

- научных сессиях факультета прикладной математики и информатики НГТУ под руководством профессора Б. Ю. Лемешко;
- научном семинаре кафедры вычислительной техники Новосибирского государственного технического университета под руководством профессора В. В. Губарева.

Публикации

По теме диссертации автором опубликовано 48 печатных работ, в том числе 11 работ, индексируемых в базах цитирования (RSCI, SCOPUS, WoS). 9 работ опубликовано без соавторов.

Личный вклад автора

Диссертационная работа выполнена непосредственно ее автором.

В совместных работах [174], [180], [175] автору диссертации принадлежат вывод расчетных формул, реализация расчетов и интерпретация их результатов; в работах [164], [167], [168] — постановка задачи, доказательство утверждений и разработка алгоритмов, интерпретация результатов расчетов; в работах [163], [177], [178] — постановка задачи и доказательство утверждений; в работах [181], [182], [191], [194] — вывод расчетных формул и разработка алгоритмов, интерпретация результатов расчетов; в работе [197] — постановка задачи и разработка вычислительных алгоритмов.

Работа выполнялась в Новосибирском государственном техническом университете в период с 1999 по 2018 год.

Обзор литературы

Мы будем всюду далее предполагать, что доступная для анализа информация — это последовательность значений $X_1, \ldots, X_n$ — имеющиеся в распоряжении исследователя данные. Число n мы будем называть объемом имеющихся данных.

Модель выборки — основная модель математической статистики

— предполагает, что имеющиеся данные являются реализацией последовательности независимых одинаково распределенных случайных величин. Мы будем дополнительно предполагать, что X_i имеют конечную ненулевую дисперсию. Согласно функциональной центральной предельной теореме (принципу инвариантности Донскера — Прохорова) случайные ломаные, построенные по центрированным и нормированным суммам случайных величин, сходятся к стандартному винеровскому процессу (стандартному процессу обычного броуновского движения) [6]. Это гауссовский процесс с нулевым математическим ожиданием и линейно растущей дисперсией. Математическое описание явления броуновского движения было выведено из законов физики Эйнштейном в 1905 году, а его точная формализация на языке теории случайных процессов получена основателем кибернетики Н. Винером в его диссертации (1918 г.) и более поздних работах [162]. Дальнейшее изучение вероятностных свойств броуновского движения было предпринято П. Леви [137], [58], К. Ито и Г. Маккином [43]. Винеровский процесс — наиболее употребительная модель случайного процесса с непрерывными почти наверное траекториями (см. [137], [58], [43], [45], [132], [82] и списки литературы в них).

Однако порой встречаются явления, для которых винеровский процесс не годится в качестве модели. Мы рассмотрим две принципиально разных модели — разладки и фрактального гауссовского шума. Для модели разладки процесс, предельный для сумм при надлежащей нормировке, оказывается детерминированным и характеризуется моментом разладки T. Для модели фрактального гауссовского шума параметром служит скорость роста стандартных отклонений сумм — так называемый показатель Херста H. Предельный процесс для нормированных сумм в этом случае называется фрак-

тальным броуновским движением.

Моделью последовательности является модель разладки, если $X_1, \ldots, X_n$ независимы и имеют конечную ненулевую дисперсию, как и в модели выборки, но $X_1, \ldots, X_{[nT]}$ имеют распределение $\mathcal{F}_1$, а $X_{[nT]+1}, \ldots, X_n$ — распределение $\mathcal{F}_2$.

Исторически первой задачей, использующей модель разладки, была задача наискорейшего обнаружения разладки, состоящей в минимизации числа наблюдений после разладки, необходимых для ее обнаружения с заданными уровнем α и мощностью β. Для решения последней задачи А. Вальд разработал метод кумулятивных сумм [14]. Задача наискорейшего обнаружения разладки анализировалась в работах Клигене и Тельксниса [48], Торговицкого [84], Бродского и Дарховского [28], [11], [25].

обзор литературы по этой теме можно найти в монографии Никифорова [73], обзорной статье [13] и книге [114] Бродского и Дарховского. В статье Дарховского [26] задача наискорейшего обнаружения разладки решена при минимальных априорных предположениях.

В ряде приложений возникает задача апостериорного обнаружения разладки, то есть выяснения, изменила ли данная последовательность независимых случайных величин свои свойства за время наблюдения.

В случае, когда момент возможной разладки известен, задача о наличии разладки превращается в классическую задачу об однородности двух выборок ([9], §§59, 60). Мы будем рассматривать постановку задачи, в которой неизвестен момент возможной разладки. В ряде работ распределения до и после разладки считаются заданными. Так, в [111] предполагается, что распределения получаются друг из друга линейным сдвигом аргумента, причем плотность распределения предполагается известной, всюду положительной и непрерыв-

но дифференцируемой. В отличие от [111] мы будем предполагать лишь, что распределения до и после разладки имеют разные математические ожидания и конечные ненулевые дисперсии.

В работах Бродского и Дарховского [12], Васильченко [15] задача обнаружения разладки обобщается на случай многократного изменения характеристик случайной последовательности. В работе Дарховского и Пирятинской [29] предложен новый подход к выделению однородных фрагментов последовательности, основанный на эпсилон-сложности.

Другая постановка задачи — оценивание момента разладки по апостериорным данным. В задаче оценивания параметра, как правило, либо предполагаются известными распределения до и после разладки, либо предполагаются различными некоторые их числовые характеристики, например математические ожидания (см. [156] для подробной библиографии). В работах Бродского и Дарховского [24], [27], [10] предложены оценки момента разладки по апостериорным данным и проведено сравнение оценок.

Отметим в этой связи работы [115], [121], где задача оценивания параметра решена в наиболее общих предположениях: требуется лишь, чтобы распределения до и после разладки различались.

Моделью последовательности является фрактальный (дробный) гауссовский шум, если дисперсия ее частичных сумм растет в соответствии с некоторой степенной функцией, отличной от линейной. Параметр Херста H — это показатель скорости роста стандартного отклонения частичных сумм (корня из дисперсии). Эффект нелинейного роста стандартного отклонения сумм был обнаружен Херстом [128] для ряда природных явлений (речного стока, колец на деревьях и т. д.). Для оценивания параметра H Херст использовал метод нормированного размаха, подробно описанный нами в главе 2.

Описание этой процедуры оценивания и ряд примеров можно найти в работах [142], [53], [86].

Вероятностную модель для роста дисперсии в соответствии со степенным законом впервые получил основатель современной аксиоматики теории вероятностей А. Н. Колмогоров в 1940 году [49], однако его работа долго оставалась незамеченной. В 1968 году Мандельброт и ван Несс [144] переоткрыли его результаты и назвали соотвотствующий случайный процесс фрактальным броуновским движением. Фрактальный гауссовский шум — это последовательность приращений фрактального броуновского движения за равные интервалы времени.

Исследования свойств фрактального броуновского движения были осуществлены в работах [140], [141], [69], [116], [110].

Фрактальное броуновское движение оказывается предельным для процессов частных сумм скользящих средних. Соответствующие теоремы (в том числе с оценками скорости сходимости) доказаны в [23], [134], [4] и будут нами использоваться для моделирования процесса фрактального броуновского движения в главе 1.

Применение модели фрактального броуновского движения к теории массового обслуживания можно найти в работах [146], [150].

Моделирование фрактального броуновского движения рассмотрено в работах [143], [160], [154], [155], [153], [80], [23], [134], [4]. Метод, предложенный в [143] — метод дискретизации интеграла Мандельброта — ван Несса. В [160] предложен метод срединного смещения. В [53], §9.3, отмечено, что этот алгоритм приводит к процессу с нестационарными приращениями, отличному от фрактального броуновского движения. В [154], [155], [153], [80] описан алгоритм случайных фаз. Нами показано в [177], что этот алгоритм приводит к случайному процессу — процесу Сопа, который, как и фрактальное бро-

уновское движение, имеет стационарные приращения, но, в отличие от него, сам является стационарным, и не обладает степенной скоростью роста дисперсии — его дисперсия в силу стационарности постоянна. Этот факт не отмечен нигде в известной нам литературе. В [23], [134], [4] для моделирования предлагается использовать процесс скользящих средних, и доказаны соответствующие теоремы сходимости: в [23] — при завышенных моментных ограничениях в регулярном случае (то есть при $H \geq 1/2$), в [134] сняты завышенные моментные ограничения, в [4] рассмотрен случай произвольных H и получены оценки скорости сходимости.

Методы оценивания параметра H, отличные от метода нормированного размаха, рассматриваются в [53], [113], [106], [124]. Метод дисперсии состоит в оценке скорости роста стандартного отклонения у группированных данных. Метод знаков оценивает параметр H с помощью подсчета частоты перемены знака приращениями процесса. Отметим, что знаковый метод использован в статье [83] для проверки гипотезы о нулевом математическом ожидании выборки из нормального распределения.

Одно из приложений описанных моделей — исследование наличия разладки и фрактальности в текстах. В частности, исследование разладки должно показывать, написан текст одним или несколькими авторами. Задача определения разладки близка к задаче атрибуции (определения авторства) текста, которая решалась рядом авторов.

Для правильной атрибуции текстов важным является метод, на основании которого тексту сопоставляется последовательность чисел. Этот метод должен удовлетворять требованиям, описанным в [89]: для разных авторов давать существенно разные значения, а для одного и того же автора давать близкие значения. Исторически

первый подход состоит в подсчете среднего числа слогов в слове или среднего числа слов в предложении [158], [70], [127], [17], [18], [101], [66], [67]. Близкий к нему подход состоит в подсчете числа предложений или слов определенной длины [77], [78], [97]. Структурный подход связан с анализом структуры предложений [71], [7]. Подход, основанный А. А. Марковым, изучает зависимости появления буквенных сочетаний в тексте, изложение его можно найти в [33], а применение к анализу текстов — в [90], [91], [54]. Для нас будет наиболее важен подход, изложенный в [89] — сравниваются частоты служебных слов в текстах. Авторами показано, что этот подход наиболее соответствует требованиям к авторскому инварианту, то есть частота появления служебных слов устойчива для конкретного автора и различается для разных авторов. Частотный и марковский подходы используются и сравниваются в [81].

Для принятия решения об атрибуции используются статистические процедуры [77], [97], [5], [81], [66], [67], [7]. Кроме того, применяются подсчет энтропии цепи Маркова и коэффициент сжатия программами архивирования [90], [91], [54]; хаотические нейронные сети и карты Кохонена [78], [97], [5].

Диссертация построена следующим образом. В главе 1 сравниваются критерии наличия разладки в модели выборки. Доказано, что в широком классе критериев наилучшим является критерий, использующий в качестве статистики супремум модуля эмпирического моста. В главе 2 разрабатываются и сравниваются различные оценки параметров фрактального гауссовского шума. В главе 3 строятся и анализируются критерии обнаружения разладки в модели фрактального гауссовского шума. В главе 4 разработанные в предыдущих главах критерии применяются к анализу текстов на естественном языке. Сначала устанавливается расхождение поведения одно-

мерных статистик текста с теоретическим поведением в модели выборки. Затем отыскивается разладка для склейки текстов разных авторов. Наконец, используется модель фрактального гауссовского шума. В главе 5 разрабатываются и сравниваются статистические критерии обнаружения разладки регрессии с циклическим трендом. В главе 6 рассматриваются применения статистических критериев, основанных на эмпирическом мосте, к выбору вероятностных моделей на основании фактических данных.

ГЛАВА 1
Обнаружение разладки в модели выборки

1.1 Вводные замечания

В ряде приложений возникает задача апостериорного обнаружения разладки, то есть выяснения, изменила ли данная последовательность независимых случайных величин свои свойства за время наблюдения. Эта задача существенно отличается от задачи наискорейшего обнаружения разладки, состоящей в минимизации числа наблюдений после разладки, необходимых для ее обнаружения с заданными уровнем α и мощностью β. Для решения последней задачи был разработан метод кумулятивных сумм [14], обзор литературы по этой теме можно найти в [73].

В случае, когда момент возможной разладки известен, задача о наличии разладки превращается в классическую задачу об однородности двух выборок ([9], §§59, 60). Мы будем рассматривать постановку задачи, в которой неизвестен момент возможной разладки. В литературе распределения до и после разладки считаются заданными. Так, в [111] предполагается, что распределения получаются друг из друга линейным сдвигом аргумента, причем плотность распределения предполагается известной, всюду положительной и непрерывно дифференцируемой. В отличие от [111] мы будем предполагать лишь, что распределения до и после разладки имеют разные математические ожидания и конечные ненулевые дисперсии.

Еще один круг работ имеет отношение к постановке задачи — работы по оцениванию момента разладки по апостериорным данным. В задаче оценивания параметра, как правило, либо предполагаются известными распределения до и после разладки, либо предполагаются различными некоторые их числовые характеристики, например

математические ожидания (см. [156] для подробной библиографии).

Отметим в этой связи работу [161], где задача оценивания параметра решена в наиболее общих предположениях о различии распределений до и после разладки.

Введем следующие обозначения. Пусть $\{\xi_i^{(1)}\}_{i=1}^{\infty}$ и $\{\xi_i^{(2)}\}_{i=1}^{\infty}$ — две взаимно независимых последовательности независимых одинаково распределенных случайных величин с распределениями $\mathcal{F}_1$ и $\mathcal{F}_2$ соответственно.

Схема серий случайных величин $\{X_i^{(n)}\}_{i=1}^{n}$ задается следующим образом:

$X_i^{(n)} = \xi_i^{(1)}$ при $1 \le i \le [nT]$;

$X_i^{(n)} = \xi_i^{(2)}$ при $[nT] + 1 \le i \le n$.

Здесь T — неизвестный параметр, $0 < T < 1$.

Нулевая гипотеза H состоит в том, что разладки не произошло, то есть $\mathcal{F}_1 = \mathcal{F}_2$. Ее альтернатива — в том, что не только распределения, но и их математические ожидания различны. Как при нулевой гипотезе, так и при альтернативе мы будем предполагать, что дисперсии распределений $\mathcal{F}_1$ и $\mathcal{F}_2$ конечны и положительны.

Каждый из рассматриваемых критериев проверки гипотез основан на последовательности статистик $V_1,\ V_2,\ \ldots$. Гипотеза H принимается, если статистика не превосходит некоторого уровня.

Для решения задачи обнаружения разладки строятся статистики, являющиеся функционалами от *эмпирического моста* $Z_n = \{Z_n(t), 0 \le t \le 1\}$ — случайной ломаной, построенной по точкам

$$\left(\frac{k}{n};\ \frac{nS_k - kS_n}{sn\sqrt{n}}\right), k = 0, \ldots, n, \tag{1}$$

где $S_n = \sum_{i=1}^{n} X_i^{(n)} = n\overline{X}$, $s^2 = \overline{X^2} - (\overline{X})^2$.

Доказывается, что для любой простой гипотезы θ_0 из H эмпирический мост C-сходится к стандартному броуновскому мосту W^0, то

есть для любого заданного на $C(0;1)$ непрерывного в равномерной метрике функционала g имеет место сходимость по распределению

$$g(Z_n) \stackrel{\mathbf{P}_{\theta_0}}{\Rightarrow} g(W^0).$$

Также доказывается, что для любой простой гипотезы θ из альтернативы эмпирический мост сходится почти наверное в равномерной метрике на отрезке $[0;\ 1]$ к детерминированному процессу $z_\theta = \{z_\theta(t),\ 0 \le t \le 1\}$, то есть для любого заданного на $C(0;1)$ непрерывного в равномерной метрике функционала J имеет место сходимость почти наверное

$$J(Z_n)/\sqrt{n} \to J(z_\theta)\ \ (\mathbf{P}_\theta) - \text{п.н.}$$

Процесс z_θ — кусочно линейный, с единственным изломом в точке T.

В качестве функционалов рассматриваются различные нормы эмпирического моста, а также нормированные взвешенные суммы случайных величин. Доказано, что нормированные взвешенные суммы с точностью до малого слагаемого (сходящегося п. н. к нулю при выполнении как основной гипотезы, так и альтернативы) представимы в виде интегральных функционалов от эмпирического моста.

В случае, когда последовательность статистик J_n сходится по распределению к случайной величине J с функцией распределения F, для нее определен *приближенный бахадуровский наклон* $c(\theta)$ равенством

$$c(\theta) = 2 \lim_{n\to\infty} (-n^{-1} \ln(1 - F(J_n)))\ \ (\mathbf{P}_\theta - \text{п. н.}) \qquad (2)$$

(см. [72], §1.4). Это определение и будет нами использоваться для сравнения критериев: чем больше $c(\theta)$, тем лучше критерий различает гипотезы. Отметим, что $c(\theta)$ в пределе (при $\theta \to \theta_0$) обратно пропорционально числу испытаний n, необходимому для достиже-

ния заданного размера критерия α и заданной мощности β при близких гипотезах $\mathcal{F}_1$ и $\mathcal{F}_2$. Этот подход, изобретенный Э. Питменом, изложен в [75] (глава 7) для статистик, имеющих в пределе нормальное распределение. Для случая, когда предельное распределение абсолютно непрерывно, но отлично от нормального, модификация подхода Питмена введена Х. С. Виэндом в [161]. Для применения подхода Э. Питмена надо проверить, что статистика является виэндовской, то есть при выполнении альтернативной гипотезы сходится к предельному значению достаточно быстро. Подробное изложение этих результатов можно найти в книге Я. Ю. Никитина [72] (глава 1).

В §1.2 приводится постановка задачи. В §1.3 рассматриваются статистики, основанные на L_p-нормах эмпирического моста, в §1.4 — статистики взвешенных сумм. Сравнение полученных критериев проведено в §1.5. Краткое изложение полученных результатов содержится в §1.6.

1.2 Постановка задачи

Заданы $\{\xi_i^{(1)}\}_{i=1}^{\infty}$ и $\{\xi_i^{(2)}\}_{i=1}^{\infty}$ — две взаимно независимых последовательности независимых одинаково распределенных случайных величин.

Случайные величины $\xi_i^{(1)}$ имеют распределение $\mathcal{F}_1$ с математическим ожиданием m_1 и дисперсией $0 < {\sigma_1}^2 < \infty$.

Случайные величины $\xi_i^{(2)}$ имеют распределение $\mathcal{F}_2$ с математическим ожиданием m_2 и дисперсией $0 < {\sigma_2}^2 < \infty$.

Схема серий случайных величин $\{X_i^{(n)}\}_{i=1}^{n}$ задается следующим образом:

$X_i^{(n)} = \xi_i^{(1)}$ при $1 \le i \le [nT]$;

$X_i^{(n)} = \xi_i^{(2)}$ при $[nT] + 1 \le i \le n$.

Предполагается, что T — неизвестная константа, $0 < T < 1$.

Простыми гипотезами θ здесь являются тройки $\theta = (\mathcal{F}_1,\ \mathcal{F}_2,\ T)$. Будем предполагать, что все множество гипотез $\Theta = \Theta_0 \cup \Theta_1$, где $\Theta_0 = \{\theta :\ \mathcal{F}_1 = \mathcal{F}_2,\ m_1 \neq 0\}$, $\Theta_1 = \{\theta :\ m_1 \neq m_2\}$. Мы будем строить критерии, различающие сложные гипотезы Θ_0 и Θ_1.

Будем использовать обозначения X_i вместо $X_i^{(n)}$ там, где это не приводит к противоречиям.

Замечание 1.1 *Сформулированные выше условия на случайные величины* $X_1^{(n)},\ \ldots,\ X_n^{(n)}$ *близки к необходимым условиям состоятельности критерия проверки гипотез. Если согласно первой гипотезе дисперсия равна 0, это приводит к вырожденной статистической задаче. Случаев бесконечной дисперсии или выполнения равенств* $m_1 = m_2$, $m_1 = 0$ *можно избежать путем подходящих преобразований фазового пространства.*

Критерий принимает вторую гипотезу в случае, когда $J_n \geq C$, где J_n — статистика, определяющая критерий.

В дальнейшем рассматриваются различные классы статистик: взвешенные суммы и L_p-нормы.

Введем топологию, порожденную полуметрикой модуля разности математических ожиданий до и после разладки: $||\theta|| = |m_1 - m_2|$. В этом случае $\partial\Theta_0 = \{\theta :\ m_1 = m_2\} = \Theta_0$.

Для сравнения критериев будем вычислять приближенные бахадуровские наклоны, определяемые равенством (2). Отметим, что в случае сходимости почти наверное (как в (2)) говорят о сильных наклонах, а в случае сходимости по вероятности — о слабых. Мы докажем, что для всех рассматриваемых последовательностей статистик есть сходимость почти наверное.

Мы будем говорить, что один критерий *лучше* другого в смысле используемого подхода, если для всех $\theta \in \Theta_1$ из некоторой окрестности $\partial\Theta_0$ значение $c(\theta)$ приближенного бахадуровского наклона

(определяемого равенством (2)) для статистики первого критерия больше, чем для второго.

Лемма 1.1 *Если для любой* $\theta_0 \in \Theta_0$ *имеет место сходимость к одному и тому же распределению* $\mathbf{P}_{\theta_0}\{J_n < t\} \to F(t)$, *а для любой* $\theta \in \Theta_1$ *— сходимость* $J_n/\sqrt{n} \to j(\theta)$ *(*$\mathbf{P}_\theta$*-п. н.), причем*

$$\ln(1 - F(t)) \sim -\frac{1}{2}at^2 \tag{3}$$

при $t \to \infty$, *то*

$$c(\theta) = aj^2(\theta) \ (\mathbf{P}_\theta - п.\ н.) \tag{4}$$

Доказательство.

Согласно формуле (2) ,

$$c(\theta) = 2 \lim_{n\to\infty}(-n^{-1}\ln(1 - F(J_n)) \ \ (\mathbf{P}_\theta - \text{п. н.})$$

В силу сделанных предположений,

$$c(\theta) = 2 \lim_{n\to\infty}(n^{-1}\frac{1}{2}aJ_n^2) = aj^2(\theta) \ \ (\mathbf{P}_\theta - \text{п. н.})$$

Доказательство завершено.

1.3 Применение принципа инвариантности

Рассмотрим эмпирический мост $Z_n = \{Z_n(t),\ 0 \le t \le 1\}$ — случайную ломаную, построенную по точкам, определенным формулой (1). Понятие эмпирического моста возникло в [174] при анализе неоднородности энергопотребления в течение суток. По определению,

$$Z_n(t) = \frac{nS_k - kS_n}{sn\sqrt{n}} + \frac{nX_{k+1} - S_n}{s\sqrt{n}}\left(t - \frac{k}{n}\right),$$

$$\frac{k}{n} \le t < \frac{k+1}{n}, \quad k = 0, \ldots, n-1.$$

Лемма 1.2 *Для любой* $\theta_0 \in \Theta_0$ *имеет место* C*-сходимость эмпирического моста* Z_n *к броуновскому мосту* W^0.

Доказательство.

Представим $Z_n(t)$ в виде

$$Z_n(t) = Z_n^*(t) - tZ_n^*(1),$$

где Z_n^* — случайная ломаная, построенная по точкам

$$\left(\frac{k}{n};\ \frac{S_k - km_1}{s\sqrt{n}}\right), \quad k = 0, \ldots, n.$$

Пусть Z_n^{**} — случайная ломаная, построенная по точкам

$$\left(\frac{k}{n};\ \frac{S_k - km_1}{\sigma_1\sqrt{n}}\right), k = 0, \ldots, n.$$

В силу принципа инвариантности (см. [6], [8]) процесс Z_n^{**} будет C-сходиться к стандартному винеровскому процессу W.

Так как $s \to \sigma_1$ ($\mathbf{P}_{\theta_0}$-п.н.), то Z_n^* будет C-сходиться к стандартному винеровскому процессу W.

В силу свойств слабой сходимости Z_n будет C-сходиться к

$$W^0 = \{W^0(t), 0 \le t \le 1\},$$

где $W^0(t) = W(t) - tW(1)$.

Доказательство завершено.

Лемма 1.3 *Для любой $\theta \in \Theta_1$ процесс $Z_n/\sqrt{n}$ сходится ($\mathbf{P}_\theta$-п. н.) в равномерной метрике на отрезке $[0;\ 1]$ к функции z_θ:*

$$z_\theta(t) = \begin{cases} \frac{tM_\theta}{T}, & 0 \le t \le T; \\ \frac{(1-t)M_\theta}{1-T}, & T \le t \le 1, \end{cases}$$

$$M_\theta = \frac{(m_1 - m_2)T(1-T)}{\sigma_\theta},$$

$$\sigma_\theta = \sqrt{T({m_1}^2 + {\sigma_1}^2) + (1-T)({m_2}^2 + {\sigma_2}^2) - (Tm_1 + (1-T)m_2)^2}.$$

Доказательство.

Докажем, что

$$\sup_{0\le t\le T}|Z_n(t)/\sqrt{n}-z(t)|\to 0$$

п. н. Так как в условиях теоремы $s\to\sigma_T$ п. н., то достаточно доказать, что

$$\max_{1\le k\le [nT]}\left|\frac{nS_k-kS_n}{n^2}-\frac{k}{n}(m_1-m_2)(1-T)\right|\to 0$$

п. н. Эта сходимость следует из неравенств

$$\max_{1\le k\le [nT]}\left|\frac{nS_k-kS_n}{n^2}-\frac{k}{n}(m_1-m_2)(1-T)\right|$$

$$\le \max_{1\le k\le [nT]}\left|\frac{S_k-km_1}{n}\right|+\max_{1\le k\le [nT]}\left|\frac{k}{n}\cdot\frac{S_n-m_1T-m_2(1-T)}{n}\right|$$

$$\le \max_{1\le k\le [nT]}\left|\frac{S_k-km_1}{n}\right|+\left|\frac{S_{[nT]}-m_1T}{n}\right|+\left|\frac{S_n-S_{[nT]}-m_2(1-T)}{n}\right|.$$

Последние 2 слагаемых сходятся п. н. к нулю в силу усиленного закона больших чисел. Докажем сходимость п. н. к нулю первого слагаемого: пусть $S_n^0=S_n-nm_1$ — сумма центрированных слагаемых. Сходимость

$$\max_{1\le k\le [nT]}\left|\frac{S_k^0}{n}\right|\to 0$$

п. н. следует из того, что для любого фиксированного $N_0\ge 1$ имеет место

$$\max_{1\le k\le [nT]}\left|\frac{S_k^0}{n}\right|\le\frac{N_0}{n}\max_{1\le k\le N_0}\left|\frac{S_k^0}{N_0}\right|+\max_{N_0\le k\le [nT]}\left|\frac{S_k^0}{n}\right|\to 0$$

п. н., так как

$$\frac{N_0}{n}\max_{1\le k\le N_0}\left|\frac{S_k^0}{N_0}\right|\to 0$$

п. н. при $n\to\infty$, и

$$\max_{N_0\le k\le [nT]}\left|\frac{S_k^0}{n}\right|\le\sup_{k\ge N_0}\left|\frac{S_k^0}{k}\right|\to 0$$

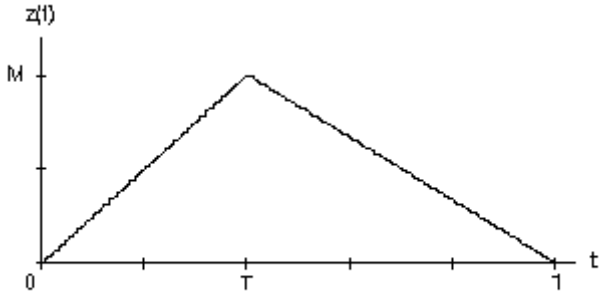

Рис. 1.1. График процесса z_θ. Здесь T — момент разладки, $M = M_\theta$ — экстремальное значение процесса.

п. н. при $N_0 \to \infty$ согласно УЗБЧ.

Аналогично

$$\sup_{T \le t \le 1} |Z_n(t)/\sqrt{n} - z(t)| \to 0$$

п. н. Лемма доказана.

Предельный процесс z_θ имеет график, изображенный на рис. 1.1.

Представляется логичным рассматривать различные нормы случайной функции Z_n на отрезке $[0;\ 1]$, а также размах этой функции на отрезке. Обозначим

$$J_n^{(r)} = ||Z_n||_{L_r} = \left(\int_0^1 |Z_n(t)|^r dt \right)^{1/r};$$

$$J_n^{\infty} = ||Z_n||_{L_\infty} = \sup_{t \in [0;\ 1]} |Z_n(t)|;$$

$$J_n^{R} = \sup_{t \in [0;\ 1]} Z_n(t) - \inf_{t \in [0;\ 1]} Z_n(t).$$

Наряду с L_r-нормами эмпирического процесса рассмотрим функционалы вида

$$J_n^{|r|} = \left| \int_0^1 (Z_n(t))^r dt \right|^{1/r}. \qquad (5)$$

При четном r эти функционалы совпадают с L_r-нормами.

При выполнении первой гипотезы распределение статистики J_n^∞ слабо сходится к распределению Колмогорова, J_n^R — к распределению размаха броуновского моста, а $J_n^{(2)}$ — к распределению корня квадратного из случайной величины, имеющей распределение омега-квадрат.

Отметим, что критерии, основанные на функционалах размаха и L_r-нормы, оказываются при выборе из альтернатив Θ_0 и Θ_1 асимптотически не лучше критериев, основанных на функционалах J_n^∞ и $J_n^{|r|}$ соответственно:

Теорема 1.1 *Статистика J_n^R не лучше статистики J_n^∞. Статистика $J_n^{(r)}$ не лучше статистики $J_n^{|r|}$.*

Доказательство.

В силу леммы 1.3 для $\theta \in \Theta_1$ имеют место сходимости

$$J_n^R/\sqrt{n} \to |M_\theta|, \qquad J_n^\infty/\sqrt{n} \to |M_\theta|$$

($\mathbf{P}_\theta$-п.н.). В то же время

$$J_n^R \geq J_n^\infty$$

почти наверное просто по определению, и при выполнении нулевой гипотезы для предельных распределений имеет место неравенство

$$\mathbf{P}\{\sup_{t\in[0;\ 1]} W^0(t) - \inf_{t\in[0;\ 1]} W^0(t) \geq x\} \geq \mathbf{P}\{\sup_{t\in[0;\ 1]} W^0(t) \geq x\}.$$

Согласно формуле (2) статистика J_n^R не лучше статистики J_n^∞.

Аналогично для интегральных функционалов: для $\theta \in \Theta_1$

$$J_n^{(r)}/\sqrt{n} \to \left(\int_0^1 |z_\theta(t)|^r dt\right)^{1/r} = \left|\int_0^1 (z_\theta(t))^r dt\right|^{1/r},$$

$$J_n^{|r|}/\sqrt{n} \to \left|\int_0^1 (z_\theta(t))^r dt\right|^{1/r}$$

($\mathbf{P}_\theta$-п.н.). В то же время

$$\mathbf{P}\{\left(\int_0^1 |W^0(t)|^r dt\right)^{1/r} \geq x\} \geq \mathbf{P}\{\left|\int_0^1 (W^0(t))^r dt\right|^{1/r} \geq x\}.$$

Доказательство завершено.

1.4 Взвешенные суммы

Рассмотрим статистику

$$J_n = \left|\frac{\sum_{k=1}^n h_{k,n} X_k}{s\sqrt{n}}\right|.$$

Здесь $h_{1,n}, \ldots, h_{k,n}$ — весовые коэффициенты. Они могут выбираться весьма произвольно, в частности, их знаки могут чередоваться, что ухудшает качество статистического критерия. Мы будем предполагать, что весовые коэффициенты заданы следующим регулярным образом:

$$h_{k,n} = g(k/n),$$

где $g(t)$ — функция ограниченной вариации на $[0,\ 1]$. Для введенных с помощью функции g весовых коэффициентов будем использовать обозначение статистики $J_n = J_n(g)$.

Теорема 1.2 *Для любой $\theta_0 \in \Theta_0$ статистика $J_n(g)$ сходится по распределению к модулю стандартного нормального закона в том и только том случае, когда выполнены условия*

$$\int_0^1 g(t)dt = 0, \qquad \int_0^1 g^2(t)dt = 1. \tag{6}$$

Доказательство.

Последовательность случайных величин $\{h_{k,n}X_k\}$ удовлетворяет условиям теоремы Линдеберга (см., например, [8], глава 16, параграф 7), поэтому

$$\frac{\sum_{k=1}^n h_{k,n}X_k - \mathbf{E}\left(\sum_{k=1}^n h_{k,n}X_k\right)}{\sqrt{\mathbf{D}\left(\sum_{k=1}^n h_{k,n}X_k\right)}}$$

сходится по распределению к стандартному нормальному закону.

Заметим, что

$$\frac{1}{n}\mathbf{D}\left(\sum_{k=1}^{n} h_{k,n}X_k\right) = \sum_{k=1}^{n} \frac{{h_{k,n}}^2}{n}{\sigma_1}^2 \to {\sigma_1}^2 \int_0^1 g^2(t)dt.$$

Так как

$$\mathbf{E}\left(\sum_{k=1}^{n} h_{k,n}X_k\right) = m_1 \sum_{k=1}^{n} g(k/n),$$

и $s \to \sigma_1$ ($\mathbf{P}_{\theta_0}$-п.н.), то $J_n(g)$ сходится по распределению к модулю стандартного нормального закона тогда и только тогда, когда

$$\frac{\sum_{k=1}^{n} g(k/n)X_k - m_1 \sum_{k=1}^{n} g(k/n)}{\sigma_1\sqrt{n}}$$

сходится по распределению к стандартному нормальному закону, то есть

$$\frac{1}{n{\sigma_1}^2}\mathbf{D}\left(\sum_{k=1}^{n} h_{k,n}X_k\right) \to 1$$

что равносильно второму из условий теоремы, и

$$\frac{m_1}{\sigma_1\sqrt{n}} \sum_{k=1}^{n} g(k/n) \to 0,$$

что равносильно первому из условий теоремы, так как по теореме о среднем существуют такие $r_k = r_k(n)$, что $r_k \in [k-1;\ k]$ и

$$\int_0^1 g(t)dt = \frac{1}{n}\sum_{k=1}^{n} g(r_k/n),$$

и поэтому

$$\sup_n \left|\sum_{k=1}^{n} g(k/n) - n\int_0^1 g(t)dt\right| \leq$$

$$\leq \sup_n \sum_{k=1}^{n} |g(k/n) - g(r_k/n)| \leq \int_0^1 |dg(t)| < \infty.$$

Теорема доказана.

Теорема 1.3 *Пусть g — функция ограниченной вариации, удовлетворяющая условиям (6) теоремы 1.2. Тогда статистику $J_n(g)$*

можно представить в виде интеграла Стилтъеса

$$J_n(g) = \left| \int_0^1 Z_n(t) dg(t) + \nu_n \right|,$$

где $\nu_n \to 0$ *($\mathbf{P}_\theta$-п.н.) для любой* $\theta \in \Theta$.

Доказательство.

По определению

$$J_n(g) = \left| \frac{\sum_{k=1}^n g(k/n) X_k}{s\sqrt{n}} \right|,$$

$$\begin{aligned}
&\int_0^1 Z_n(t) dg(t) = \\
&= \sum_{k=0}^{n-1} \int_{\frac{k}{n}}^{\frac{k+1}{n}} \frac{nS_k - kS_n}{sn\sqrt{n}} dg(t) + \\
&+ \sum_{k=0}^{n-1} \int_{\frac{k}{n}}^{\frac{k+1}{n}} \frac{nX_{k+1} - S_n}{s\sqrt{n}} \left(t - \frac{k}{n} \right) dg(t) = \\
&= \sum_{k=0}^{n-1} \int_{\frac{k}{n}}^{\frac{k+1}{n}} \left(\frac{S_k}{s\sqrt{n}} - \frac{k}{n} \frac{S_n}{s\sqrt{n}} \right) dg(t) + \\
&+ \sum_{k=0}^{n-1} \int_{\frac{k}{n}}^{\frac{k+1}{n}} \frac{nX_{k+1} - S_n}{s\sqrt{n}} \left(t - \frac{k}{n} \right) dg(t) = \\
&= \frac{1}{s\sqrt{n}} \sum_{k=0}^{n-1} \left(S_k - \frac{k}{n} S_n \right) \left(g\left(\frac{k+1}{n} \right) - g\left(\frac{k}{n} \right) \right) + \\
&+ \sum_{k=0}^{n-1} \int_{\frac{k}{n}}^{\frac{k+1}{n}} \frac{nX_{k+1} - S_n}{s\sqrt{n}} \left(t - \frac{k}{n} \right) dg(t) = \\
&= \frac{1}{s\sqrt{n}} \left(\sum_{k=1}^{n} \left(S_{k-1} - \frac{k-1}{n} S_n \right) g\left(\frac{k}{n} \right) - \sum_{k=1}^{n} \left(S_k - \frac{k}{n} S_n \right) g\left(\frac{k}{n} \right) \right) + \\
&+ \sum_{k=0}^{n-1} \int_{\frac{k}{n}}^{\frac{k+1}{n}} \frac{nX_{k+1} - S_n}{s\sqrt{n}} \left(t - \frac{k}{n} \right) dg(t) = \\
&= \frac{1}{s\sqrt{n}} \sum_{k=1}^{n} g\left(\frac{k}{n} \right) \left(-X_k + \frac{S_n}{n} \right) + \\
&+ \sum_{k=0}^{n-1} \int_{\frac{k}{n}}^{\frac{k+1}{n}} \frac{nX_{k+1} - S_n}{s\sqrt{n}} \left(t - \frac{k}{n} \right) dg(t) =
\end{aligned}$$

$$= \frac{1}{s\sqrt{n}} \left(\frac{S_n}{n} \sum_{k=1}^{n} g\left(\frac{k}{n}\right) - \sum_{k=1}^{n} g\left(\frac{k}{n}\right) X_k \right) +$$
$$+ \sum_{k=0}^{n-1} \int_{\frac{k}{n}}^{\frac{k+1}{n}} \frac{nX_{k+1} - S_n}{s\sqrt{n}} \left(t - \frac{k}{n}\right) dg(t).$$

Итак,

$$\nu_n = -\frac{S_n}{sn} \frac{\sum_{k=1}^{n} g(\frac{k}{n})}{\sqrt{n}} - \sum_{k=0}^{n-1} \int_{\frac{k}{n}}^{\frac{k+1}{n}} \frac{nX_{k+1} - S_n}{s\sqrt{n}} \left(t - \frac{k}{n}\right) dg(t).$$

В силу УЗБЧ и условия теоремы получаем, что

$$\frac{S_n}{sn} \frac{\sum_{k=1}^{n} g(\frac{k}{n})}{\sqrt{n}} \to 0$$

п. н. с ростом n.

Справедлива оценка

$$\left| \sum_{k=0}^{n-1} \int_{\frac{k}{n}}^{\frac{k+1}{n}} \frac{nX_{k+1} - S_n}{s\sqrt{n}} \left(t - \frac{k}{n}\right) dg(t) \right| \leq$$
$$\leq \frac{\max_{k=0}^{n-1} |X_{k+1}| + |S_n| \int_0^1 d|g(t)|}{sn\sqrt{n}} \leq$$
$$\leq \frac{\sum_{k=1}^{n} |X_k| \left(1 + \int_0^1 d|g(t)|\right)}{sn\sqrt{n}}.$$

Правая часть неравенства сходится п. н. к нулю в силу усиленного закона больших чисел.

Теорема доказана.

Следствие 1.1 *Для любой $\theta \in \Theta_1$ имеет место сходимость*

$$\frac{J_n(g)}{\sqrt{n}} \to \left| \frac{m_1 \int_0^T g(t)dt + m_2 \int_T^1 g(t)dt}{\sigma_\theta} \right|$$

($\mathbf{P}_\theta$-п. н.)

Доказательство.

Так как $\int_0^1 Z_n(t)dg(t)$ — непрерывный в равномерной метрике функционал от Z_n, лемма 1.3 обосновывает сходимость ($\mathbf{P}_\theta$-п. н.) $\int_0^1 Z_n(t)dg(t)/\sqrt{n} \to \int_0^1 z_\theta(t)dg(t)$. Интегрирование по частям дает

$$\int_0^1 z_\theta(t)dg(t) = \frac{m_1 \int_0^T g(t)dt + m_2 \int_T^1 g(t)dt}{\sigma_\theta}.$$

Применяя теорему 1.3, получаем требуемое утверждение.

Следствие доказано.

1.5 Сравнение критериев

Напомним, что, согласно введенной терминологии, критерий тем *лучше*, чем больше его приближенный бахадуровский наклон, вычисление которого основано на лемме 1.1.

Для сравнения статистик, введенных в предыдущем параграфе, рассмотрим байесовскую постановку задачи — будем предполагать, что T — случайная величина с известной функцией распределения $F_T(t)$.

Для формулировки следующей теоремы обозначим

$$s_k(u) = \sin(2\pi k u), \quad c_k(u) = \cos(2\pi k u).$$

Теорема 1.4 *Если известна функция распределения F_T случайной величины T, то функция*

$$g(t) = -\sqrt{2}\sum_{k=1}^{\infty} \frac{k^{-1}\left(c_k(u)\int\limits_0^1 s_k(u)dF_T(u) + s_k(t)\int\limits_0^1 (1-c_k(u))dF_T(u)\right)}{\sqrt{\sum\limits_{j=1}^{\infty} j^{-2}\left(\left(\int\limits_0^1 s_k(u)dF_T(u)\right)^2 + \left(\int\limits_0^1 (1-c_k(u))dF_T(u)\right)^2\right)}}$$

определяет наилучший критерий в классе критериев, удовлетворяющих условиям теоремы 1.2.

Доказательство.

В силу теоремы 1.2, при $\theta_0 \in \Theta_0$ имеет место сходимость по распределению к стандартному нормальному закону. Поэтому в силу леммы 1.1 требуется найти функцию g, на которой достигается максимум предела последовательности $J_n(g)/\sqrt{n}$.

Согласно следствию 1.1, нужно решить следующую оптимизационную задачу: найти функцию g, максимизирующую выражение

$$\left|(m_1 - m_2)\int_0^1 \left(\int_0^u g(t)dt\right) dF_T(u)\right|,$$

где

$$g(t) = \sum_{k=1}^{\infty} (a_k \cos(2\pi kt) + b_k \sin(2\pi kt))$$

(слагаемое $a_0 = 0$ в силу условия $\int_0^1 g(t)dt = 0$) при ограничении

$$\sum_{k=1}^{\infty} ({a_k}^2 + {b_k}^2) = 2$$

(равенство Парсеваля, соответствующее условию $\int_0^1 g^2(t)dt = 1$).

Будем искать функцию $g(t)$ такую, что

$$\int_0^1 \left(\int_0^u g(t)dt\right) dF_T(u) \to \max$$

(функция $-g(t)$ дает симметричное решение).

Для отыскания условного экстремума составляем функцию Лагранжа, дифференцируем по a_k и b_k, приравниваем частные производные к 0 и используем условие нормировки (равенство Парсеваля). Приходим к решению, приведенному в формулировке теоремы.

Теорема доказана.

Следствие 1.2 *При известном T функция*

$$g_T(t) = \begin{cases} -K_1, & 0 \le t < T; \\ K_2, & T \le t \le 1 \end{cases}$$

определяет наилучший критерий в классе критериев, удовлетворяющих условиям теоремы 1.2. Здесь K_1 и K_2 определяются как

решения системы уравнений

$$K_1 T = K_2(1-T);$$

$$K_1{}^2 T + K_2{}^2(1-T) = 1,$$

то есть

$$K_1 = \sqrt{T/(1-T)}, \qquad K_2 = \sqrt{(1-T)/T}.$$

Доказательство.

Подставляя в формулировку теоремы 1.4 функцию распределения $F_T(t) = \mathbf{I}\{T < t\}$, получаем разложение в ряд Фурье ступенчатой функции, приведенной в формулировке теоремы.

Доказательство завершено.

Замечание 1.2 *Важно отметить, что функция $g(t)$ не зависит от значений m_1, m_2. Однако она зависит от значения T, которое в ряде практических задач неизвестно. Следствие 1.2 показывает, что не существует равномерного по T наилучшего критерия. В случае $m_1 = 0$ к асимптотически оптимальному критерию приводит другая функция $g(t)$, зависящая от m_2.*

Следствие 1.3 *Если T имеет равномерное распределение на $[0;\ 1]$, то функция $g^{(u)}(t) = \sqrt{12}(t - 1/2)$, $t \in [0;\ 1]$, обеспечивает наилучший критерий.*

Доказательство.

Подставляя $\int_0^1 \sin(2\pi k u)du = 0$ и $\int_0^1 (1 - \cos(2\pi k u))du = 1$ в утверждение теоремы 1.4, получаем

$$g(t) = -\sum_{k=1}^{\infty} \frac{\sqrt{2}\sin(2\pi k t)}{k\sqrt{\sum_{j=1}^{\infty} j^{-2}}} =$$

$$= -\sum_{k=1}^{\infty} \frac{\sqrt{12}}{k\pi}\sin(2\pi k t),$$

что совпадает с разложением функции $g^{(u)}(t)$ в ряд Фурье на $[0;\ 1]$.

Следствие доказано.

При отсутствии информации о распределении случайной величины наиболее логичным представляется использование либо статистики $J_n(g_{1/2})$ (в качестве предполагаемого значения T выбирается середина интервала), либо статистики $J_n(g^{(u)})$ (предполагается равномерное распределение случайной величины T). В силу теоремы 1.3, изучение этих статистик сводится к анализу частного случая функционалов от Z_n. Такой язык является вполне естественным. Так, статистики $J_n(g^{(u)})$ и $J_n(g_{1/2})$, введенные в следствиях 1.2 (при $T = 1/2$) и 1.3, приводят (с точностью до константы) к функционалам $|Z_n(1/2)|$ и $\left|\int_0^1 Z_n(t)dt\right|$ соответственно. Отметим, что последний функционал — это функционал $J_n^{|1|}$, введенный равенством (5) при $r = 1$.

В результате проведенного исследования (теорема 1.1, следствия 1.2, 1.3) нами отобраны следующие функционалы:

$J_n^{\infty} = \sup_{t\in[0;\ 1]} |Z_n(t)|$ — супремум модуля эмпирического моста;

$J_n^{|r|} = \left|\int_0^1 (Z_n(t))^r dt\right|^{1/r},\ r = 1,\ 2,\ \ldots,$ — интегральные функционалы, совпадающие при четных r с L_r-нормами, а при нечетных оказывающиеся более предпочтительными в силу теоремы 1.1 (при $r = 1$ функционал совпадает с наилучшим в предположении о равномерности распределения случайной величины T — результат следствия 1.3 и теоремы 1.3);

$J_n(g_{1/2}) = |Z_n(1/2)|$ — оптимальная статистика для случая, когда разладка происходит в середине интервала.

Для этих функционалов найдем приближенные бахадуровские наклоны, определенные формулой (2). В формулировке теоремы использованы обозначения, введенные в лемме 1.3.

Теорема 1.5 *Приближенные бахадуровские наклоны для функционалов* J_n^∞, $J_n^{|r|}$, $J_n(g_{1/2})$ *равны соответственно*

$$c_\infty(\theta) = 4{M_\theta}^2, \qquad c_r(\theta) = \frac{2^{2-2/r} B^2\left(\frac{1}{r},\ \frac{1}{2}\right)}{r(r+2)^{1-2/r}(r+1)^{2/r}} {M_\theta}^2,$$

$$c_{g_{1/2}}(\theta) = \left(\min\left\{\frac{1}{T};\ \frac{1}{1-T}\right\}\right)^2 {M_\theta}^2.$$

Здесь $B(\cdot,\cdot)$ *— бета-функция.*

Доказательство.

Согласно лемме 1.1, $c(\theta) = a(j(\theta))^2$ п.н.

Для статистики J_n^∞ получаем

$$a_\infty = -\lim_{x\to\infty} \frac{2}{x^2} \ln \mathbf{P}\{\sup_{t\in[0;\ 1]} |W^0(t)| \geq x\} = -\lim_{x\to\infty} \frac{2}{x^2} \ln(2e^{-2x^2}) = 4,$$

так как $\sup_{t\in[0;\ 1]} |W^0(t)|$ имеет распределение Колмогорова.

Согласно лемме 1.3,

$$j_\infty(\theta) = \sup_{t\in[0;\ 1]} z_\theta(t) = M_\theta.$$

Отсюда $c_\infty(\theta) = 4{M_\theta}^2$.

Для статистики $J_n^{|r|}$ коэффициенты a_r найдены в [131] (см. также [72], §2.6):

$$a_r = \frac{2^{2-2/r} B^2\left(\frac{1}{r},\ \frac{1}{2}\right)}{r(r+2)^{1-2/r}}.$$

Вычислим $j_r(\theta)$:

$$j_r(\theta) = \left(\int_0^1 (z_\theta(t))^r\right)^{1/r} dt$$

$$= \left(\int_0^T \left(\frac{M_\theta}{T}\right)^r t^r dt + \int_0^{1-T} \left(\frac{M_\theta}{1-T}\right)^r t^r dt\right)^{1/r} = \frac{M_\theta}{(r+1)^{1/r}}.$$

Подставляя в выражение для $c_r(\theta)$, получаем соответствующее утверждение теоремы.

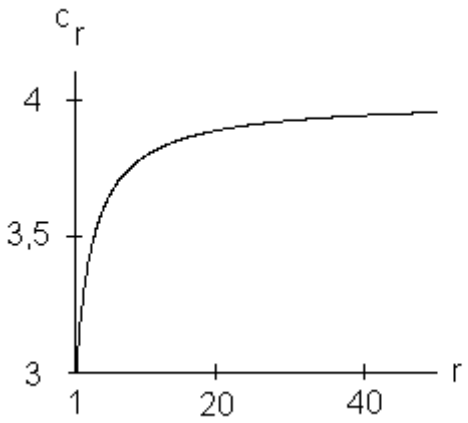

Рис. 1.2. График $c_r(\theta)$ как функции переменной r (при $M_\theta = 1$). Отметим, что $c_1(\theta) = 3{M_\theta}^2$.

Вычислим $a_{g_{1/2}}$. Так как $\mathbf{D}W^0(1/2) = 1/4$, то

$$a_{g_{1/2}} = -\lim_{x\to\infty} \frac{2}{x^2} \ln \mathbf{P}\{|W^0(1/2)| \geq x\} = 4.$$

Согласно лемме 3.2,

$$j_{g_{1/2}}(\theta) = |z_\theta(1/2)| = \min\left\{\frac{M_\theta}{2T};\ \frac{M_\theta}{2(1-T)}\right\}.$$

Подставляя в выражение для $c_{g_{1/2}}(\theta)$, получаем последнее утверждение теоремы.

Теорема доказана.

Отметим, что наилучшим является критерий, использующий L_∞-норму. Действительно,

$$\min\left\{\frac{1}{T};\ \frac{1}{1-T}\right\} \leq 2,$$

а график $c_r(\theta)$ как функции переменной r (при $M_\theta = 1$) изображен на рис. 1.2 — функция строго возрастает.

Подходы к практическому применению найденного наилучшего критерия (основанного на L_∞-норме эмпирического моста) к анализу однородности временных рядов изложены в соответствующих главах. Там же приведены результаты анализа конкретных временных рядов на однородность.

1.6 Результаты главы 1

1. Введен класс статистических критериев апостериорного обнаружения разладки в модели выборки. Эти критерии используют размах и различные нормы эмпирического моста, а также интегралы от положительных целых степеней эмпирического моста.

2. Доказано, что использование размаха и L_r-нормы эмпирического моста не имеет преимуществ перед использованием L_∞-нормы (супремум-нормы) и интеграла от эмпирического моста, возведенного в степень r, соотвтетственно.

3. Введен другой класс статистических критериев, основанных на статистике взвешенных сумм с коэффициентами, задаваемыми регулярным образом. Найдены необходимые и достаточные условия слабой сходимсоти распределения этой статистики к распределению модуля стандартного нормального закона.

4. Доказано, что статистика взвешенных сумм представима в виде интеграла Стилтьеса от эмпирического моста с остатком, сходящимся с вероятностью единица к нулю при выполнении как основной, так и альтернативной гипотезы.

5. В условиях байесовской постановки задачи (известного распределения момента разладки) найден наилучший критерий, основанный на взвешенных суммах. Доказано, что в случае равно-

мерного распределения момента разладки этот критерий приводит к интегралу от эмпирического моста.

6. Вычислены приближенные бахадуровские наклоны для критериев, основанных на супремум-норме, интегралов от положительных целых степеней эмпирического моста, а также наилучший критерий для случая, когда разладка происходит в середине интервала. Доказано, что наилучшим в этом широком классе критериев является критерий, основанный на L_∞-норме эмпирического моста.

ГЛАВА 2
Оценивание параметра Херста в модели фрактального гауссовского шума

2.1 Вводные замечания

Нередко встречаются ситуации, когда последовательность независимых одинаково распределенных случайных величин не годится в качестве модели приращений наблюдаемого процесса в дискретном времени. Несоответствие этой модели исследуемым данным может быть выявлено методами предыдущей главы.

Для временных рядов, встречающихся, в частности, в экономических исследованиях, характерна долговременная зависимость элементов. Стандартным средством моделирования этой зависимости служит фрактальное броуновское движение — модель, введенная почти одновременно А. Н. Колмогоровым и Винером. Название «фрактальное броуновское движение» восходит к статье Мандельброта и Ван Несса [144]. Отметим, что последовательность приращений фрактального броуновского движения в дискретном времени может быть задана как скользящее среднее от независимых одинаково распределенных гауссовских величин, взятых с коэффициентами, убывающими по степенному закону, регулирующему характер и силу зависимости.

По определению, фрактальное броуновкое движение с параметром Херста $H,\ 0 < H \leq 1$ — это гауссовский процесс $X_H(t)$, $t \geq 0$, с нулевым математическим ожиданием и корреляционной функцией

$$\mathbf{E}X_H(t)X_H(s) = \frac{\sigma^2}{2}\left(t^{2H} + s^{2H} - |t-s|^{2H}\right). \tag{7}$$

В частности, $X_H(0)$ имеет вырожденное распределение.

Из формулы (7) сразу же следует, что

$$\mathbf{D}X_H(t) = \mathbf{E}{X_H}^2(t) = \sigma^2 t^{2H}, \tag{8}$$

то есть дисперсия такого процесса возрастает в соответствии со степенным законом. Случай $H = 1/2$ соответствует обычному винеровскому процессу (броуновскому движению).

Стандартное фрактальное броуновское движение $B_H(t)$ характеризуется условиями $B_H(0) = 0,\ \ \sigma = 1$. Произвольное фрактальное броуновское движение может быть задано через стандартное в соответствии с формулой

$$X_H(t) = X_H(0) + \sigma B_H(t). \tag{9}$$

Существуют интегральные представления стандартного фрактального броуновского движения. Классическим является выведенное в статье Мандельброта и Ван Несса [144] представление

$$B_H(t) = L_H\Bigg(\int_{-\infty}^{0} \left((t-y)^{H-1/2} - |y|^{H-1/2}\right) dW(y) + \\ + \int_0^t (t-y)^{H-1/2} dW(y)\Bigg),$$

где

$$L_H = \left(\frac{1}{2H} + \int_0^\infty \left((1+s)^{H-1/2} - s^{H-1/2}\right)\right)^{-1/2}.$$

Здесь $W(t)$ — двусторонний стандартный винеровский процесс, который может быть задан через два независимых стандартных винеровских процесса (один для положительных, а другой для отрицательных значений аргумента).

Константа L_H может быть выражена через гамма-функцию Эйлера (см. работу Норроса и др. [151]):

$$L_H = \sqrt{\frac{2H\Gamma(3/2-H)}{\Gamma(H+1/2)\Gamma(2-2H)}}.$$

В статье [151] также можно найти другое интегральное представление фрактального броуновского движения, не использующее интегрирование по бесконечному интервалу.

Фрактальное броуновское движение является автомодельным процессом в смысле следующего определения. Автомодельными называют процессы $Y(t)$, обладающие свойством масштабной инвариантности: существует $\alpha > 0$ такое, что для любого $\lambda > 0$ выполнено равенство

$$Y(\lambda t) \stackrel{d}{=} \lambda^{1/\alpha} Y(t).$$

Здесь $\stackrel{d}{=}$ означает совпадение распределений.

Последовательность приращений $X_1, \ldots, X_n$ фрактального броуновского движения на единичных интервалах времени называется фрактальным гауссовским шумом:

$$X_i = B_H(i) - B_H(i-1).$$

Отметим, что для стационарной (в широком смысле) последовательности ([53], с.272) при $j > 0$

$$\mathbf{E} X_i X_{i+j} = \frac{1}{2}(\mathbf{D} S_{j+1} + \mathbf{D} S_{j-1} - 2\mathbf{D} S_j).$$

Здесь $S_n = X_1 + \ldots + X_n$, $S_0 = 0$.

Если стационарная гауссовская последовательность $X_1, \ldots, X_n$ — фрактальный гауссовский шум, то есть $S_n = B_H(n)$,

$$\mathbf{D} S_n = \sigma^2 n^{2H},$$

где H — показатель Херста ($0 < H \leq 1$), то

$$\mathbf{cov}(X_i;\ X_{i+j}) = \mathbf{E} X_i X_{i+j} = \frac{\sigma^2}{2}\left(|j+1|^{2H} + |j-1|^{2H} - 2|j|^{2H}\right), \tag{10}$$

и корреляционная функция фрактального гауссовского шума имеет вид

$$r(j) = \frac{\mathbf{cov}(X_i;\ X_{i+j})}{\mathbf{D}X_i} = \frac{1}{2}\left(|j+1|^{2H} + |j-1|^{2H} - 2|j|^{2H}\right). \qquad (11)$$

Задаче проверки гипотез относительно фрактального броуновского движения предшествуют задачи оценивания его параметров. В параграфе 2.2 рассмотрены существующие методы оценивания параметра Херста. В параграфе 2.3 разработаны знаковые методы оценивания параметра Херста, то есть методы, основанные на связи знаков приращений и их корреляций. В параграфе 2.4 рассматриваются и сравниваются различные модификации знакового метода. В результате проведенного анализа выделен бинарный знаковый метод, имеющий наименьшую дисперсию оценки среди тех из рассмотренных методов, число операций которых линейно зависит от объема выборки. В параграфе 2.5 суммированы результаты главы.

2.2 Методы оценивания параметра

Оцениванию параметров фрактального броуновского движения посвящена обширная литература, обзор которой приведен во введении.

Оценка максимального правдоподобия строится следующим образом.

Необходимо максимизировать плотность совместного распределения $f(\mathbf{X})$ фрактального гауссовского шума $X_1, \ldots, X_n$, имеющую вид

$$f(\mathbf{X}) = \frac{1}{(2\pi)^{n/2}\sigma^n\sqrt{\det R}} \exp(-\frac{1}{2\sigma^2}\mathbf{X}^T R^{-1}\mathbf{X}), \qquad (12)$$

где $\mathbf{X} = (X_1, \ldots, X_n)$.

Здесь

$$R = (r(i-j))_{i,j=1}^{n}$$

— корреляционная матрица, элементы которой — это значения корреляционной функции фрактального гауссовского шума

$$r(k) = \mathbf{corr}(X_i;\ X_{i+k}) = \frac{1}{2}\left(|k+1|^{2H} + |k-1|^{2H} - 2|k|^{2H}\right).$$

Максимизацию нужно проводить по совокупности параметров $H,\ \sigma$.

В работе [130] проведено исследование методов оценивания параметра H фрактального броуновского движения и показано, что другие оценки (методом нормированного размаха, методом дисперсии) имеют в 2–3 раза большую дисперсию по сравнению с оценкой максимального правдоподобия.

В качестве логарифмической функции правдоподобия здесь выступает логарифм совместной плотности распределения (12), взятой в точке $\mathbf{X}$:

$$L(\mathbf{X};\ H,\ \sigma) = -\ln(2\pi)^{n/2} - n\ln\sigma - \frac{1}{2}\ln\det R - \frac{1}{2\sigma^2}\mathbf{X}^T R^{-1}\mathbf{X}.$$

Из условия $\partial L/\partial\sigma = 0$ находим оценку максимального правдоподобия

$$\widehat{\sigma^2} = \frac{1}{n}\mathbf{X}^T R^{-1}\mathbf{X}$$

и, подставляя ее в предыдущее равенство, получаем

$$L(\mathbf{X};\ H) = -\ln(2\pi)^{n/2} - \frac{n}{2}\ln\mathbf{X}^T R^{-1}\mathbf{X} - \frac{1}{2}\ln\det R - \frac{n}{2}.$$

Итак, нам достаточно найти минимум функции

$$l(\mathbf{X};\ H) = n\ln\mathbf{X}^T R^{-1}\mathbf{X} + \ln\det R,$$

то есть оценка максимального правдоподобия параметра H равна

$$\widehat{H} = \arg\max_{0<H<1} l(\mathbf{X};\ H).$$

Эту оценку отыскиваем методом золотого сечения (см., например, [87], параграф 5.5). В качестве значений на первом шаге выбираем

точки $u_-^{(1)} = (3 - \sqrt{5})/2$, $u_+^{(1)} = (\sqrt{5} - 1)/2$, образующие золотые сечения интервала $(0;\ 1)$. Значения на втором шаге — золотые сечения интервала, выбираемого в зависимости от знака разности $l(\mathbf{X};\ u_-^{(1)}) - l(\mathbf{X};\ u_+^{(1)})$.

При поиске оценки параметра H обращение теплицевой симметричной матрицы R реализовывалось по экономичному алгоритму, предложенному в [20], с. 142–160, в ходе реализации которого одновременно вычисляется и детерминант.

Эту процедуру удается реализовать на персональном компьютере для выборок объемом до 1024 значений.

Для выборок большого объема предлагаются другие, менее точные оценки.

Первым стремлением при анализе долговременной зависимости множества данных может быть желание вычислить, насколько быстро автокорреляционная функция $r_H(k)$ убывает при больших лагах k. Однако это оказывается неудачным подходом, так как антиперсистентные данные трудно отличить от некоррелированных данных. Корреляции могут быть ошибочными из-за флуктуаций шума вокруг нуля. Метод надежного оценивания коэффициента Херста для временных рядов должен быть устойчивым методом измерения долговременных корреляций.

Метод нормированного размаха

Старейший и до сих пор наиболее используемый метод оценивания параметра H принадлежит Херсту, который заметил, что размах $R_L = \max_{k \le L} S_k - \min_{k \le L} S_k$ уровня (или накопленного стока) воды в водохранилище за промежуток времени L относится к выборочному стандартному отклонению $\sigma_L^* = \sqrt{\overline{X_L{}^2} - \left(\overline{X_L}\right)^2}$ стока за тот же период как величина, пропорциональная L^H, где H должно равняться $1/2$ для некоррелированных данных. Метод Херста

масштабированного размаха состоит в выборе неперекрывающихся подмножеств длины L из временного ряда и вычислении среднего значения статистики R_L/σ_L^*. При изображении данных на графике с логарифмическим масштабом по каждой из координат показатель Херста — это коэффициент угла наклона прямой линии $\ln(R_L/\sigma_L^*) = H \ln L + C$. К сожалению, несмотря на его широкое использование, анализ масштабированного размаха по Херсту является плохой оценкой для H с систематическим уклонением в сторону $H = 0,7$ и логарифмической скоростью сходимости [53].

Мы будем пользоваться следующим алгоритмом оценивания методом нормированного размаха [113].

Алгоритм 2.1 *Для набора данных* $X_1, \ldots, X_n$*:*

1. *Центрировать данные:* $X_i^* = X_i - \overline{X}$.
2. *Для длины блока* $L = 1, 2, 4, \ldots, 2^{m-2}$ $(L \leq n/4)$ *осуществить масштабирование данных:*

$$X_j^{(L)} = \sum_{i=L(j-1)+1}^{Lj} X_i^*, \quad j = 1, 2, 3, \ldots, 2^m/L.$$

3. *Вычислить размах сумм и стандартное отклонение масштабированных данных*

$$R_L = \max_{k \leq 2^m/L} \sum_{j=1}^{k} X_j^{(L)} - \min_{k \leq 2^m/L} \sum_{j=1}^{k} X_j^{(L)};$$

$$\sigma_L^* = \sqrt{\frac{L}{2^m} \sum_{j=1}^{L} (X_j^{(L)})^2 - \left(\frac{L}{2^m} \sum_{j=1}^{L} X_j^{(L)} \right)^2};$$

4. *Вычислить оценку* $\tilde{H}_1$ *параметра* H *как коэффициент наклона линии линейной регрессии* $\ln R_L - \ln \sigma_L^*$ *на* L *по всем*

$L = 1, 2, 4, \ldots, 2^{m-2}$ *($L \leq n/4$):*

$$\tilde{H}_1 = \frac{\overline{(\ln R - \ln \sigma^*)L} - \overline{\ln R - \ln \sigma^*} \cdot \overline{L}}{\overline{L^2} - (\overline{L})^2}.$$

Метод дисперсии

В литературе предлагается несколько альтернатив методу масштабированного размаха, включая автокорреляционный анализ, фурье-анализ и оценки максимального правдоподобия. Преимущество последних в том, что это не графические методы, а численные — они просто отвечают наилучшей оценке показателя Херста напрямую. К сожалению, они требуют (по крайней мере) предположения о форме долговременной зависимости (такого, например, как у фрактального броуновского движения) и дают неадекватные оценки, если это предположение некорректно. Каждый из вышеперечисленных методов дает смещенные и медленно сходящиеся оценки (большой объем данных требуется для того, чтобы снизить смещение).

Однако существует метод, который оказался существенно лучше, требуя меньших объемов данных и давая меньшие смещения оценки. Это метод дисперсии. Он допускает графическую интерпретацию — воспроизводит соотношение степенного закона, из которого показатель H может быть найден как наклон прямой при использовании логарифмического масштаба по каждой из координат.

Метод дисперсии основан на усреднении данных по ячейкам длины L и вычислении выборочной дисперсии усредненного набора данных. Алгоритм основан на том, что дисперсия должна расти пропорционально L^{2H}, следовательно, коэффициент H и в этом алгоритме может быть вычислен через коэффициент угла наклона прямой линейной регрессии $\ln \sigma_L^* = H \ln L + C$.

Алгоритм 2.2 *Для набора данных $X_1, \ldots, X_n$:*

1. *Для длины блока* $L = 1, 2, 4, \ldots, 2^{m-2}$ $(L \leq n/4)$ *осуществить масштабирование данных:*

$$X_j^{(L)} = \sum_{i=L(j-1)+1}^{Lj} X_i, \quad j = 1, 2, 3, \ldots, 2^m/L.$$

2. *Вычислить стандартное отклонение масштабированных данных*

$$\sigma_L^* = \sqrt{\frac{L}{2^m}\sum_{j=1}^{L}(X_j^{(L)})^2 - \left(\frac{L}{2^m}\sum_{j=1}^{L} X_j^{(L)}\right)^2}.$$

3. *Вычислить оценку* $\tilde{H}_2$ *параметра* H *как коэффициент наклона линии линейной регрессии* $\ln \sigma_L^*$ *на* $\ln L$ *по всем* $L = 1, 2, 4, \ldots, 2^{m-2}$ $(L \leq n/4)$*:*

$$\tilde{H}_2 = \frac{\overline{\ln \sigma^* \cdot \ln L} - \overline{\ln \sigma^*} \cdot \overline{\ln L}}{\overline{(\ln L)^2} - (\overline{\ln L})^2}.$$

Разработаем теперь метод, позволяющий по знакам элементов последовательности, образующей фрактальный гауссовский шум, получать оценку параметра Херста.

2.3 Элементарный знаковый метод и центрированный знаковый метод

Если известно, что двумерный нормальный вектор имеет нулевой вектор математического ожидания (является центрированным), то существует взаимно однозначное соответствие между коэффициентом корреляции его компонент и вероятностью того, что эти компоненты имеют разный знак. Это соответствие дается следующей леммой, которую можно найти в [103], с. 236.

Лемма 2.1 *Если* (ξ, η) *— двумерная нормальная случайная величина с нулевым вектором математического ожидания, то*

$$P\{\xi\eta < 0\} = \frac{1}{\pi} \arccos \ corr\,(\xi, \eta).$$

Лемма 2.1 служит основой для построения знаковых процедур оценивания параметра Херста, различные модификации которых составляют содержание следующих двух параграфов.

Применяя лемму 2.1 к фрактальному гауссовскому шуму $X_1, \ldots, X_n$, получаем формулу

$$r(1) = \cos(\pi \mathbf{P}\{X_1 X_2 < 0\}). \tag{13}$$

Здесь $r(1)$ — значение корреляционной функции фрактального гауссовского шума в точке 1. Подстановкой в формулу (11) получаем

$$r(1) = 2^{2H-1} - 1. \tag{14}$$

Из формул (14) и (13) получаем

$$2^{2H-1} - 1 = \cos(\pi \mathbf{P}\{X_1 X_2 < 0\}). \tag{15}$$

Подставим в формуле (15) частоту вместо вероятности перемены знака. В левой части получится оценка $\tilde{H}$ параметра H по частоте перемены знака:

$$2^{2\tilde{H}-1} - 1 = \cos(\pi\nu). \tag{16}$$

Здесь ν — частота перемены знака, подсчитанная по выборке:

$$\nu = \frac{1}{n-1} \sum_{i=1}^{n-1} \mathbf{I}\{X_i X_{i+1} < 0\}.$$

Эта оценка может быть выражена в явном виде через частоту перемены знака.

На этой формуле основан метод получения оценки параметра H методом знаков:

$$\tilde{H} = \frac{1}{2} + \frac{1}{2} \log_2 \left(1 + \cos(\pi\nu)\right), \tag{17}$$

где ν — частота перемены знака в последовательности приращений фрактального броуновского движения.

Введем случайные величины $\{J_i\}_{i=1}^{n-1}$, где $J_i = I\{X_i X_{i+1} < 0\}$ — индикатор того, что соседние X_i разного знака.

Согласно введенным нами обозначениям, $\nu = \frac{1}{n-1}\sum_{i=1}^{n-1} J_i$.

Докажем, что для всех $H \in (0;\ 1)$ оценка (17) будет состоятельной.

Доказательству предпошлем лемму, позволяющую применять результаты для последовательностей с сильным перемешиванием к оценкам такого рода.

Лемма 2.2 *Если* $\{X_i\},\ i = 1,\ 2,\ \ldots$ — *стационарная гауссовская последовательность с корреляционной функцией* $r(n)$, *обладающей свойством* $r(n) \to 0$ *при* $n \to \infty$, *а* $g :\ \mathbf{R}^k \to \mathbf{R}$ — *произвольная борелевская функция* k *переменных, то* $\{g(X_i,\ \ldots,\ X_{i+k-1})\}$, $i = 1,\ 2,\ \ldots$ — *стационарная последовательность с сильным перемешиванием.*

Доказательство.

Стационарность последовательности $\{g(X_i,\ \ldots,\ X_{i+k-1})\}$ сразу же следует из определения стационарности.

Согласно [42], последовательность $\{X_i\}$ обладает свойством сильного перемешивания, то есть для любого фиксированного числа m и для любых $\mathcal{A} \in \sigma(X_1,\ \ldots,\ X_m),\ \mathcal{B} \in \sigma(X_{n+1},\ X_{n+2},\ \ldots)$ выполнено свойство

$$\mathbf{P}(\mathcal{AB}) - \mathbf{P}(\mathcal{A})\mathbf{P}(\mathcal{B}) \to 0$$

при $n \to \infty$. Но это свойство выполняется и для последовательности $\{g(X_i,\ \ldots,\ X_{i+k-1})\},\ i = 1,\ 2,\ \ldots$ в силу вложения сигма-алгебр

$$\sigma(g(X_1,\ \ldots,\ X_{k-1}),\ \ldots,\ g(X_m,\ \ldots,\ X_{m+k-1})\} \subseteq \sigma(X_1,\ \ldots,\ X_{m+k-1}),$$

$$\sigma(g(X_{n+1},\ \ldots,\ X_{n+k}),\ \sigma(g(X_{n+2},\ \ldots,\ X_{n+k+1}),\ldots\} \subseteq$$
$$\subseteq \sigma(X_{n+1},\ X_{n+2},\ \ldots).$$

Лемма доказана.

Теорема 2.1 *Оценка (17) параметра $H \in (0;\ 1)$ сильно состоятельна.*

Доказательство. Последовательность $X_1,\ X_2,\ \ldots$ удовлетворяет условиям сильного перемешивания [42] в силу гауссовости и стремления корреляционной функции к нулю.

Согласно предыдущей лемме, последовательность $\{J_i\}$ удовлетворяет условиям сильного перемешивания, поэтому для нее имеет место усиленный закон больших чисел: $\nu \to p$ п. н. [42], где p — вероятность перемены знака фрактальным гауссовским шумом. В силу непрерывности функции, выражающей H через p, оценка (17) параметра $H \in (0;\ 1)$ сходится п. н. к истинному значению параметра, то есть является сильно состоятельной.

Теорема доказана.

Сформулируем следующий алгоритм получения оценки элементарным знаковым методом.

Алгоритм 2.3 *Для набора данных $X_1, \ldots, X_n$:*

1. *центрировать данные $X_i^* = X_i - \overline{X}$;*
2. *по центрированным данным вычислить частоту перемены знака $\nu_n^* = \frac{1}{n-1}\sum_{i=1}^{n-1} J_i$, где $J_i = I\{X_i^* X_{i+1}^* < 0\}$;*
3. *вычислить оценку по методу знаков:*

$$\tilde{H}_3 = \frac{1}{2} + \frac{1}{2}\log_2\left(1 + \cos(\pi\nu_n^*)\right).$$

В качестве более точного метода оценивания, учитывающего корреляции, возникающие при центрировании исходных данных, нами предложен следующий *центрированный знаковый метод*, использующий конструкцию *фрактального броуновского моста*.

Наряду с фрактальным броуновским движением введем понятие *фрактального броуновского моста* B_H^0:

$$B_H^0(t) = B_H(t) - tB_H(1).$$

Он обладает тем свойством, что $B_H^0(1) = B_H^0(0) = 0$.

Здесь полная аналогия с обычным броуновским мостом. В частности, при $H = 1/2$ фрактальный броуновский мост является обычным броуновским мостом. Однако при $H \neq 1/2$ фрактальный мост не обладает симметрией по пространству: распределения $B_H^0(t)$ и $B_H^0(1-t)$ не совпадают при $t \in (0;\ 1/2)$.

Будем обозначать $X_i^* = X_i - \overline{X}$, $S_k^* = X_1^* + \ldots + X_k^*$. Тогда

$$B_H^0(k/n) = S_k^*/n^H.$$

Коэффициент корреляции между X_k^* и X_{k+1}^* обозначим через $\rho_1^*(k)$. Прямыми вычислениями получаем

$$\rho_1^*(k) = \frac{c_1^*(k)}{\sigma_1^*(k)\sigma_1^*(k+1)},$$

где

$$c_1^*(k) = \frac{2^{2H}-2}{2n^{2H}} - \frac{1}{2n}\left(\left(\frac{k+1}{n}\right)^{2H} + \left(\frac{k-1}{n}\right)^{2H} - 2\left(\frac{k}{n}\right)^{2H}\right) +$$
$$+\frac{1}{2n}\left(\left(\frac{n-k-1}{n}\right)^{2H} + \left(\frac{n-k+1}{n}\right)^{2H} - 2\left(\frac{n-k}{n}\right)^{2H}\right) + \frac{1}{n^2},$$

$$\sigma_1^*(k) = \sqrt{\frac{1}{n^{2H}} - \frac{k^{2H} - (k-1)^{2H} - (n-k)^{2H} + (n-k+1)^{2H}}{n^{2H+1}} + \frac{1}{n^2}}.$$

Соответственно,

$$\sigma_1^*(k{+}1) = \sqrt{\frac{1}{n^{2H}} - \frac{(k+1)^{2H} - k^{2H} - (n-k-1)^{2H} + (n-k)^{2H}}{n^{2H+1}} + \frac{1}{n^2}}.$$

Так как получение напрямую оценки для H из этого равенства — трудоемкая вычислительная процедура, получим аппроксимацию приведенного выше выражения. Пусть $n \to \infty$. Тогда равномерно по $k \leq n$

$$\left(\frac{k+1}{n}\right)^{2H} - \left(\frac{k}{n}\right)^{2H} = \left(\frac{k}{n}\right)^{2H-1} \cdot \frac{2H}{n} + o\left(\frac{1}{n}\right),$$

$$\left(\frac{n-k+1}{n}\right)^{2H} - \left(\frac{n-k}{n}\right)^{2H} = \left(\frac{n-k}{n}\right)^{2H-1} \cdot \frac{2H}{n} + o\left(\frac{1}{n}\right).$$

Отсюда

$$\rho_1^*(k) = \frac{\dfrac{2^{2H}-2}{2n^{2H}} + \dfrac{1}{n^2} + o\left(\dfrac{1}{n^2}\right)}{\dfrac{1}{n^{2H}} - \dfrac{2H}{n^2}\left(\left(\dfrac{k}{n}\right)^{2H-1} + \left(\dfrac{n-k}{n}\right)^{2H-1}\right) + \dfrac{1}{n^2} + o\left(\dfrac{1}{n^2}\right)}.$$

Умножив на n^{2H}, получаем:

$$\rho_1^*(k) = \frac{\dfrac{2^{2H}-2}{2} + n^{2H-2} + o\left(n^{2H-2}\right)}{1 - n^{2H-2}\left(2H\left(\dfrac{k}{n}\right)^{2H-1} + 2H\left(\dfrac{n-k}{n}\right)^{2H-1} - 1\right) + o\left(n^{2H-2}\right)} =$$

$$= \left(\frac{2^{2H}-2}{2} + n^{2H-2}\right) \times$$

$$\times \left(1 + n^{2H-2}\left(2H\left(\frac{k}{n}\right)^{2H-1} + 2H\left(\frac{n-k}{n}\right)^{2H-1} - 1\right) + o(1)\right).$$

Обозначим $\tilde{p}_1^*$ — частоту перемены знака последовательностью $X_1^*, \ldots, X_n^*$. Представляя арккосинус с помощью дифференциала, получаем

$$\mathbf{E}\tilde{p}_1^* = \frac{1}{\pi(n-1)} \sum_{k=1}^{n-1} \arccos \rho_1^*(k) =$$

$$= \frac{1}{\pi} \left(\arccos \frac{2^{2H}-2}{2} - \frac{\sum_{k=1}^{n-1} d_k}{(n-1)\sqrt{1-\left(\frac{2^{2H}-2}{2}\right)^2}} + o\left(\frac{1}{n-1} \sum_{k=1}^{n-1} d_k \right) \right),$$

где

$$d_k = \rho_1^*(k) - \frac{2^{2H}-2}{2} =$$

$$= n^{2H-2} \left(1 + \frac{2^{2H}-2}{2} \left(2H \left(\left(\frac{k}{n}\right)^{2H-1} + \left(\frac{n-k}{n}\right)^{2H-1} \right) - 1 \right) + o(1) \right).$$

Переходя от суммы к интегралу, получаем

$$\mathbf{E}\tilde{p}_1^* = \frac{1}{\pi} \arccos \frac{2^{2H}-2}{2} - \frac{n^{2H-2}}{\pi\sqrt{1-\left(\frac{2^{2H}-2}{2}\right)^2}} \times$$

$$\times \left(1 + \frac{2^{2H}-2}{2} \left(2H \int_0^1 \left(x^{2H-1} + (1-x)^{2H-1} \right) dx - 1 \right) \right) + o\left(n^{2H-2}\right) =$$

$$= \frac{1}{\pi} \left(\arccos\left(2^{2H-1}-1\right) - \frac{2^{H-1}}{\sqrt{1-2^{2H-2}}} n^{2H-2} \right) + o\left(n^{2H-2}\right).$$

Полученный результат позволяет отыскивать оценку $\tilde{H}$ параметра H как решение уравнения

$$\tilde{p}_1^* = \frac{1}{\pi} \left(\arccos\left(2^{2\tilde{H}-1}-1\right) - \frac{2^{\tilde{H}-1}}{\sqrt{1-2^{2\tilde{H}-2}}} n^{2\tilde{H}-2} \right).$$

В силу монотонности арккосинуса это уравнение при достаточно больших n имеет единственный корень на отрезке $[0;\ 1]$ и может быть решено методом дихотомии.

Изложенный центрированный метод знаков полезен в силу того обстоятельства, что он позволяет работать с данными, имеющими произвольное математическое ожидание. Он позволяет в значительной мере компенсировать систематическую погрешность, возникающую при центрировании данных.

Алгоритм 2.4 *Для набора данных* $X_1, \ldots, X_n$*:*

1. *центрировать данные* $X_i^* = X_i - \overline{X}$*;*
2. *по центрированным данным вычислить частоту перемены знака* $\nu_n^* = \frac{1}{n-1}\sum_{i=1}^{n-1} J_i$*, где* $J_i = I\{X_i^* X_{i+1}^* < 0\}$*;*
3. *вычислить оценку* $\tilde{H}_4$ *параметра* H *по центрированному методу знаков как решение методом дихотомии уравнения*

$$\tilde{p}_1^* = \frac{1}{\pi}\left(\arccos\left(2^{2\tilde{H}_4-1}-1\right) - \frac{2^{\tilde{H}_4-1}}{\sqrt{1-2^{2\tilde{H}_4-2}}} n^{2\tilde{H}_4-2}\right).$$

Две наиболее употребительные модели для временных рядов, возникающих в финансовой математике — обычное броуновское движение (винеровский процесс) и его обобщение — фрактальное броуновское движение. Естественным образом возникает вопрос, какую из этих моделей следует предпочесть для каждого конкретного временного ряда. На этот вопрос можно ответить, построив критерий проверки основной гипотезы о том, что параметр Херста равен $1/2$ против альтернативной гипотезы о том, что параметр отличается от $1/2$. Рассмотрим критерий, основанный на анализе частоты перемены знака приращениями процесса.

В качестве основной гипотезы предлагается модель выборки, для которой $H = 1/2$. Альтернативная модель — модель фрактального гауссовского шума, для которой $H \neq 1/2$.

Обозначим через J_i индикаторы того, что X_i и X_{i+1} разного знака:

$$J_i = \mathbf{I}\{X_i X_{i+1} < 0\}.$$

Лемма 2.3 *В модели выборки случайные величины J_i имеют распределение Бернулли с параметром 1/2.*

Доказательство.

$P\{J_i = 1\} = P\{X_i > 0, X_{i+1} < 0\} + P\{X_i < 0, X_{i+1} > 0\} = \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2}$.

Лемма доказана.

Лемма 2.4 *В модели выборки J_i независимы.*

Доказательство.

Докажем попарную независимость J_i. Очевидно, что J_k и J_{k+s}, где $s > 1$, независимы.

Рассмотрим J_k и J_{k+1}:

$$P\{J_k = 0, J_{k+1} = 0\} = 2P\{X_k > 0, X_{k+1} > 0, X_{k+2} > 0\} = \\ = \frac{1}{4} = P\{J_k = 0\}P\{J_{k+1} = 0\}.$$

Далее,

$$P\{J_k = 0, J_{k+1} = 1\} = P\{J_k = 0\} - P\{J_k = 0, J_{k+1} = 0\} = \frac{1}{4} = \\ = P\{J_k = 0\}P\{J_{k+1} = 1\}.$$

Аналогично $P\{J_k = 1, J_{k+1} = 1\} = P\{J_k = 1, J_{k+1} = 0\} = \frac{1}{4}$.

Предположим, что для любых $i_1 < i_2 < \ldots < i_k$ случайные величины $J_{i_1}, J_{i_2}, \ldots, J_{i_k}$ независимы, т.е.

$$P\{J_{i_1} = a_1, J_{i_2} = a_2, \ldots, J_{i_k} = a_k\} = \prod_{m=1}^{k} P\{J_{i_m} = a_m\}, \text{где } a_m \in \{0, 1\}.$$

Рассмотрим произвольную комбинацию из $k + 1$ индикатора $J_{i_1}, J_{i_2}, \ldots, J_{i_k+1}$ для любых $i_1 < i_2 < \ldots < i_{k+1}$. Возможны два случая:

а) для некоторого l существует s такое, что выполнено $i_l < s$ и $i_{l+1} > s$, тогда

$$P\{J_{i_1} = a_1, \ldots, J_{i_l} = a_l, J_{i_{l+1}} = a_{l+1}, \ldots, J_{i_k+1} = a_{k+1}\} =$$
$$= P\{J_{i_1} = a_1, \ldots, J_{i_l} = a_l\} P\{J_{i_{l+1}} = a_{l+1}, \ldots, J_{i_k+1} = a_{k+1}\} =$$
$$= \prod_{m=1}^{l} P\{J_{i_m} = a_m\} \prod_{m=l+1}^{k+1} P\{J_{i_m} = a_m\} =$$
$$= \prod_{m=1}^{k+1} P\{J_{i_m} = a_m\},$$

где $a_m \in \{0;\ 1\}$;

б) для всех $l \in \{1, \ldots, k\}$ выполнено $i_{l+1} = i_l + 1$, тогда

$$P\{J_{i_1} = a_1, \ldots, J_{i_k+1} = a_{k+1}\} =$$
$$= P\{J_{i_1} = a_1\} P\{J_{i_2} = a_2 | J_{i_1} = a_1\} \cdot \ldots \cdot P\{J_{i_{k+1}} = a_{k+1} | J_{i_k} = a_k\} =$$
$$= P\{J_{i_1} = a_1\} P\{J_{i_2} = a_2\} \cdot \ldots \cdot P\{J_{i_{k+1}} = a_{k+1}\} = \prod_{m=1}^{k+1} P\{J_{i_m} = a_m\}.$$

Лемма доказана.

По ЦПТ получаем, что $P\{\frac{S_n - n/2}{\sqrt{n}/2} < t\} \to \Phi_{0,1}(t)$, где $S_n = \sum\limits_{i=1}^{n} J_i$.

Так как

$$\tilde{H}_3 = \frac{1}{2}\left(1 + \log_2(1 + \cos(\pi\nu))\right),$$

то $\tilde{H}_3$ — оценка по методу моментов. Она является асимптотически нормальной в силу теоремы об асимптотической нормальности [9], так как

$$h(p) = \frac{1}{2}\left(1 + \log_2(1 + \cos(\pi p))\right)$$

— непрерывно дифференцируемая функция при $p \in [0;\ 1]$. Коэффициент асимптотической нормальности равен

$$\sigma_{1/2} = |h'(\mathbf{E}\nu)|\sqrt{\mathbf{D}\nu} = \frac{\pi \sin(\pi p)}{\log_2(1 + \cos(\pi p))}\Big|_{p=1/2} \cdot \sqrt{1/4} = \frac{2\pi}{\ln 2}.$$

Итак, теоретическое значение стандартного отклонения

$$\sigma = \frac{\sigma_{1/2}}{\sqrt{N}} = \frac{2\pi}{\sqrt{N}\ln 2}.$$

Алгоритм 2.5 *(Элементарный знаковый критерий). По имеющимся данным* $X_1, \ldots, X_n$*:*

1. *Вычисляем среднее значение и вычитаем его из данных:* $X_i^* = X_i - \overline{X}$.

2. *Укрупняем данные для обеспечения нормальности: принимая* $\tau \geq 8$*, вычисляем* $X_1^\tau, \ldots, X_N^\tau$*, где* $N = [n/\tau]$*, по формуле*

$$X_i^\tau = \sum_{j=1}^{\tau} X^*_{(i-1)\tau+j}.$$

3. *Находим частоту перемены знака:*

$$\nu_N^1 = \frac{1}{N-1} \sum_{i=1}^{N-1} \mathbf{I}\{X_i^\tau X_{i+1}^\tau < 0\}.$$

4. *Находим достигаемый уровень значимости гипотезы о независимости:*

$$\varepsilon^* = 2\overline{\Phi}(2\sqrt{n}|\nu_N^1 - 1/2|).$$

   Здесь $\overline{\Phi}(x) = \frac{1}{\sqrt{2\pi}} \int_x^{+\infty} e^{-t^2/2} dt$.

5. *Гипотеза о некоррелированности принимается на уровне* ε*, если* $\varepsilon^* \geq \varepsilon$*; отвергается на уровне* ε*, если* $\varepsilon^* < \varepsilon$.

2.4 Модифицированный знаковый метод и бинарный знаковый метод

Оценки знаковым методом не получили широкого распространения ввиду их большой дисперсии. Из-за большой дисперсии оценки и критерий, предложеннный в предыдущем параграфе, имеет малую мощность. Рассмотрим две модификации оценки знаковым методом, существенно понижающие дисперсию оценки параметра H. Эти

оценки будут использоваться для построения критерия, различающего основную гипотезу об отсутствии зависимости и альтернативную гипотезу о том, что зависимость соответствует модели фрактального гауссовского шума.

Модификации оценок основаны на использовании свойства самоподобия фрактального броуновского движения.

Отметим, что для стационарной (в широком смысле) последовательности ([53], с.272) при $j > 0$

$$\mathbf{E}X_iX_{i+j} = \frac{1}{2}(\mathbf{D}S_{j+1} + \mathbf{D}S_{j-1} - 2\mathbf{D}S_j).$$

Здесь $S_n = X_1 + \ldots + X_n$, $S_0 = 0$.

Если стационарная гауссовская последовательность $X_1, \ldots, X_n$ — фрактальный гауссовский шум, то есть

$$\mathbf{D}S_n = \sigma^2 n^{2H},$$

где H — показатель Херста ($0 < H \leq 1$), то

$$\mathbf{cov}(X_i;\ X_{i+j}) = \mathbf{E}X_iX_{i+j} = \frac{\sigma^2}{2}\left(|j+1|^{2H} + |j-1|^{2H} - 2|j|^{2H}\right),$$

и корреляционная функция последовательности имеет вид

$$r(j) = \frac{\mathbf{cov}(X_i;\ X_{i+j})}{\mathbf{D}X_i} = \frac{1}{2}\left(|j+1|^{2H} + |j-1|^{2H} - 2|j|^{2H}\right). \qquad (18)$$

В частности, коэффициент корреляции между соседними случайными величинами равен $r(1) = 2^{2H-1} - 1$. Поэтому вероятность того, что последовательные случайные величины имеют противоположные знаки, равна

$$\mathbf{P}\{X_iX_{i+1} < 0\} = \frac{1}{\pi}\arccos(2^{2H-1} - 1). \qquad (19)$$

Но и последовательные суммы k случайных величин обладают тем же свойством. Обозначим через $S_{k,i}$ сумму k случайных величин,

начиная с номера $i+1$:

$$S_{k,i} = \sum_{j=i+1}^{i+k} X_j = S_{k+i} - S_i.$$

Тогда $\mathbf{D}\sum_{j=1}^{n} S_{k,k(j-1)} = \mathbf{D}S_{kn} = \sigma^2 k^{2H} n^{2H}$, и

$$\mathbf{E}S_{k,0}S_{k,kj} = \frac{\sigma^2 k^{2H}}{2}\left(|j+1|^{2H} + |j-1|^{2H} - 2|j|^{2H}\right).$$

Поэтому коэффициент корреляции и вероятность иметь противоположные знаки для $S_{k,0}$ и $S_{k,k}$ такие же, как для X_i и X_{i+1}. В силу стационарности то же верно для $S_{k,i}$ и $S_{k,k+i}$ при произвольном $i > 0$.

Таким образом, появляется возможность улучшить оценку параметра H, используя агрерирование случайной последовательности, то есть суммируя индикаторы перемены знака блоками равной длины, полученными суммированием соседних элементов последовательности.

В работе предлагается алгоритм построения критерия, основанного на статистике числа перемен знака соседними суммами элементов последовательности

$$V_n = V_n(K) = \sum_{k=1}^{K}\sum_{j=0}^{n-2k} \mathbf{I}\{S_{k,j} \cdot S_{k,j+k} < 0\}.$$

Здесь $K \le n/2$ — максимальное число слагаемых в рассмативаемых суммах элементов последовательности.

Введем необходимые обозначения, вычислим ковариации индикаторов перемены знака и докажем центральную предельную теорему при основной гипотезе. Коэффициент асимптотической нормальности будет вычислен в явном виде. Затем докажем общее утверждение относительно асимптотики ковариации индикаторов перемены знака элементами стационарной гауссовской последовательности с

нулевым математическим ожиданием и медленно сходящейся к нулю корреляционной функцией. Затем применим эти результаты к доказательству центральной предельной теоремы для статистики $V_n(K)$ при $H \neq 1/2$. Докажем непрерывность коэффициента асимптотической нормальности $\sigma_K(H)$ в точке $H = 1/2$. Затем проведем аналитическое сравнение построенных оценок при разных K с оценками по методу периодограммы.

Итак, пусть $X_1, \ldots, X_n$ — фрактальный гауссовский шум, то есть стационарная гауссовская последовательность с нулевым математическим ожиданием и корреляционной функцией, задаваемой формулой (11). Согласно основной гипотезе, $H = 1/2$, то есть случайные величины $X_1, \ldots, X_n$ независимы — образуют выборку из нормального распределения. Согласно альтернативной гипотезе, $H \neq 1/2$.

Напомним, что $S_{k,j}$ — сумма k случайных величин, начиная с номера $j + 1$. Обозначим

$$J_{k,j} = \mathbf{I}\{S_{k,j} \cdot S_{k,j+k} < 0\}$$

— индикатор того, что соседние суммы k случайных величин имеют разные знаки,

$$L_k = \sum_{j=0}^{n-2k} J_{k,j}$$

— число перемен знака соседними суммами k слагаемых,

$$p(\rho) = \frac{1}{\pi} \arccos(\rho)$$

— вероятность того, что нормальные случайные величины с нулевыми математическими ожиданиями и коэффициентом корреляции ρ имеют разные знаки (согласно лемме 2.1),

$$q(\rho) = 1 - p(\rho) = \frac{1}{\pi} \arccos(-\rho)$$

— вероятность противоположного события.

Мы будем рассматривать статистики

$$V_n = V_n(K) = \sum_{k=1}^{K} L_k.$$

Это общее число перемен знака соседними суммами k слагаемых, $1 \le k \le K$. Здесь K — наибольшее допустимое число слагаемых в сумме, $K \le n/2$.

В качестве оценки вероятности перемены знака предлагается частота перемены знака, подсчитанная на основании статистики V_n:

$$p_n^*(K) = \frac{V_n(K)}{\sum_{k=1}^{K}(n-2k+1)} = \frac{V_n(K)}{K(n-K)}.$$

Отметим, что при $H = 1/2$, то есть при выполнении нулевой гипотезы, V_n подчиняется центральной предельной теореме, если $\mathbf{D}V_n \to \infty$: разобьем V_n на блоки T_j индикаторов перемены знака суммами, первая из которых начинается с номера $Kj + i, \;\; i = 1, \ldots, K$. Тогда T_j — 1-зависимые одинаково распределенные ограниченные случайные величины, и их сумма подчиняется центральной предельной теореме, если дисперсия суммы стремится к бесконечности [42].

Для проверки последнего условия, а также для вычисления коэфициента асимптотической нормальности оценки $p_n^*(K)$ изучим асимптотику $\mathbf{D}V_n$ при нулевой гипотезе.

В следующем параграфе при нулевой гипотезе вычисляются ковариации $\mathbf{cov}\,(J_{k,j},\; J_{k',j'})$, знание которых необходимо для подсчета $\mathbf{D}V_n$.

Будем предполагать, что выполнена нулевая гипотеза $H = 1/2$, то есть исходные случайные величины независимы — образуют выборку из нормального распределения с нулевым математическим ожиданием.

Для последующих вычислений будет полезно следующее обозна-

чение: $G_{k,k',l}$ — ковариация индикаторов перемены знака парами блоков из k и k' слагаемых, середины которых смещены на l друг относительно друга.

Введя обозначение $j' = k + j + l - k'$, определим $G_{k,k',l}$ формально как

$$G_{k,k',l} = \mathbf{cov}\ (J_{k,j},\ J_{k',j'}) =$$
$$= \mathbf{P}\{S_{k,j} \cdot S_{k,j+k} < 0,\ S_{k',j'} \cdot S_{k',j'+k'} < 0\} -$$
$$-\mathbf{P}\{S_{k,j} \cdot S_{k,j+k} < 0\} \cdot \mathbf{P}\{S_{k',j'} \cdot S_{k',j'+k'} < 0\} =$$
$$= \mathbf{P}\{S_{k,j} \cdot S_{k,j+k} < 0,\ S_{k',j'} \cdot S_{k',j'+k'} < 0\} - 1/4.$$

С учетом симметрии

$$G_{k,k',l} = 2\mathbf{P}\{S_{k,j} < 0,\ S_{k,j+k} > 0,\ S_{k',j'} < 0,\ S_{k',j'+k'} > 0\} + \qquad (20)$$
$$+2\mathbf{P}\{S_{k,j} > 0,\ S_{k,j+k} < 0,\ S_{k',j'} < 0,\ S_{k',j'+k'} > 0\} - 1/4.$$

Рис. 2.1 поясняет введенное обозначение.

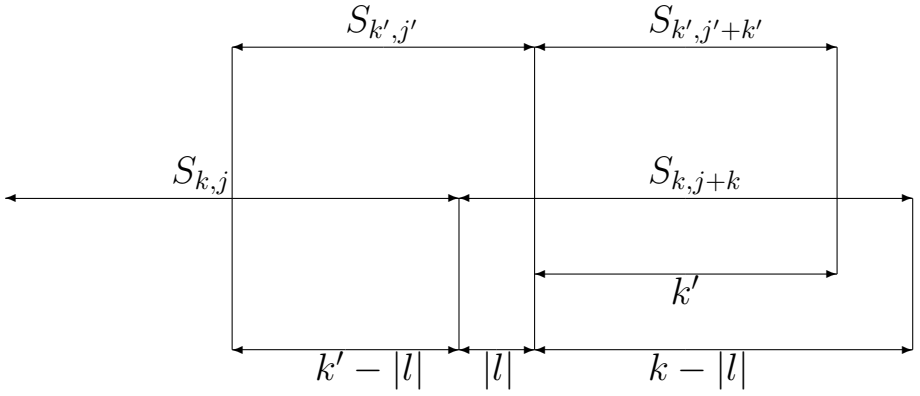

Рис. 2.1.

Отметим, что по определению $G_{k,k',l} = G_{k',k,l}$.

Вычислим коэффициенты $G_{k,k',l}$ в случае, когда $k' \leq k$.

Теорема 2.2 *Пусть* $0 < k' \le k$.

Если $|l| \ge k'$, *то* $G_{k,k',l} = 0$.

Если $|l| < k'$, *то*

$$G_{k,k',l} = p\left(\frac{k'-|l|}{\sqrt{kk'}}\right) q\left(\frac{|l|}{\sqrt{kk'}}\right) p\left(\frac{k''}{\sqrt{kk'}}\right) +$$

$$+q\left(\frac{k'-|l|}{\sqrt{kk'}}\right) p\left(\frac{|l|}{\sqrt{kk'}}\right) q\left(\frac{k''}{\sqrt{kk'}}\right) - 1/4,$$

где $k'' = \min\{k',\ k-|l|\}$.

Доказательство.

Будем использовать обозначение $j' = j + k + l - k'$.

Отметим, что в формуле (20) при $|l| \ge k'$ либо $S_{k,j}$ (при $l \ge k'$), либо $S_{k,j+k}$ (при $l \le -k'$) не зависит от $S_{k',j'},\ \ S_{k',j'+k'}$.

Например, при $l \ge k'$ (то есть при $j' \ge j + k$)

$$G_{k,k',l} = 2\mathbf{P}\{S_{k,j} < 0,\ S_{k,j+k} > 0,\ S_{k',j'} < 0,\ S_{k',j'+k'} > 0\}+$$

$$+2\mathbf{P}\{S_{k,j} > 0,\ S_{k,j+k} < 0,\ S_{k',j'} < 0,\ S_{k',j'+k'} > 0\} - 1/4 =$$

$$= \frac{1}{4}\mathbf{P}\{S_{k,j+k} > 0 \mid S_{k',j'} < 0,\ S_{k',j'+k'} > 0\}+$$

$$+\frac{1}{4}\mathbf{P}\{S_{k,j+k} < 0 \mid S_{k',j'} < 0,\ S_{k',j'+k'} > 0\} - 1/4 =$$

$$= \frac{1}{4}\mathbf{P}\{S_{k,j+k} \ne 0\} - 1/4 = 0.$$

В случае $|l| < k'$ воспользуемся тем, что суммы $S_{k,j}$ и $S_{k',j'}$, имеющие m общих слагаемых, имеют коэффициент корреляции $r = m/\sqrt{kk'}$. Используя лемму 2.1 и расписывая формулу (20) через условные вероятности, получаем (см. рис. для случая $l > 0$):

$$G_{k,k',l} = 2\mathbf{P}\{S_{k,j} < 0,\ S_{k,j+k} > 0,\ S_{k',j'} < 0,\ S_{k',j'+k'} > 0\}+$$

$$+2\mathbf{P}\{S_{k,j} > 0,\ S_{k,j+k} < 0,\ S_{k',j'} < 0,\ S_{k',j'+k'} > 0\} - 1/4 =$$

$$= p\left(\frac{k'-|l|}{\sqrt{kk'}}\right) q\left(\frac{|l|}{\sqrt{kk'}}\right) p\left(\frac{k''}{\sqrt{kk'}}\right) +$$

$$+q\left(\frac{k'-|l|}{\sqrt{kk'}}\right) p\left(\frac{|l|}{\sqrt{kk'}}\right) q\left(\frac{k''}{\sqrt{kk'}}\right) - 1/4,$$

где $k'' = \min\{k',\ k-|l|\}$ — при $0 \le l < k'$ это число общих слагаемых в суммах $S_{k,j+k}$ и $S_{k',j'+k'}$.

Доказательство завершено.

Обозначим

$$\sigma_K = \frac{1}{K}\sqrt{\sum_{k=1}^{K}\sum_{k'=1}^{K}\sum_{|l|<\min\{k,\ k'\}} G(k,\ k',\ l)}, \tag{21}$$

где $G(k,\ k',\ l)$ определяются согласно предыдущей теореме.

Следствие 2.1 *Если для некоторого $K \ge 1$ выполнено $\sigma_K > 0$, то при $H = 1/2$ оценка $p_n^*(K)$ вероятности перемены знака $p = 1/2$ является асимптотически нормальной с параметром σ_K, то есть*

$$\sqrt{n}\frac{p_n^*(K) - \frac{1}{2}}{\sigma_K}$$

сходится по распределению к стандартному нормальному закону при $n \to \infty$.

Доказательство.

Заметим, что

$$\mathbf{D}V_n(K) = \sum_{k=1}^{K}\sum_{k'=1}^{K} \mathbf{cov}(L_k,\ L_{k'}).$$

При $n \to \infty$ имеют место эквивалентности

$$\mathbf{cov}(L_k,\ L_{k'}) = \sum_{j=0}^{n-2k}\sum_{j'=0}^{n-2k'} \mathbf{cov}(J_{k,j},\ J_{k',j'})$$

$$\sim \sum_{j=0}^{n-2k}\sum_{|l|<\min\{k,\ k'\}} G(k,\ k',\ l) \sim n \sum_{|l|<\min\{k,\ k'\}} G(k,\ k',\ l).$$

Получаем $\mathbf{D}V_n(K) \sim nK^2{\sigma_K}^2$, и, пользуясь замечанием в конце параграфа 2 и определением $p_n^*(K)$, доказываем асимптотическую нормальность. Доказательство завершено.

Для применения критерия важно знать поведение статистики при выполнении альтернативной гипотезы.

Докажем общую теорему о корреляции индикаторов перемены знака элементами стационарной гауссовской последовательности с нулевым математическим ожиданием и медленно убывающей корреляционной функцией. Затем применим эту теорему к доказательству асимптотической нормальности при альтернативной гипотезе и анализу непрерывности коэффициента асимптотической нормальности при $H \to 1/2$.

Теорема 2.3 *Пусть* $(X_1,\ X_2,\ X_3(k),\ X_4(k))$ — *последовательность четырехмерных нормальных случайных векторов,* $k = 3,\ 4,\ \ldots,$ *с нулевым математическим ожиданием и корреляционной матрицей*

$$R = \begin{pmatrix} 1 & r(1) & r(k) & r(k+1) \\ r(1) & 1 & r(k-1) & r(k) \\ r(k) & r(k-1) & 1 & r(1) \\ r(k+1) & r(k) & r(1) & 1 \end{pmatrix}. \qquad (22)$$

Будем предполагать, что $r(k) \neq 0$ *для всех* $k \geq 1$, $r(k) \to 0$ *при* $k \to \infty$, *и что* $r(k+1) \sim r(k)$ *при* $k \to \infty$.

Обозначим $I_1 = \mathbf{I}\{X_1X_2 < 0\},\ \ I_2 = \mathbf{I}\{X_3(k)X_4(k) < 0\}$. *Тогда существует константа* C *такая, что для всех* $k \geq 3$ *выполнено*

$$|\mathbf{cov}(I_1,\ I_2)| \leq Cr^2(k). \qquad (23)$$

Доказательство.

В тривиальном случае $r(1) = 1$ неравенство выполнено. В дальнейшем будем предполагать, что $r(1) < 1$.

Без ограничения общности будем считать, что случайные величины $X_1,\ X_2,\ X_3(k),\ X_4(k)$ имеют стандартные нормальные распределения — в противном случае их можно стандартизировать, разделив на стандартное отклонение.

Отметим, что (с учетом симметрии)

$$\begin{aligned}\mathbf{cov}(I_1,\ I_2) = \mathbf{P}\{X_1 \cdot X_2 < 0,\ X_3(k) \cdot X_4(k) < 0\}\\ -\mathbf{P}\{X_1 \cdot X_2 < 0\}\mathbf{P}\{X_3(k) \cdot X_4(k) < 0\}\\ = 2\mathbf{P}\{X_1 > 0,\ X_2 < 0,\ X_3(k) < 0,\ X_4(k) > 0\}\\ +2\mathbf{P}\{X_1 > 0,\ X_2 < 0,\ X_3(k) > 0,\ X_4(k) < 0\}\\ -2\mathbf{P}\{X_1 > 0,\ X_2 < 0\}\mathbf{P}\{X_3(k) < 0,\ X_4(k) > 0\}\\ -2\mathbf{P}\{X_1 > 0,\ X_2 < 0\}\mathbf{P}\{X_3(k) > 0,\ X_4(k) < 0\}.\end{aligned}$$

Найдем матрицу A преобразования, переводящего случайный вектор $\vec{\xi} = (\xi_1,\ \xi_2,\ \xi_3,\ \xi_4)$ с четырехмерным стандартным нормальным распределением в вектор $(X_1,\ X_2,\ X_3(k),\ X_4(k))$. Согласно (5.5) в [88], $R = AA^T$. Здесь R определяется формулой (22). Матрица A определяется с точностью до умножения на произвольную ортогональную матрицу. Найдем A в виде нижней треугольной матрицы. Пусть

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ c_1 & d_2 & 0 & 0 \\ b_1 & c_2 & d_3 & 0 \\ a_1 & b_2 & c_3 & d_4 \end{pmatrix}.$$

Тогда

$$R = \begin{pmatrix} 1 & c_1 & b_1 & a_1 \\ c_1 & {c_1}^2 + {d_2}^2 & b_1c_1 + c_2d_2 & a_1c_1 + b_2d_2 \\ b_1 & b_1c_1 + c_2d_2 & {b_1}^2 + {c_2}^2 + {d_3}^2 & a_1b_1 + b_2c_2 + c_3d_3 \\ a_1 & a_1c_1 + b_2d_2 & a_1b_1 + b_2c_2 + c_3d_3 & {a_1}^2 + {b_2}^2 + {c_3}^2 + {d_4}^2 \end{pmatrix}.$$

Решая систему уравнений, получаем:

$$\begin{cases} c_1 = r(1) \\ b_1 = r(k) \\ a_1 = r(k+1) \\ d_2 = \sqrt{1 - r^2(1)} \\ c_2 = \frac{r(k-1) - r(1)r(k)}{\sqrt{1-r^2(1)}} \\ b_2 = \frac{r(k) - r(1)r(k+1)}{\sqrt{1-r^2(1)}} \\ d_3 = \sqrt{1 - r^2(k) - {c_2}^2} \\ c_3 = \frac{r(1) - r(k+1)r(k) - b_2 c_2}{d_3} \\ d_4 = \sqrt{1 - r^2(k+1) - {b_2}^2 - {c_3}^2} \end{cases}$$

Здесь из-за громоздкости выражений элементы d_3, c_3 и d_4 выражены через предыдущие элементы.

Исследуем асимптотику элементов матрицы A при $k \to \infty$.

Обозначим

$$\beta = \frac{1 - r(1)}{d_2} = \sqrt{\frac{1 - r(1)}{1 + r(1)}}.$$

Тогда

$$c_2 \sim b_2 \sim \beta r(k), \qquad d_3 = 1 + O(r^2(k)),$$

$$c_3 = r(1) + O(r^2(k)), \qquad d_4 = d_2 + O(r^2(k)).$$

Отметим, что

$$X_1 > 0 \;\Leftrightarrow\; \xi_1 > 0;$$

$$X_2 < 0 \;\Leftrightarrow\; c_1\xi_1 + d_2\xi_2 < 0 \;\Leftrightarrow\; \xi_2 < -\frac{c_1}{d_2}\xi_1;$$

$$X_3(k) > 0 \;\Leftrightarrow\; b_1\xi_1 + c_2\xi_2 + d_3\xi_3 > 0 \;\Leftrightarrow\; \xi_3 > -\frac{b_1}{d_3}\xi_1 - \frac{c_2}{d_3}\xi_2;$$

$$X_4(k) < 0 \;\Leftrightarrow\; \xi_4 < -\frac{a_1}{d_4}\xi_1 - \frac{b_2}{d_4}\xi_2 - \frac{c_3}{d_4}\xi_3.$$

Обозначим $c_1^0 = \frac{c_1}{d_2}$, $b_1^0 = \frac{b_1}{d_3}$, $c_2^0 = \frac{c_2}{d_3}$, $a_1^0 = \frac{a_1}{d_4}$, $b_2^0 = \frac{b_2}{d_4}$, $c_3^0 = \frac{c_3}{d_4}$.

Заметим, что

$$c_1^0 = \frac{r(1)}{\sqrt{1-r^2(1)}},$$
$$b_1^0 = r(k) + O(r^3(k)),$$
$$c_2^0 \sim \beta r(k),$$
$$a_1^0 \sim \frac{r(k)}{\sqrt{1-r^2(1)}},$$
$$b_2^0 \sim \frac{\beta r(k)}{\sqrt{1-r^2(1)}},$$
$$c_3^0 = \frac{r(1)}{\sqrt{1-r^2(1)}} + O(r^2(k)).$$

Оценим $\mathbf{cov}(I_1,\ I_2)$, представив ее в виде четырехкратного интеграла от плотности четырехмерного стандартного нормального распределения:

$$\begin{aligned}\mathbf{cov}(I_1,\ I_2) &= 2\mathbf{P}\{X_1>0,\ X_2<0,\ X_3(k)<0,\ X_4(k)>0\}\\ &+2\mathbf{P}\{X_1>0,\ X_2<0,\ X_3(k)>0,\ X_4(k)<0\}\\ &-2\mathbf{P}\{X_1>0,\ X_2<0\}\mathbf{P}\{X_3(k)<0,\ X_4(k)>0\}\\ &-2\mathbf{P}\{X_1>0,\ X_2<0\}\mathbf{P}\{X_3(k)>0,\ X_4(k)<0\}\end{aligned}$$

$$= 2\int\limits_0^\infty dt_1 \int\limits_{-\infty}^{-c_1^0 t_1} dt_2 \left(\int\limits_{-b_1^0 t_1 - c_2^0 t_2}^{\infty} dt_3 \int\limits_{-\infty}^{-a_1^0 t_1 - b_2^0 t_2 - c_3^0 t_3} \varphi dt_4 - \int\limits_0^\infty dt_3 \int\limits_{-\infty}^{-c_1^0 t_3} \varphi dt_4 \right)$$

$$+2\int\limits_0^\infty dt_1 \int\limits_{-\infty}^{-c_1^0 t_1} dt_2 \left(\int\limits_{-\infty}^{-b_1^0 t_1 - c_2^0 t_2} dt_3 \int\limits_{-a_1^0 t_1 - b_2^0 t_2 - c_3^0 t_3}^{\infty} \varphi dt_4 - \int\limits_{-\infty}^{0} dt_3 \int\limits_{-c_1^0 t_3}^{\infty} \varphi dt_4 \right),$$

где

$$\varphi = \varphi(t_1,\ t_2,\ t_3,\ t_4) = \frac{1}{(2\pi)^2} e^{-\frac{{t_1}^2+{t_2}^2+{t_3}^2+{t_4}^2}{2}}$$

— плотность четырехмерного стандартного нормального закона.

Представим $\mathbf{cov}(I_1,\ I_2)$ в виде $\mathbf{cov}(I_1,\ I_2) = \alpha_1 + \alpha_2$, где α_1 получено заменой во внутренних интегралах c_1^0 на c_3^0, а α_2 — остаток после такой замены:

$$\alpha_1 = 2\int\limits_0^\infty dt_1 \int\limits_{-\infty}^{-c_1^0 t_1} dt_2 \left(\int\limits_{-b_1^0 t_1 - c_2^0 t_2}^{\infty} dt_3 \int\limits_{-\infty}^{-a_1^0 t_1 - b_2^0 t_2 - c_3^0 t_3} \varphi dt_4 - \int\limits_0^\infty dt_3 \int\limits_{-\infty}^{-c_3^0 t_3} \varphi dt_4 \right)$$

$$+2\int\limits_0^\infty dt_1 \int\limits_{-\infty}^{-c_1^0 t_1} dt_2 \left(\int\limits_{-\infty}^{-b_1^0 t_1 - c_2^0 t_2} dt_3 \int\limits_{-a_1^0 t_1 - b_2^0 t_2 - c_3^0 t_3}^{\infty} \varphi dt_4 - \int\limits_{-\infty}^{0} dt_3 \int\limits_{-c_3^0 t_3}^{\infty} \varphi dt_4 \right);$$

$$\alpha_2 = 2\int\limits_0^\infty dt_1 \int\limits_{-\infty}^{-c_1^0 t_1} dt_2 \Big(\int\limits_0^\infty dt_3 \left(\int\limits_{-\infty}^{-c_3^0 t_3} \varphi dt_4 - \int\limits_{-\infty}^{-c_1^0 t_3} \varphi dt_4 \right)$$

$$+ \int\limits_{-\infty}^{0} dt_3 \left(\int\limits_{-c_3^0 t_3}^{\infty} \varphi dt_4 - \int\limits_{-c_1^0 t_3}^{\infty} \varphi dt_4 \right) \Big).$$

Исследуем поведение α_2 при $k \to \infty$. Переходя в полярные системы координат по парам переменных $(t_1,\ t_2)$ и $(t_3,\ t_4)$, получаем:

$$\alpha_2 = \frac{4}{(2\pi)^2} \int\limits_{-\pi/2}^{-\arcsin r(1)} d\phi_1 \int\limits_0^\infty e^{-{\rho_1}^2/2} \rho_1 d\rho_1 \times$$

$$\times \int\limits_{-\arcsin r(1)}^{-\arcsin r(1) + O(r^2(k))} d\phi_2 \int\limits_0^\infty e^{-{\rho_2}^2/2} \rho_2 d\rho_2$$

$$= \frac{1}{\pi^2} \arccos r(1) \cdot O(r^2(k)) = O(r^2(k)).$$

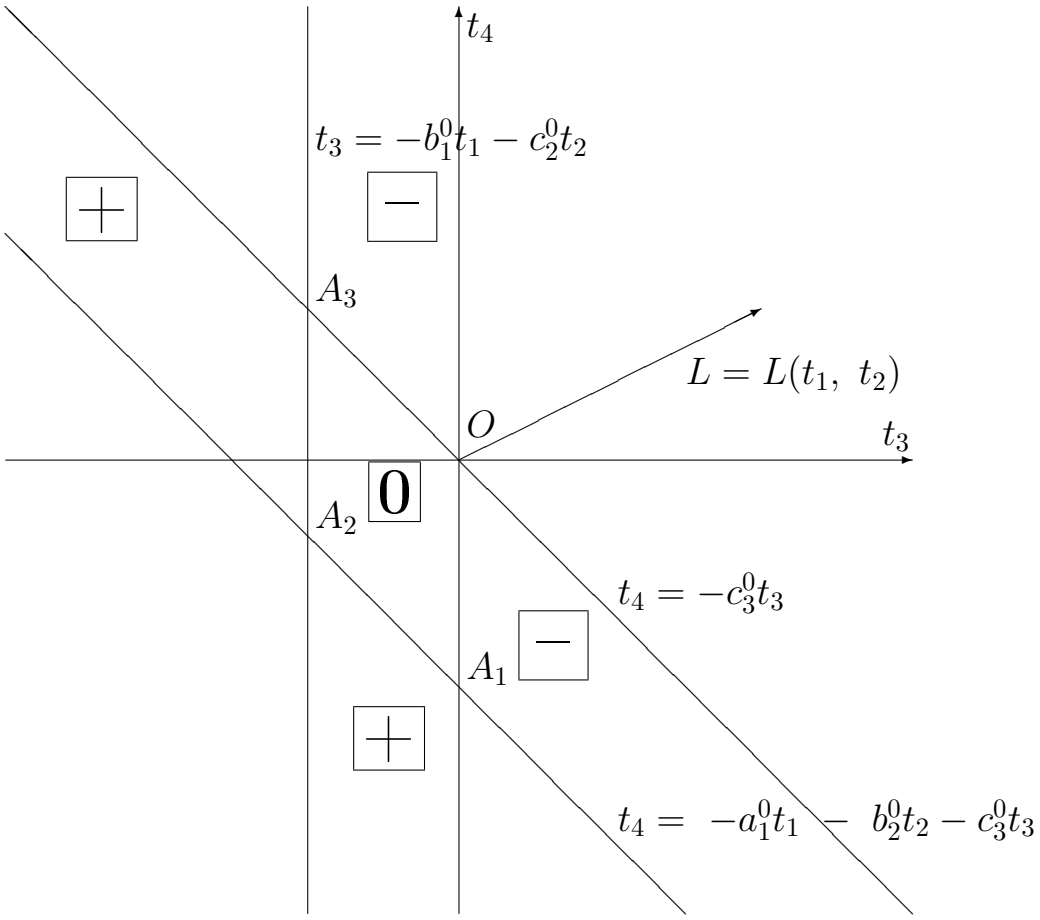

Рис. 2.2.

Исследуем поведение α_1 при $k \to \infty$. На рис. 2.2 показаны области положительности и отрицательности разности интегралов

$$\int\limits_{-b_1^0 t_1 - c_2^0 t_2}^{\infty} dt_3 \int\limits_{-\infty}^{-a_1^0 t_1 - b_2^0 t_2 - c_3^0 t_3} \varphi dt_4 - \int\limits_{0}^{\infty} dt_3 \int\limits_{-\infty}^{-c_3^0 t_3} \varphi dt_4$$

для некоторых $t_1,\ t_2$. Заметим, что для любых $t_1,\ t_2$ можно выбрать достаточно большое $L = L(t_1,\ t_2)$ такое, что приведенную выше разность интегралов можно оценить сверху через двойной интеграл от φ по кругу ${t_3}^2 + {t_4}^2 \le L^2$, так как вне этого круга разность интегралов равна 0 в силу симметрии функции φ.

Достаточно взять L такое, что

$$|OA_1| \le L;\ |OA_2| \le L;\ |OA_3| \le L.$$

Точки $A_1,\ A_2,\ A_3$ имеют координаты (зависящие от $t_1,\ t_2$)

$$A_1(0;\ -a_1^0 t_1 - b_2^0 t_2);$$
$$A_2(-b_1^0 t_1 - c_2^0 t_2;\ -a_1^0 t_1 - b_2^0 t_2 + c_3^0(b_1^0 t_1 + c_2^0 t_2));$$
$$A_2(-b_1^0 t_1 - c_2^0 t_2;\ c_3^0(b_1^0 t_1 + c_2^0 t_2)).$$

Поэтому достаточно выполнения неравенства

$$L = L(t_1,\ t_2) \ge (a_1^0 + b_1^0 + c_3^0 b_1^0)|t_1| + (b_2^0 + c_2^0 + c_3^0 c_2^0)|t_2|.$$

Положим

$$L = L(t_1,\ t_2) = c\sqrt{{t_1}^2 + {t_2}^2},$$

где $c = a_1^0 + b_1^0 + b_2^0 + c_2^0 + c_3^0(b_1^0 + c_2^0)$.

В силу найденных асимптотик, $c = O(r(k))$. Итак,

$$\alpha_1 \le 4 \int\limits_0^\infty dt_1 \int\limits_{-\infty}^{-c_1^0 t_1} dt_2 \int\int_{{t_3}^2 + {t_4}^2 \le L^2(t_1,\ t_2)} \varphi dt_3 dt_4$$

$$= \frac{4}{(2\pi)^2} \int\limits_{-\pi/2}^{-\arcsin r(1)} d\phi_1 \int\limits_0^\infty e^{-{\rho_1}^2/2} \rho_1 d\rho_1 \int\limits_0^{2\pi} d\phi_2 \int\limits_0^{c\rho_1} e^{-{\rho_2}^2/2} \rho_2 d\rho_2$$

$$= \frac{1}{\pi^2} \arccos r(1) \cdot 2\pi \int\limits_0^\infty e^{-{\rho_1}^2/2} \rho_1 (1 - e^{-c{\rho_1}^2/2}) d\rho_1$$

$$= \frac{2}{\pi} \arccos r(1) \cdot \frac{c^2}{1 + c^2} = O(r^2(k)).$$

Разделив $\mathbf{cov}(I_1,\ I_2)$ на $r^2(k)$, получаем, что существует верхний предел

$$\lim_{n \to \infty} \sup \frac{|\mathbf{cov}(I_1,\ I_2)|}{r^2(k)} < \infty.$$

Существование константы C из утверждения теоремы следует из ограниченности сходящейся последовательности.

Теорема доказана.

В качестве очевидного следствия теоремы получим

Следствие 2.2 *Пусть* $\{X_i\}$ *— стационарная гауссовская последовательность с нулевым математическим ожиданием и корреляционной функцией* $r(k)$*. Будем предполагать, что* $r(k) \neq 0$ *для всех* $k \geq 1$*,* $r(k) \to 0$ *при* $k \to \infty$*, и что* $r(k+1) \sim r(k)$ *при* $k \to \infty$*. Тогда существует константа* C *такая, что для всех* $k \geq 1$ *выполнено*

$$|\mathbf{cov}(I\{X_1X_2 < 0\};\ I\{X_{k+1}X_{k+2} < 0\})| \leq Cr^2(k). \qquad (24)$$

Теперь будем предполагать, что выполнена альтернативная гипотеза $H \neq 1/2$. Без ограничения общности можно полагать, что $X_1,\ X_2,\ \ldots$ имеют стандартное нормальное распределение. Корреляционная функция этой последовательности $r(k)$ определяется формулой (11).

Используя теорему предыдущего параграфа, получим при альтернативной гипотезе аналог предыдущего следствия. Напомним, что константа σ_K определяется формулой (21).

Следствие 2.3 *Если для некоторого* $K \geq 1$ *выполнено* $\sigma_K > 0$*, то при* H *из некоторой окрестности* $1/2$ *оценка* $p_n^*(K)$ *вероятности перемены знака* $p = p(r(1))$ *является асимптотически нормальной с параметром* $\sigma_K(H)$*, причем* $\lim_{H\to 1/2} \sigma_K(H) = \sigma_K$*.*

Доказательство.

Отметим, что согласно (11) корреляционная функция $r(k)$ монотонна при $k \geq 0$: убывает при $H > 1/2$, и возрастает при $H < 1/2$. Этот результат можно получить непосредственно дифференцированием, считая k неотрицательной действительной переменной.

Кроме того, $r(k)$ при $H \neq 1/2$ удовлетворяет условиям теоремы: $r(k) \neq 0,\ \ r(k) \to 0$ при $k \to \infty$, $r(k) \sim r(k+1)$.

Вынося множитель k^{2H} из формулы (11) и раскладывая остаток в ряд по степеням k^{-1}, получаем:

$$r(k) = H(2H-1)k^{2H-2} + O(k^{2H-4}) \tag{25}$$

при $k \to \infty$.

Для доказательства центральной предельной теоремы для $V_n(K)$ используем теорему 18.5.4 из [42]. Для этого доопределим стационарную последовательность $\{X_i\}$ для всех целых i и для $K \geq 1$ рассмотрим последовательность случайных величин

$$J_i = J_{1,i} + \ldots + J_{K,i}$$

и оценим коэффициент сильного перемешивания $\alpha(n)$:

$$\alpha(n) = \sup_{A \in \mathcal{F}_{-\infty}^0,\ B \in \mathcal{F}_n^\infty} |\mathbf{P}(AB) - \mathbf{P}(A)\mathbf{P}(B)|$$

согласно формуле (17.2.1) в [42]. Здесь в качестве сигма-алгебр $\mathcal{F}_{-\infty}^0$ и $\mathcal{F}_n^\infty$ возьмем сигма-алгебры, порожденные случайными векторами $\mathbf{J}_i = (J_{1,i};\ \ldots\ ; J_{K,i})$ с номерами от $-\infty$ до 0 и от n до ∞ соответственно. Заметим, что

$$\alpha(n) \leq \varrho(n) = \sup_{j \geq n} |\mathbf{cov}(J_0;\ J_j)| \leq \sup_{j \geq n} \sum_{k=1}^{K} \sum_{k'=1}^{K} |\mathbf{cov}(J_{k,0};\ J_{k',j})|.$$

Применим теорему к четырехмерному вектору $(S_{k,0},\ S_{k,k},\ S_{k',j},\ S_{k',j+k'})$. Условия теоремы выполнены, так как

$$|\mathbf{corr}(S_{k,0},\ S_{k',j})| = \left| \frac{\sum_{s=1}^{k} \sum_{s'=1}^{k'} \mathbf{cov}(X_s,\ X_{j+s'})}{\sqrt{\mathbf{D}S_{k,0}\mathbf{D}S_{k',j}}} \right|$$

$$= \left| \frac{\sum_{s=1}^{k} \sum_{s'=1}^{k'} r(j+s'-s)}{k^H (k')^H} \right| \leq |r(j-k)| k^{1-H} (k')^{1-H}.$$

Кроме того, имеет место эквивалентность

$$\mathbf{corr}(S_{k,0},\ S_{k',j}) \sim r(j) k^{1-H} (k')^{1-H}.$$

Поэтому корреляции сумм обладают теми же свойствами, что и корреляционная функция исходной последовательности. Используя теорему и монотонность корреляционной функции $r(k)$, получаем:

$$|\mathbf{cov}(J_{k,0};\ J_{k',j})| \le Cr^2(j-k)k^{2-2H}(k')^{2-2H},$$
$$\alpha(n) \le \sup_{j\ge n}\sum_{k=1}^{K}\sum_{k'=1}^{K}|\mathbf{cov}(J_{k,0};\ J_{k',j})| \le$$
$$\le Cr^2(n-k)k^{3-2H}(k')^{3-2H} \le C'r^2(n).$$

Итак, при $H < 3/4$ справедлива оценка

$$\sum_{n\ge n_0}\alpha(n) \le C'\sum_{n\ge n_0}r^2(n) \le C''\sum_{n\ge n_0}n^{4H-4} \to 0$$

при $n_0 \to \infty$ в силу (25). Поэтому условия теоремы 18.5.4 из [42] выполнены, если $\sigma_K^2(H) = \mathbf{D}J_0 + 2\sum_{j=1}^{\infty}\mathbf{cov}(J_0,\ J_j) \ne 0$.

Докажем сходимость $\sigma_K(H) \to \sigma_K$ при $H \to 1/2$: на любом конечном промежутке $[0;\ n_0]$ сходимость ковариаций $\mathbf{cov}(J_0,\ J_j)$ имеет место в силу непрерывности $r(j)$ по H, а остаток $\sum_{j\ge n_0}\mathbf{cov}(J_0,\ J_j)$ сходится к нулю в силу сделанных оценок. Если $\sigma_K > 0$, то в некоторой окрестности $H = 1/2$ выполнены все условия теоремы 18.5.4 из [42].

Следствие доказано.

Согласно вышеизложенному, предлагаемые критерии проверки гипотезы $H = 1/2$ будут иметь асимптотический уровень ε, если их критическая область определяется условием

$$\left|p_n^*(K) - \frac{1}{2}\right| \ge \frac{A\sigma_K}{\sqrt{n}},$$

где A — квантиль уровня $1-\varepsilon/2$ стандартного нормального распределения.

Так как вероятность перемены знака $p = p(r(1))$ и параметр H связаны соотношением

$$H = \frac{1}{2}\left(1 + \log_2(1+\cos(\pi p))\right),$$

то оценка $\tilde{H}_n(K)$, вычисляемая на основании статистики $p_n^*(K)$, имеет вид

$$\tilde{H}_n(K) = \frac{1}{2}\left(1 + \log_2(1 + \cos(\pi p_n^*(K)))\right).$$

Согласно теореме об асимптотической нормальности [9], эта оценка является асимптотически нормальной с коэффициентом $b_K(H) = |H'(p)|\sigma_K(H)$. При $p = H = 1/2$ получаем

$$b_K = b_K(1/2) = \frac{\pi}{2\ln 2}\sigma_K. \tag{26}$$

Отметим, что в работе [119] предложена асимптотически эффективная оценка параметра H, основанная на методе периодограммы. Коэффициент асимптотической нормальности для этой оценки согласно теореме 2.1 в [119] равен $b = \Gamma^{-1/2}$, где

$$\Gamma = \frac{1}{4\pi}\int_{-\pi}^{\pi}(\frac{d}{dH}\ln f_H(x))^2 dx,$$

где $f_H(x) = \frac{1}{2\pi}\left(r(0) + 2\sum_{k=1}^{\infty} r(k)\cos kx\right)$ — спектральная плотность.

$$\frac{d}{dH}\ln f_H(x) = \frac{1}{\pi f_H(x)}\sum_{k=1}^{\infty}\frac{d}{dH}r(k)\cos kx =$$

$$= \frac{2}{\pi f_H(x)}\sum_{k=1}^{\infty}\Big(|k+1|^{2H}\ln|k+1| +$$

$$+|k-1|^{2H}\ln|k-1| - 2|k|^{2H}\ln|k|\Big)\cos kx.$$

При $H = 1/2$ получаем $f_H(x) \equiv 1/2\pi$,

$$\frac{d}{dH}\ln f_H(x) = 4\sum_{k=1}^{\infty}(|k+1|\ln|k+1| +$$

$$+ |k-1|\ln|k-1| - 2|k|\ln|k|)\cos kx,$$

$$\Gamma = \frac{4}{\pi}\sum_{k=1}^{\infty}(|k+1|\ln|k+1| + |k-1|\ln|k-1| -$$

$$-2|k|\ln|k|)^2\int_{-\pi}^{\pi}\cos^2 kx\ dx =$$

$$= 4\left((2\ln 2)^2 + \sum_{k=2}^{\infty}\left((k+1)\ln(k+1) + (k-1)\ln(k-1) - 2k\ln k\right)^2\right).$$

Расчеты дают $\Gamma \approx 10,39,\ \ b \approx 0,3102$.

Вычислим σ_K и b_K по формулам (21) и (26). Результаты вычислений представлены в таблице.

K	1	2	4	8	16	32
σ_K	0,5	0,377	0,294	0,239	0,204	0,178
b_K	1,133	0,855	0,665	0,542	0,462	0,404

Итак, при больших K коэффициент асимптотической нормальности становится близок к оптимальному.

Отметим, что можно модифицировать статистику $V_n(K)$, суммируя индикаторы перемены знака с различными весами. Расчеты показывают, что такой подход не приносит существенных преимуществ, так как при больших K оптимальные веса становятся почти равными.

В качестве альтернативы модифицированному знаковому методу рассмотрим следующий *бинарный знаковый метод* оценивания параметра Херста. Он оказывается более удачным, чем модифицированный знаковый метод (оценки этим методом имеют меньшую дисперсию) в силу того, что здесь устранены большие положительные корреляции слагаемых.

Вычислительная сложность этого алгоритма имеет порядок n в отличие от алгоритма максимального правдоподобия, имеющего сложность порядка n^2.

Введем в рассмотрение бинарный знаковый алгоритм оценивания параметра H, при котором исходные значения суммируются блоками длины 2^k.

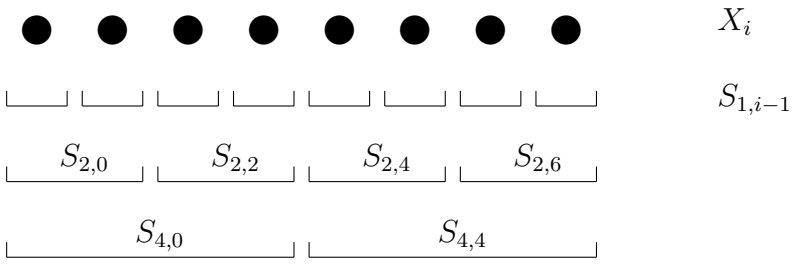

Рис. 2.3.

Для упрощения изложения будем предполагать, что объем данных n является целой степенью числа 2, т. е. $\log_2 n$ — целое число.

Рассмотрим статистику

$$U_n = \sum_{k=0}^{\log_2 n-1} \sum_{j=0}^{n2^{-k}-2} \mathbf{I}\{S_{2^k,2^k j} \cdot S_{2^k,2^k(j+1)} < 0\}.$$

В качестве оценки вероятности перемены знака используется частота перемены знака, подсчитанная на основании статистики U_n:

$$p_n^* = \frac{U_n}{\sum_{k=0}^{\log_2 n-1}(n2^{-k}-1)} = \frac{U_n}{2n-2-\log_2 n}.$$

Оценка H_n^*, вычисляемая на основании статистики p_n^*, имеет вид

$$H_n^* = \frac{1}{2}\left(1+\log_2(1+\cos(\pi p_n^*))\right).$$

Будем предполагать, что выполнена нулевая гипотеза $H = 1/2$, то есть исходные случайные величины независимы — образуют выборку из нормального распределения с нулевым математическим ожиданием.

Для последующих вычислений будет полезно следующее обозначение: $G_{k,k'}$ — ковариация индикаторов перемены знака парами блоков из k и k' слагаемых, середины которых совпадают.

Введя обозначение $j' = k + j - k'$, определим $G_{k,k'}$ формально как

$$G_{k,k'} = \mathbf{cov}\ (J_{k,j},\ J_{k',j'}) =$$

$$= \mathbf{P}\{S_{k,j} \cdot S_{k,j+k} < 0,\ S_{k',j'} \cdot S_{k',j'+k'} < 0\} -$$

$$-\mathbf{P}\{S_{k,j} \cdot S_{k,j+k} < 0\} \cdot \mathbf{P}\{S_{k',j'} \cdot S_{k',j'+k'} < 0\} =$$

$$= \mathbf{P}\{S_{k,j} \cdot S_{k,j+k} < 0,\ S_{k',j'} \cdot S_{k',j'+k'} < 0\} - 1/4.$$

Здесь $J_{k,j} = \mathbf{I}\{S_{k,j} \cdot S_{k,j+k} < 0\}$ — индикатор перемены знака последовательными суммами k слагаемых, первое из которых имеет номер $j+1$.

Рис. 2.4 поясняет введенное обозначение.

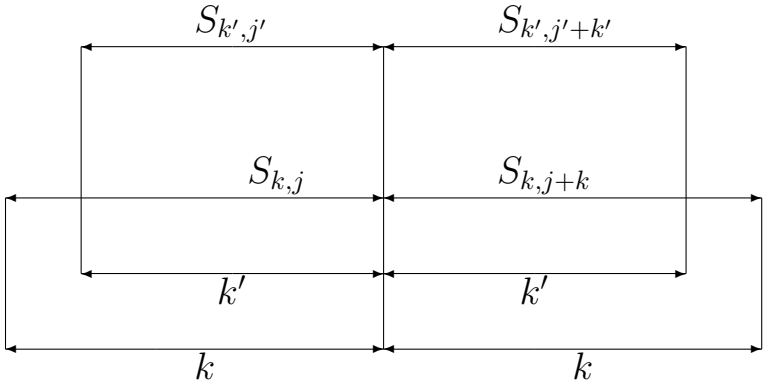

Рис. 2.4.

Отметим, что по определению $G_{k,k'} = G_{k',k}$.

Вычислим коэффициенты $G_{k,k'}$ в случае, когда $k' \le k$.

Теорема 2.4 *Пусть* $0 < k' \le k$. *Тогда*

$$G_{k,k'} = \left(\frac{1}{\pi} \arccos \sqrt{k'/k} - \frac{1}{2}\right)^2.$$

Доказательство.

Будем использовать обозначение $j' = j + k - k'$.

С учетом симметрии

$$G_{k,k'} = 2\mathbf{P}\{S_{k,j} < 0,\ S_{k,j+k} > 0,\ S_{k',j'} < 0,\ S_{k',j'+k'} > 0\} +$$

$$+2\mathbf{P}\{S_{k,j} > 0,\ S_{k,j+k} < 0,\ S_{k',j'} < 0,\ S_{k',j'+k'} > 0\} - 1/4.$$

В силу независимости левой и правой частей,

$$G_{k,k'} = 2\mathbf{P}\{S_{k,j} < 0,\ S_{k',j'} < 0\}\mathbf{P}\{S_{k,j+k} > 0,\ S_{k',j'+k'} > 0\}+$$

$$+2\mathbf{P}\{S_{k,j} > 0,\ S_{k',j'} < 0\}\mathbf{P}\{S_{k,j+k} < 0,\ S_{k',j'+k'} > 0\} - 1/4.$$

В силу симметрии

$$G_{k,k'} = 2\mathbf{P}\{S_{k,j} < 0\}\mathbf{P}\{S_{k,j} \cdot S_{k',j'} > 0 \mid S_{k,j} < 0\}\cdot$$

$$\cdot\mathbf{P}\{S_{k,j+k} > 0\}\mathbf{P}\{S_{k,j+k} \cdot S_{k',j'+k'} > 0 \mid S_{k,j+k} > 0\}+$$

$$+2\mathbf{P}\{S_{k,j} > 0\}\mathbf{P}\{S_{k,j} \cdot S_{k',j'} < 0 \mid S_{k,j} > 0\}\cdot$$

$$\cdot\mathbf{P}\{S_{k,j+k} < 0\}\mathbf{P}\{S_{k,j+k} \cdot S_{k',j'+k'} < 0 \mid S_{k,j+k} < 0\} - 1/4 =$$

$$= 2 \cdot \frac{1}{2}\mathbf{P}\{S_{k,j} \cdot S_{k',j'} > 0\} \cdot \frac{1}{2}\mathbf{P}\{S_{k,j+k} \cdot S_{k',j'+k'} > 0\}+$$

$$+2 \cdot \frac{1}{2}\mathbf{P}\{S_{k,j} \cdot S_{k',j'} < 0\} \cdot \frac{1}{2}\mathbf{P}\{S_{k,j+k} \cdot S_{k',j'+k'} < 0\} - 1/4 =$$

$$= \frac{1}{2}\left(1 - \mathbf{P}\{S_{k,j} \cdot S_{k',j'} < 0\}\right)^2 + \frac{1}{2}\left(\mathbf{P}\{S_{k,j} \cdot S_{k',j'} < 0\}\right)^2 - 1/4.$$

Обозначим $p = \mathbf{P}\{S_{k,j} \cdot S_{k',j'} < 0\}$. Тогда

$$G_{k,k'} = \frac{1}{2}((1-p)^2 + p^2) - 1/4 = (p - 1/2)^2.$$

Так как суммы $S_{k,j}$ и $S_{k',j'}$ имеют k' общих слагаемых, то их коэффициент корреляции $r = \sqrt{k'/k}$. Используя лемму 1, получаем утверждение теоремы.

Доказательство завершено.

Теорема 2.5 *Если $H = 1/2$, то при $n \to \infty$ имеет место слабая сходимость последовательности случайных величин $\sqrt{n}(p_n^* - 1/2)$ к нормальному закону с нулевым математическим ожиданием и дисперсией*

$$\sigma^2 = \frac{1}{8} + \sum_{s=1}^{\infty} 2^{-s}\left(\frac{1}{\pi}\arccos 2^{-s/2} - \frac{1}{2}\right)^2 \approx 0,1654.$$

Доказательству теоремы предпошлем следующую лемму. Она формализует утверждение о том, что оценка, являющаяся в некотором смысле предельной для последовательности асимптотически нормальных оценок, обладает свойством асимптотической нормальности.

Лемма 2.5 *Пусть выполнены следующие 3 условия:*

1) для любого $K < \infty$ оценка $\theta_n^*(K)$ является асимптотически нормальной оценкой параметра θ с коэффициентом σ_K;

2) имеет место сходимость $\sigma_K \to \sigma$ при $K \to \infty$, $0 < \sigma < \infty$;

3) для любого $\varepsilon > 0$ имеет место сходимость

$$\lim_{K\to\infty} \lim_{n\to\infty} \mathbf{P}\{\sqrt{n}|\theta_n^*(K) - \theta_n^*| \geq \varepsilon\} = 0.$$

Тогда θ^* — асимптотически нормальная оценка параметра θ с коэффициентом σ.

Доказательство.

Для любого $\varepsilon > 0$ выполнено

$$\mathbf{P}\{\sqrt{n}(\theta_n^* - \theta) < t\} \geq \mathbf{P}\{\sqrt{n}(\theta_n^*(K) - \theta) < t - \varepsilon,\ \sqrt{n}|\theta_n^* - \theta_n^*(K)| < \varepsilon\}.$$

Так как для любых событий A и B выполнено $\mathbf{P}(AB) \geq \mathbf{P}(A) - \mathbf{P}(\overline{B})$, то, переходя к пределу, получаем:

$$\lim_{n\to\infty} \mathbf{P}\{\sqrt{n}(\theta_n^* - \theta) < t\} \geq$$

$$\geq \lim_{n\to\infty} \mathbf{P}\{\sqrt{n}(\theta_n^*(K) - \theta) < t - \varepsilon)\} - \lim_{n\to\infty} \mathbf{P}\{\sqrt{n}|\theta_n^* - \theta_n^*(K)| \geq \varepsilon\}.$$

Согласно условию (1) леммы, первый из пределов правой части равен $\Phi(\sigma_K(t - \varepsilon))$, где $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2} dt$ — функция Лапласа.

Согласно условию (3) леммы, второй из пределов правой части равен $\delta_K > 0$, где δ_K можно выбором K сделать сколь угодно малым.

Итак,

$$\lim_{n\to\infty} \mathbf{P}\{\sqrt{n}(\theta_n^* - \theta) < t\} \geq \Phi(\sigma_K(t-\varepsilon)) - \delta_K.$$

Аналогично

$$\lim_{n\to\infty} \mathbf{P}\{\sqrt{n}(\theta_n^* - \theta) < t\} \leq \Phi(\sigma_K(t+\varepsilon)) + \delta_K.$$

Так как $\sigma_K \to \sigma,\ \delta_K \to 0$ при $K \to \infty$, функция Φ непрерывна, а константы $\varepsilon > 0$ и $K < \infty$ выбираются произвольно, то

$$\lim_{n\to\infty} \mathbf{P}\{\sqrt{n}(\theta_n^* - \theta) < t\} = \Phi(\sigma t).$$

Лемма доказана.

Следствие 2.4 *Если в условиях леммы 2 заменить условие (3) на условие*

$$\lim_{K\to\infty} \lim_{n\to\infty} n \cdot \mathbf{E}(\theta_n^*(K) - \theta_n^*)^2 = 0,$$

то утверждение леммы сохранится.

Доказательство очевидно из неравенства Чебышева.

Доказательство теоремы.

1) Вычислим $\mathbf{D}U_n$. Из определения получаем

$$\mathbf{D}U_n = \sum_{m=0}^{\log_2 n-1} \mathbf{D}J_{2^m,0} \cdot (n2^{-m} - 1) +$$

$$+2 \sum_{m=1}^{\log_2 n-1} \sum_{m'=0}^{m-1} G_{2^{m'},2^m} \cdot (n2^{-m} - 1) =$$

$$= \frac{1}{4}(2n - 2 - \log_2 n) +$$

$$+2 \sum_{m=1}^{\log_2 n-1} \sum_{m'=0}^{m-1} \left(\frac{1}{\pi} \arccos 2^{(m'-m)/2} - \frac{1}{2}\right)^2 (n2^{-m} - 1).$$

При $n \to \infty$ имеет место асимптотика

$$\mathbf{D}U_n \sim \frac{n}{2} + 2n \sum_{m=1}^{\infty} 2^{-m} \sum_{m'=0}^{m-1} \left(\frac{1}{\pi} \arccos 2^{(m'-m)/2} - \frac{1}{2}\right)^2.$$

Сделаем замену $s = m - m'$ и поменяем порядок суммирования:

$$\mathbf{D}U_n \sim \frac{n}{2} + 2n \sum_{s=1}^{\infty} \sum_{m=s}^{\infty} 2^{-m} \left(\frac{1}{\pi} \arccos 2^{-s/2} - \frac{1}{2} \right)^2 =$$

$$= \frac{n}{2} + 4n \sum_{s=1}^{\infty} 2^{-s} \left(\frac{1}{\pi} \arccos 2^{-s/2} - \frac{1}{2} \right)^2 .$$

Следовательно,

$$\mathbf{D}p_n^* \sim \frac{1}{4n^2} \left(\frac{n}{2} + 4n \sum_{s=1}^{\infty} 2^{-s} \left(\frac{1}{\pi} \arccos 2^{-s/2} - \frac{1}{2} \right)^2 \right) =$$

$$= \frac{1}{n} \left(\frac{1}{8} + \sum_{s=1}^{\infty} 2^{-s} \left(\frac{1}{\pi} \arccos 2^{-s/2} - \frac{1}{2} \right)^2 \right).$$

Докажем сходимость по распределению. Обозначим

$$U_n(K) = \sum_{m=0}^{K-1} \sum_{j=0}^{n2^{-m}-2} \mathbf{I}\{S_{2^m,2^m j} \cdot S_{2^m,2^m(j+1)<0}.$$

Тогда

$$\mathbf{D}U_n(K) = \sum_{m=0}^{K-1} \mathbf{D}J_{2^m,0}(n2^{-m} - 1) +$$

$$+2 \sum_{m=1}^{K-1} \sum_{m'=0}^{m-1} G_{2^{m'},2^m}(n2^{-m} - 1) =$$

$$= \frac{1}{4} \left(\frac{n(1 - 2^{-K})}{1 - 1/2} - K \right) +$$

$$+2 \sum_{m=1}^{K-1} \sum_{m'=0}^{m-1} (n2^{-m} - 1) \left(\frac{1}{\pi} \arccos 2^{(m'-m)/2} - \frac{1}{2} \right)^2 .$$

При $n \to \infty$

$$\mathbf{D}U_n(K) \sim Cn,$$

$C > 0$.

Обозначим

$$p_n^*(K) = \frac{U_n(K)}{\sum_{m=0}^{K-1}(n2^{-m} - 1)} = \frac{U_n(K)}{2n(1 - 2^{-K}) - K}.$$

Проверим условия предыдущей леммы.

1) Центральная предельная теорема для $U_n(K)$ имеет место в силу того, что $U_n(K)$ представима в виде суммы элементов стационарной последовательности 1-зависимых случайных величин (числа перемен знаков в последовательных блоках длины 2^{K+1}), при этом $\mathbf{D}U_n(K) \to \infty$.

Следовательно,

$$\lim_{n\to\infty} \mathbf{P}\{\sqrt{n}(p_n^* - 1/2) < t\} = \Phi(\sigma_K t).$$

2) Так как

$$\sigma_K^2 = \lim_{n\to\infty} n \cdot \mathbf{D}p_n^*(K) =$$

$$= \frac{1}{4(1-2^{-K})^2}\left(\frac{1}{2}(1-2^{-K}) + 2\sum_{m=1}^{K}\sum_{m'=0}^{m-1} 2^{-m}\left(\frac{1}{\pi}\arccos 2^{(m'-m)/2} - \frac{1}{2}\right)^2\right).$$

Сделаем замену $s = m - m'$ и поменяем порядок суммирования:

$$\sigma_K^2 = \frac{1}{4(1-2^{-K})^2}\left(\frac{1}{2}(1-2^{-K}) + 2\sum_{s=1}^{K}\sum_{m=s}^{K} 2^{-m}\left(\frac{1}{\pi}\arccos 2^{s/2} - \frac{1}{2}\right)^2\right) =$$

$$= \frac{1}{4(1-2^{-K})^2}\left(\frac{1}{2}(1-2^{-K}) + 2\sum_{s=1}^{K}(2^{-s+1} - 2^{-K})\left(\frac{1}{\pi}\arccos 2^{s/2} - \frac{1}{2}\right)^2\right).$$

Переходя к пределу при $K \to \infty$, устанавливаем сходимость $\sigma_K \to \sigma$.

3) Заметим, что разность

$$p_n^* - p_n^*(K) = \frac{U_n}{2n - 2 - \log_2 n} - \frac{U_n(K)}{2n(1-2^{-K}) - K}$$

имеет нулевое математическое ожидание.

Обозначим $\Delta_n(K) = U_n - U_n(K)$. Тогда

$$p_n^* - p_n^*(K) =$$

$$= \frac{\Delta_n(K)}{2n - 2 - \log_2 n} - \frac{U_n(K)(2n \cdot 2^{-K} + K - \log_2 n)}{(2n(1-2^{-K}) - K)(2n - 2 - \log_2 n)} \stackrel{def}{=} A_n + B_n.$$

Заметим, что

$$\mathbf{E}(p_n^* - p_n^*(K))^2 = \mathbf{D}(p_n^* - p_n^*(K)) \le \mathbf{D}A_n + \mathbf{D}B_n + 2\sqrt{\mathbf{D}A_n\mathbf{D}B_n},$$

$$\mathbf{D}A_n \sim \frac{\mathbf{D}U_n(K)2^{-2K}}{4n^2},$$
$$\mathbf{D}B_n \sim \frac{\mathbf{D}\Delta_n(K)}{4n^2}.$$

Так как

$$\Delta_n(K) = \sum_{m=K}^{\log_2 n-1} \sum_{j=0}^{n2^{-m}-2} \mathbf{I}\{S_{2^m,2^m j} \cdot S_{2^m,2^m(j+1)} < 0\},$$

то

$$\mathbf{D}\Delta_n(K) = \sum_{m=K}^{\log_2 n-1} \mathbf{D}J_{2^m,0}(n2^{-m}-1)+2\sum_{m=K}^{\log_2 n-1}\sum_{m'=0}^{m-1} G_{2^{m'},2^m}(n2^{-m}-1) =$$
$$= \sum_{m=K}^{\log_2 n-1} (n2^{-m}-1)\left(\frac{1}{4}+2\sum_{m'=0}^{m-1} G_{2^{m'},2^m}\right) \le$$
$$\le C_0 \sum_{m=K}^{\log_2 n-1} (n2^{-m}-1) \le 2C_0 n \cdot 2^{-K}$$

в силу сходимости ряда $\sum_{s=0}^{\infty} G_{1,2^s}$.

Переходя к пределу при $n \to \infty$, а затем при $K \to \infty$, обосновываем выполнение условия (3) предыдущей леммы.

Доказательство завершено.

Так как вероятность перемены знака p и параметр H связаны соотношением

$$H = \frac{1}{2}\left(1 + \log_2(1 + \cos(\pi p))\right),$$

то оценка $\tilde{H}_n$, вычисляемая на основании статистики p_n^*, имеет вид

$$\tilde{H}_n = \frac{1}{2}\left(1 + \log_2(1 + \cos(\pi p_n^*)\right).$$

Согласно теореме об асимптотической нормальности [9], эта оценка является асимптотически нормальной с коэффициентом $B(H) = |H'(p)|\sigma(H)$. При $p = H = 1/2$ получаем

$$B = B(1/2) = \frac{\pi}{2\ln 2}\sigma.$$

Итак, приходим к следующему утверждению.

Следствие 2.5 *Если $H = 1/2$, то при $n \to \infty$ имеет место слабая сходимость последовательности случайных величин $\sqrt{n}(\tilde{H}_n - 1/2)$ к нормальному закону с нулевым математическим ожиданием и дисперсией*

$$B^2 = \frac{\pi^2}{4\ln^2 2}\left(\frac{1}{8} + \sum_{s=1}^{\infty} 2^{-s}\left(\frac{1}{\pi}\arccos 2^{-s/2} - \frac{1}{2}\right)^2\right) \approx 0,8494.$$

Таким образом, при $n = 1024$ получаем стандартное отклонение $B/32 \approx 0,02880$. Этот результат оказывается лучше, чем для оценок модифицированным знаковым методом, и существенно лучше, чем для оценок элементарным знаковым методом и центрированным знаковым методом. Сравнение дисперсий также показывает, что эта оценка лучше других оценок (оценки методом дисперсии, оценки нормированного размаха), см. Женнан и др. [130].

2.5 Результаты главы 2

В главе 2 получены следующие основные результаты:

- Доказана сильная состоятельность оценок знаковым методом.
- Получены алгоритмы оценивания элементарным знаковым методом и центрированным знаковым методом.
- Получен алгоритм оценивания модифицированным знаковым методом, доказана асимптотическая нормальность оценки и вычислен коэффициент асимптотической нормальности при $H = 1/2$.
- Получен алгоритм оценивания бинарным знаковым методом, доказана асимптотическая нормальность оценки и вычислен коэффициент асимптотической нормальности при $H = 1/2$.
- Показано преимущество бинарного знакового метода.

ГЛАВА 3
Проверка гипотез для фрактального гауссовского шума и его обобщений

3.1 Вводные замечания

Напомним, что фрактальный гауссовский шум — это стационарная гауссовская последовательность с корреляционной функцией такого вида, что дисперсия частичных сумм растет по степенному закону.

В связи с использованием модели фрактального гауссовского шума требуют статистического тестирования следующие гипотезы:

- нормальность приращений (элементов выборки);
- наличие зависимости приращений (отличие коэффициента Херста от $1/2$);
- отсутствие разладки фрактального гауссовского шума.

Для проверки этих гипотез в параграфе 3.2 строится новый критерий проверки нормальности, основанный на модификации знакового метода. В параграфе 3.3 на основании бинарного знакового метода разрабатывается и исследуется знаковый критерий для проверки основной гипотезы об отсутствии зависимости против гипотезы о соответствии модели фрактального броуновского движения. В связи с тем, что не все гипотезы допускают аналитическую проверку, в параграфе 3.4 разработаны две процедуры моделирования фрактального броуновского движения. В параграфе 3.5 построен критерий выявления разладки фрактального броуновского движения, использующий результаты моделирования. В параграфе 3.6 предложен критерий проверки гипотезы о стационарности гауссовской последовательности. Критерий основан на вычислении плотности распределения отношения сумм элементов последовательности при

выполнении гипотезы. В параграфе 3.7 процедуры моделирования, оценивания параметров и проверки гипотез распространяются на последовательности, получаемые покоординатным преобразованием фрактального гауссовского шума для приведения в соответствие с устойчивым негауссовским законом распределения. Результаты главы 3 собраны в параграфе 3.8.

3.2 Новый статистический критерий тестирования нормальности малых выборок

Отметим, что как правило проверку нормальности предлагают проводить при объеме выборки n не менее 8 (см. [105]). Это связано с тем, что для отыскания критических уровней используется нормальное приближение для используемой статистики, а оно оказывается весьма неточным. В частности, ниже будет продемонстрирована величина погрешности, которая возникает при использовании нормальной аппроксимации для статистики критерия Шапиро—Уилка [157] при объеме выборки от 3 до 5. В частности, видимо, по этой причине согласно ГОСТу [21] признано допустимым проверять нормальность только при $n \geq 8$. Сходимость статистик других критериев проверки нормальности к предельному распределению еще медленнее, чем для критерия Шапиро—Уилка.

Однако уже при $n = 2$ можно (с нулевой вероятностью ошибки) отвергнуть гипотезу о нормальности, если выборочные значения совпадают. Хотелось бы для $n > 2$ построить критерий, наследующий это полезное свойство, но позволяющий также (с достаточно малой вероятностью ошибки) отвергать гипотезу о нормальности в некоторых случаях, когда все выборочные значения различны.

Обозначим

$$R_n = \max_{1 \le i,j \le n} |X_i - X_j|$$

— наибольшее расстояние между элементами (размах) выборки;

$$L_n = \min_{1 \le i < j \le n} |X_i - X_j|$$

— наименьшее расстояние между элементами выборки;

$$d_n = R_n / L_n$$

— их отношение. Будем полагать $d_n = +\infty$ при $L_n = 0$.

Согласно основной гипотезе, элементы выборки $X_1, \ldots, X_n$ имеют нормальное распределение. Критерий отвергает основную гипотезу, если $d_n \ge C$, где $0 < C < \infty$.

Величину L_n можно назвать минимальным спейсингом выборки. Наиболее близким аналогом предложенного здесь критерия может служить критерий Шапиро—Чена [117], статистика которого основана на сумме спейсингов, нормированных выборочным среднеквадратическим отклонением, с весовыми коэффициентами, выбираемыми специальным образом. Идеи использования минимального спейсинга для построения критерия проверки нормальности в литературе найти не удалось.

В данном параграфе доказано, что в широких предположениях при $n = 3$ статистика, инвариантная относительно перестановки аргументов, преобразований сдвига и масштаба, является функцией от d_3. Вычислена функция распределения статистики d_3. Для этого разности между компонентами выборки выражены через компоненты стандартного двумерного нормального распределения, и искомая формула получена сведением к равномерному распределению на окружности. Рассмотрен случай $n = 4$. Распределение статистики d_4 вычислено с использованием формулы Люилье (l'Huillier)

площади сферического треугольника. Получены верхняя и нижняя оценки для хвоста распределения d_n при $n > 3$. Для этого доказана лемма о рапределении частного модулей компонент двумерного центрированного нормального вектора. Показано, что границы для хвоста распределения убывают по закону обратной пропорциональности.

Приведены результаты моделирования, подсчет квантили уровня 0,95 для $n = 5$ и эмпирической мощности критерия на альтернативах. Кроме того, проведено сравнение с критерием Шапиро—Уилка. Квантили уровня 0,05 для этого критерия исправлены по результатам моделирования. Анализ результатов показывает, что предложенный критерий мощнее критерия Шапиро—Уилка при $n = 4,\ 5$ и резко асимметричных альтернативах. При $n = 3$ мощности критериев совпадают, но для предложенного критерия существует простая явная формула для квантилей.

Если выполнена основная гипотеза, то $d_2 = 1$ п.н. Рассмотрим статистику d_3.

$$d_3 = \frac{\max\{|X_1 - X_2|, |X_1 - X_3|, |X_2 - X_3|\}}{\min\{|X_1 - X_2|, |X_1 - X_3|, |X_2 - X_3|\}}.$$

В случае выборки объема 3 любая статистика, не определенная при $R_3 = 0$, симметричная относительно элементов выборки и инвариантная относительно преобразований сдвига и масштаба, является функцией от d_3. Докажем это.

Теорема 3.1 *Если на каждом элементарном исходе $Z(X_1, X_2, X_3)$ — функция, не определенная при $R_3 = 0$, инвариантная относительно перестановки аргументов, и для любых $b \in \mathbf{R}$, $c > 0$ выполнено $Z(b+cX_1, b+cX_2, b+cX_3) = Z(X_1, X_2, X_3)$, то Z является функцией переменной d_3, где $d_3 \in [2,\ \infty]$.*

Доказательство

На данном элементарном исходе переставим аргументы таким образом, чтобы $X_1 \leq X_2 \leq X_3$, $X_2 - X_1 \leq X_3 - X_2$. Тогда $L_3 = X_2 - X_1$, $R_3 = X_3 - X_1$,

$$Z(X_1, X_2, X_3) = Z(X_1/R_3, X_2/R_3, X_3/R_3)$$
$$= Z(0, (X_2 - X_1)/R_3, (X_3 - X_1)/R_3)$$
$$= Z(0, L_3/R_3, 1) = Z(0, 1/d_3, 1).$$

Случай $L_3 = 0$ соответствует значению $d_3 = +\infty$ согласно принятому соглашению.

Теорема доказана.

Отметим, что статистика

$$\frac{{R_3}^2}{\sum_{i=1}^3 (X_i - \overline{X})^2},$$

возникающая при $n = 3$ в критерии Шапиро—Уилка, удовлетворяет условиям теоремы, и в силу равенства

$$\sum_{i=1}^3 (X_i - \overline{X})^2$$
$$= \frac{1}{9}(((X_1 - X_2) + (X_1 - X_3))^2$$
$$+((X_2 - X_1) + (X_2 - X_3))^2 + ((X_3 - X_1) + (X_3 - X_2))^2)$$
$$= \frac{1}{9}\left((L_3 + R_3)^2 + (2L_3 - R_3)^2 + (2R_3 - L_3)^2\right)$$
$$= \frac{2({3L_3}^2 + {3R_3}^2 + R_3 L_3)}{9},$$

представима в виде

$$\frac{9{R_3}^2}{2({3L_3}^2 + {3R_3}^2 + R_3 L_3)} = \frac{9}{2(3d_3^{-2} + 3 + d_3^{-1})}.$$

Распределение этой статистики может быть вычислено в явном виде с помощью распределения статистики d_3.

Вычислим функцию распределения статистики d_3.

$$F_{d_3}(x) = \mathbf{P}\left\{\forall i \neq j \;\; \frac{\max_{1\leq k,l\leq n}|X_k - X_l|}{|X_i - X_j|} < x\right\}$$

$$= 3!\mathbf{P}\left\{X_1 < X_2 < X_3,\; \frac{X_3 - X_1}{X_2 - X_1} < x,\; \frac{X_3 - X_1}{X_3 - X_2} < x\right\}$$

$$= 6\mathbf{P}\{X_2 - X_1 > 0,\; X_3 - X_2 > 0,$$

$$x(X_2 - X_1) > X_3 - X_1,\; x(X_3 - X_2) > X_3 - X_1\}.$$

Обозначим $Y_{ij} = X_i - X_j$. Тогда $\mathbf{E}Y_{ij} = 0,\;\; \mathbf{D}Y_{ij} = 2\sigma^2$ при $i \neq j$, где σ^2 — дисперсия элементов выборки.

$$F_{d_3}(x) = 6\mathbf{P}\{Y_{21} > 0,\; Y_{32} > 0,\; xY_{21} > Y_{31},\; xY_{32} > Y_{31}\} =$$

$$= 1 - 2 \cdot 6\mathbf{P}\{Y_{21} > 0,\; xY_{21} < Y_{31}\}.$$

Так как коэффициенты корреляции равны

$$\mathbf{corr}(Y_{21};\; Y_{32}) = -\mathbf{corr}(Y_{21};\; Y_{31}) = -\mathbf{corr}(Y_{32};\; Y_{31}) = -\frac{1}{2},$$

то компоненты нормального случайного вектора $(Y_{21},\; Y_{32},\; Y_{31})$ можно задать через компоненты стандартного нормального случайного вектора $(\eta_1,\; \eta_2)$ следующим образом:

$$Y_{21} = \sigma\sqrt{2}\eta_1;\; Y_{32} = \sigma\sqrt{2}\left(-\frac{1}{2}\eta_1 + \frac{\sqrt{3}}{2}\eta_2\right);\; Y_{31} = \sigma\sqrt{2}\left(\frac{1}{2}\eta_1 + \frac{\sqrt{3}}{2}\eta_2\right).$$

Следовательно,

$$\mathbf{P}\{d_3 < x\} = 1 - 12\mathbf{P}\{\eta_1 > 0;\; x\eta_1 < \frac{1}{2}\eta_1 + \frac{\sqrt{3}}{2}\eta_2\}$$

$$= 1 - 12\mathbf{P}\{\eta_1 > 0;\; \eta_2 > \frac{2x-1}{\sqrt{3}}\eta_1\}$$

$$= 1 - \frac{12}{2\pi}\mathrm{arcctg}\frac{2x-1}{\sqrt{3}}.$$

Итак, доказана

Теорема 3.2 *Для всех* $x \geq 2$ *выполнено*

$$F_{d_3}(x) = 1 - \frac{6}{\pi}\mathrm{arcctg}\frac{2x-1}{\sqrt{3}}.$$

В качестве следствия отметим, что $\mathbf{P}\{d_3 \geq x\} \sim \frac{3\sqrt{3}}{\pi x}$ при $x \to \infty$.

Рассмотрим случай $n = 4$. Выполнено $d_4 > 3$ п.н. Для $x > 3$

$$\mathbf{P}\{d_4 < x\} = 4! \cdot \mathbf{P}\{X_4 - X_1 > 0,\ x(X_2 - X_1) > X_4 - X_1,$$

$$x(X_3 - X_2) > X_4 - X_1,\ x(X_4 - X_3) > X_4 - X_1\}$$

$$= \mathbf{P}\{d_4 < x\} = 4! \cdot \mathbf{P}\{X_4 - X_1 > 0,\ x(X_2 - X_1) > X_4 - X_1,$$

$$x(X_3 - X_2) > X_4 - X_1,\ x(X_4 - X_3) > X_4 - X_1\}$$

$$= 24\mathbf{P}\{Y_{21} + Y_{32} + Y_{43} > 0,\ xY_{21} > Y_{21} + Y_{32} + Y_{43},$$

$$xY_{32} > Y_{21} + Y_{32} + Y_{43},\ xY_{43} > Y_{21} + Y_{32} + Y_{43}\}.$$

Просуммировав последние 3 неравенства, получаем, что

$$(x-3)(Y_{21} + Y_{32} + Y_{43}) > 0,$$

откуда следует первое неравенство, так как $x > 3$.

Обозначим $Z_{ij} = Y_{ij}/(\sigma\sqrt{2})$. Тогда

$$\mathbf{P}\{d_4 < x\} = 24\mathbf{P}\{xZ_{21} > Z_{21} + Z_{32} + Z_{43},$$

$$xZ_{32} > Z_{21} + Z_{32} + Z_{43},\ xZ_{43} > Z_{21} + Z_{32} + Z_{43}\}. \qquad (27)$$

Найдем вероятность выполнения системы неравенств

$$\begin{cases} xZ_{21} > Z_{21} + Z_{32} + Z_{43}; \\ xZ_{32} > Z_{21} + Z_{32} + Z_{43}; \\ xZ_{43} > Z_{21} + Z_{32} + Z_{43}. \end{cases} \qquad (28)$$

Корреляционная матрица вектора $(Z_{21};\ Z_{32};\ Z_{43})$ имеет вид:

$$\begin{pmatrix} 1 & -1/2 & 0 \\ -1/2 & 1 & -1/2 \\ 0 & -1/2 & 1 \end{pmatrix}.$$

Удобно представить компоненты вектора следующим симметричным образом:

$$Z_{32} = \eta_2;$$
$$Z_{21} = -\frac{1}{2}\eta_2 + \frac{\sqrt{2}+1}{2\sqrt{2}}\eta_1 - \frac{\sqrt{2}-1}{2\sqrt{2}}\eta_3;$$
$$Z_{43} = -\frac{1}{2}\eta_2 + \frac{\sqrt{2}+1}{2\sqrt{2}}\eta_3 - \frac{\sqrt{2}-1}{2\sqrt{2}}\eta_1.$$

Здесь вектор $(\eta_1,\ \eta_2,\ \eta_3)$ имеет стандартное нормальное распределение.

Так как $Z_{21} + Z_{32} + Z_{43} = \frac{1}{\sqrt{2}}\eta_1 + \frac{1}{\sqrt{2}}\eta_3$, то система неравенств (28) записывается в виде

$$\begin{cases} x\left(-\frac{1}{2}\eta_2 + \frac{\sqrt{2}+1}{2\sqrt{2}}\eta_1 - \frac{\sqrt{2}-1}{2\sqrt{2}}\eta_3\right) > \frac{1}{\sqrt{2}}\eta_1 + \frac{1}{\sqrt{2}}\eta_3; \\ x\eta_2 > \frac{1}{\sqrt{2}}\eta_1 + \frac{1}{\sqrt{2}}\eta_3; \\ x\left(-\frac{1}{2}\eta_2 + \frac{\sqrt{2}+1}{2\sqrt{2}}\eta_3 - \frac{\sqrt{2}-1}{2\sqrt{2}}\eta_1\right) > \frac{1}{\sqrt{2}}\eta_1 + \frac{1}{\sqrt{2}}\eta_3. \end{cases} \tag{29}$$

В силу симметрии многомерного нормального распределения, вероятность выполнения системы неравенств (29) равна отношению площади соответствующего сферического треугольника к площади единичной сферы. Грани сферического треугольника образованы плоскостями

$$\begin{cases} x\left(-\frac{1}{2}t_2 + \frac{\sqrt{2}+1}{2\sqrt{2}}t_1 - \frac{\sqrt{2}-1}{2\sqrt{2}}t_3\right) = \frac{1}{\sqrt{2}}t_1 + \frac{1}{\sqrt{2}}t_3; \\ xt_2 = \frac{1}{\sqrt{2}}t_1 + \frac{1}{\sqrt{2}}t_3; \\ x\left(-\frac{1}{2}t_2 + \frac{\sqrt{2}+1}{2\sqrt{2}}t_3 - \frac{\sqrt{2}-1}{2\sqrt{2}}t_1\right) = \frac{1}{\sqrt{2}}t_1 + \frac{1}{\sqrt{2}}t_3. \end{cases}$$

Нормальные векторы этих плоскостей:

$$\mathbf{u}_1 = \left(\frac{(\sqrt{2}+1)x - 2}{2\sqrt{2}},\ -\frac{x}{2},\ -\frac{(\sqrt{2}-1)x + 2}{2\sqrt{2}}\right);$$
$$\mathbf{u}_2 = \left(-\frac{1}{\sqrt{2}}, x,\ -\frac{1}{\sqrt{2}}\right);$$
$$\mathbf{u}_3 = \left(-\frac{(\sqrt{2}-1)x + 2}{2\sqrt{2}},\ -\frac{x}{2},\ \frac{(\sqrt{2}+1)x - 2}{2\sqrt{2}}\right).$$

Найдем направляющие векторы прямых, по которым пересекаются плоскости, с помощью векторного произведения:

$$\mathbf{v}_{12} = \mathbf{u}_1 \times \mathbf{u}_2 = \left(-(\sqrt{2}-1)x^2 - 3x,\ -2x,\ 3x - (\sqrt{2}+1)x^2\right);$$
$$\mathbf{v}_{13} = \mathbf{u}_1 \times \mathbf{u}_3 = \left((\sqrt{2}+1)x^2 - 3x,\ 2x,\ (\sqrt{2}-1)x^2 + 3x\right);$$
$$\mathbf{v}_{23} = \mathbf{u}_2 \times \mathbf{u}_3 = \left(-4x^2,\ -4\sqrt{2}x(x-2),\ -4x^2\right).$$

Используем формулу Люилье ([159], §36) для площади ε сферического треугольника со сторонами $a,\ b,\ c$:

$$\varepsilon = 4\mathrm{arctg}\sqrt{\mathrm{tg}\frac{s}{2}\mathrm{tg}\frac{s-a}{2}\mathrm{tg}\frac{s-b}{2}\mathrm{tg}\frac{s-c}{2}},$$

где $s = (a+b+c)/2$.

Сторонами сферического треугольника называются углы между прямыми, образующими трехгранный угол. Поэтому

$$a = c = \arccos\frac{|\mathbf{v}_{12}\mathbf{v}_{13}|}{|\mathbf{v}_{12}||\mathbf{v}_{13}|} = \arccos\frac{x^2+6x-7}{3x^2-6x+11};$$

$$b = \arccos\frac{|\mathbf{v}_{12}\mathbf{v}_{23}|}{|\mathbf{v}_{12}||\mathbf{v}_{23}|} = \arccos\frac{x^2+x-2}{\sqrt{3x^2-6x+11}\sqrt{x^2-2x+2}}.$$

Согласно формуле (27), умножим результат вычисления по формуле Люилье на 24 и разделим на площадь единичной сферы 4π. Получаем следующую теорему.

Теорема 3.3 *При* $x \geq 3$

$$F_{d_4}(x) = \frac{24}{\pi}\mathrm{arctg}\left(\mathrm{tg}\frac{a}{4}\sqrt{\mathrm{tg}\left(\frac{b}{2}+\frac{a}{4}\right)\mathrm{tg}\left(\frac{b}{2}-\frac{a}{4}\right)}\ \right),$$

где

$$a = \arccos\frac{x^2+6x-7}{3x^2-6x+11},$$
$$b = \arccos\frac{x^2+x-2}{\sqrt{3x^2-6x+11}\sqrt{x^2-2x+2}}.$$

Отметим, что осуществить вычисление распределения d_5 не удается ввиду отсутствия общей формулы, выражающей объем тетраэдра в пространстве постоянной кривизны через длины его ребер [138].

Выполним сравнение с критерием Шапиро—Уилка. Будем моделировать выборки заданного объема из нормального распределения, основываясь на известном алгоритме

$$X_i = \sum_{j=1}^{12} U_{12(i-1)+j} - 6,$$

где $U_1,\ U_2,\ \ldots$ — независимые случайные величины с равномерным распределением на $[0,\ 1]$. Для их моделирования использовался встроенный датчик случайных чисел Excel. Квантили уровня 0,95 для d_3 и d_4 отыскиваются из теорем 2 и 3. Квантиль уровня 0,95 для d_5 отыскивается как выборочная квантиль по результатам моделирования $2 \cdot 10^6$ выборок и приближенно равняется 214.

Рассмотрим мощность критерия на альтернативах. Обозначим $\mathcal{LN}_\sigma$ логнормальное распределение с параметрами 0, σ^2; $\mathcal{C}$ — распределение Коши.

Для логнормального распределения моделирование осуществляется по формуле

$$X_i = \exp\left(\sigma\left(\sum_{j=1}^{12} U_{12(i-1)+j} - 6\right)\right),$$

а для распределения Коши

$$X_i = \tan(\pi(U_i - 1/2)).$$

В таблице 3.1 приведены эмпирические значения мощности R/L-критерия на альтернативах, полученные по результатам моделирования 10^6 выборок. Здесь n — объем выборки, $t_{0,95}$ — квантиль уровня 0,95 (для $n = 5$ подсчитана эмпирически).

Таблица 3.1 Эмпирическая мощность R/L—критерия.

n	$t_{0,95}$	$\mathcal{LN}_1$	$\mathcal{LN}_2$	$\mathcal{LN}_5$	$\mathcal{LN}_{10}$	$\mathcal{LN}_{20}$	$\mathcal{C}$
3	33,57	0,08	0,20	0,53	0,73	0,86	0,10
4	103,60	0,09	0,25	0,72	0,91	0,98	0,12
5	214	0,09	0,29	0,83	0,97	0,995	0,13

Мощность критерия при логнормальной альтернативе существенно зависит от параметра σ. Наилучшие результаты получаются при $\sigma \geq 10$.

Проведем сравнение с критерием Шапиро—Уилка [157]. Отметим, что статья [157] содержит досадную неточность относительно значений квантилей уровня 0,05 (табл. 6, с. 605). Поэтому пришлось пересчитать квантили, основываясь на моделировании 10^6 выборок заданного объема. Квантиль исправлялась таким образом, чтобы эмпирический критический уровень был наиболее близок к 0,05. Результаты вычислений приведены в таблице 3.2. Через $X_{(i)}$ обозначены порядковые статистики, через $\tilde{q}_{0,05}$ — квантиль уровня 0,05 в табл. 6 из [157], через $\tilde{\varepsilon}$ — эмпирический уровень, достигнутый для этой квантили по результатам моделирования, через $q_{0,05}$ — квантиль уровня 0,05, исправленная по результатам моделирования.

Таблица 3.2 Эмпирические квантили для критерия Шапиро—Уилка.

n	Статистика b	$\tilde{q}_{0,05}$	$\tilde{\varepsilon}$	$q_{0,05}$
3	$0,7071(X_{(3)} - X_{(1)})$	0,767	0,038	0,772
4	$0,6872(X_{(4)} - X_{(1)}) + 0,1677(X_{(3)} - X_{(2)})$	0,748	0,038	0,761
5	$0,6646(X_{(5)} - X_{(1)}) + 0,2413(X_{(4)} - X_{(2)})$	0,762	0,037	0,777

Используем исправленные значения для подсчета эмпирической мощности критерия Шапиро—Уилка на альтернативах. Напомним, что критерий отвергает гипотезу о нормальности на уровне 0,05, если значение $W = b^2 / \sum_{i=1}^{n}(X_i - \overline{X})^2$ окажется меньше, чем $q_{0,05}$.

Результаты приведены в таблице 3.3.

Таблица 3.3 Эмпирическая мощность критерия Шапиро—Уилка.

n	$q_{0,05}$	$\mathcal{LN}_1$	$\mathcal{LN}_4$	$\mathcal{LN}_{25}$	$\mathcal{LN}_{100}$	$\mathcal{LN}_{400}$	$\mathcal{C}$
3	0,772	0,08	0,20	0,53	0,73	0,86	0,10
4	0,761	0,17	0,38	0,70	0,84	0,92	0,21
5	0,777	0,24	0,52	0,85	0,95	0,988	0,29

Сравнение табл. 3.1 и 3.3 приводит к выводу о том, что R/L—критерий оказывается полезным при $n = 4,\ 5$ в случае существенной асимметрии распределения при альтернативной гипотезе (в частности, для логнормального распределения при $\sigma^2 \geq 100$). При $n = 3$ мощности критериев совпадают. Это совпадение подтверждает результаты теоремы 3.1 и следующего за ним замечания.

3.3 Алгоритм тестирования бинарным знаковым методом

На основании оценки бинарным знаковым методом получаем следующий алгоритм проверки основной гипотезы о соответствии данных модели выборки против альтернативной гипотезы о соответствии модели фрактального гауссовского шума с параметром $H \neq 1/2$.

Алгоритм оказывается технически сложным ввиду того, что n может не являться целой степенью числа 2.

Алгоритм 3.1 *Алгоритм проверки гипотезы о фрактальном гауссовском шуме по последовательности значений* $X_1, \ldots, X_n$.

1. *Центрировать данные* $X_i^* = X_i - \overline{X}$.
2. *Ввести обозначения для сумм* $S_i^* := X_i^*$, $i = 1, \ldots, n$.
3. *Вычислить число перемен знака* $m^*(k)$, *число слагаемых* $n^*(k)$ k, *просуммировать соседние суммы (процесс продолжается до*

тех пор, пока не останется две суммы; в этом предельном случае число слагаемых — индикаторов перемены знака — равно единице):

FOR $k = 0, \ldots, [\log_2 n] - 2$

BEGIN

$m^*(k) = \sum_{i=1}^{[n2^{-k}]-1} \mathbf{I}\{S_i^* S_{i+1}^* < 0\}$;

$n^*(k) = [n2^{-k}] - 1$;

$S_1^* := S_1^* + S_2^*$;

$\ldots$;

$S_{[n2^{-k}]}^* := S_{[n2^{-k+1}]-1}^* + S_{[n2^{-k+1}]}^*$

END.

4. *Вычислить общее число перемен знака* m^* *и общее число слагаемых* n^**:*

   $m^* = \sum_{k=0}^{[\log_2 n]-2} m^*(k)$;

   $n^* = \sum_{k=0}^{[\log_2 n]-2} n^*(k)$.

5. *Вычислить частоту перемены знака и оценку параметра* H*:*

   $\nu^{**} = m^*/n^*$;

   $H^{**} = \frac{1}{2} + \frac{1}{2}\log_2\left(1 + \cos(\pi\nu_n^{**})\right)$.

6. *Вычислить достигнутый уровень значимости:*

   $\varepsilon^{**} = 2\overline{\Phi}(\sqrt{n}|H^{**} - 1/2|/B)$,

   $B = \sqrt{\frac{\pi^2}{4\ln^2 2}\left(\frac{1}{8} + \sum_{s=1}^{\infty} 2^{-s}\left(\frac{1}{\pi}\arccos 2^{-s/2} - \frac{1}{2}\right)^2\right)}$,

   $\overline{\Phi}(x) = \frac{1}{\sqrt{2\pi}}\int_x^{+\infty} e^{-t^2/2}dt$.

7. *Гипотеза о некоррелированности принимается на уровне* ε*, если* $\varepsilon^{**} \geq \varepsilon$*; отвергается на уровне* ε*, если* $\varepsilon^{**} < \varepsilon$.

3.4 Моделирование фрактального гауссовского шума

Для сравнения вышеприведенной оценки бинарным знаковым методом и оценки максимального правдоподобия параметра H необходимо научиться моделировать фрактальный гауссовский шум. Моделирование будет применяться и в следующих параграфах для вычисления характеристик (смещения и дисперсии) разности оценок.

Рассмотрим два алгоритма моделирования: точный, но трудоемкий алгоритм разложения корреляционной матрицы на нижнюю и верхнюю треугольные применяется при $n = 1024$.

Алгоритм, использующий C–сходимость случайных ломаных, построенных по скользящим средним, используется при генерировании данных большего объема.

Алгоритм, основанный на разложении ковариационной матрицы

Алгоритм 3.2

Генерирование вектора $\mathbf{W} = (W_1, \ldots, W_n)^T$ *независимых случайных величин со стандартным нормальным распределением.*

Преобразование вектора $\mathbf{W}$ *во фрактальный гауссовский шум с параметром* H, $0 < H < 1$ *на основе разложения корреляционной матрицы [130]: матрицу*

$$R = (r(i-j))_{i,j=1}^{n},$$

где

$$r(k) = \mathbf{corr}(X_i;\ X_{i+k}) = \frac{1}{2}\left(|k+1|^{2H} + |k-1|^{2H} - 2|k|^{2H}\right),$$

подвергаем декомпозиции по Холецкому к виду $R = AA^T$, *где* A *— нижняя треугольная матрица. Далее формируем вектор фрактального гауссовского шума по схеме* $\mathbf{X} = A\mathbf{W}$.

Действительно, корреляционная матрица этого вектора совпадает с R:

$$\mathbf{E}(\mathbf{X}\mathbf{X}^{\mathbf{T}}) = \mathbf{E}(A\mathbf{W}\mathbf{W}^{\mathbf{T}}A^T) = R.$$

Для корректной работы алгоритма необходима положительность собственных чисел матрицы. Она была доказана в работе Dietrich, Newsam [120] для $H > 1/2$ и в работе Craigmile [118] для $H < 1/2$.

В ряде работ (по существу начиная с работы Мандельброта и Ван Несса [144]; подробное описание алгоритма и результаты моделирования приводятся в статье А. П. Ковалевского и Н. С. Закревской [178]) используется метод скользящего среднего для моделирования фрактального броуновского движения. Этот метод является асимптотическим, то есть фрактальное броуновское движение появляется лишь в пределе. В работе Ю. А. Давыдова [23] доказана сходимость метода. Оценки скорости сходимости найдены в работах Т. Константопоулоса и А. И. Саханенко[134], Н. С. Аркашова и И. С. Борисова [4].

Скользящие средние строятся следующим образом:

$$X_j = \sum_{k=-\infty}^{\infty} a_{j-k}\xi_k, \tag{30}$$

где ξ_k — независимые одинаково распределенные случайные величины, относиельно которых будем дополнительно предполагать $\mathbf{E}\xi_k = 0$, $\mathbf{E}\xi_k^2 = 1$.

Зададим коэффициенты так, чтобы дисперсия частичных сумм возрастала по асимптотически степенному закону. Наиболее простой и точный способ достижения этого состоит в следующем:

$$a_k = L_H^{-1/2}((k+1)^{H-1/2} - k^{H-1/2}), \quad k \geq 0; \tag{31}$$

$$a_k = 0, \quad k < 0.$$

Здесь L_H — константа, равная

$$L_H = \frac{1}{2H} + \int_0^\infty ((s+1)^{H-1/2} - s^{H-1/2})^2 ds. \tag{32}$$

Константа L_H может быть выражена через гамма-функцию Эйлера (см. работу Норроса и др. [151]):

$$L_H = \sqrt{\frac{2H\Gamma(3/2-H)}{\Gamma(H+1/2)\Gamma(2-2H)}}.$$

Так как $a_k = 0$ при $k < 0$, получаем формулу

$$X_j = \sum_{k=-\infty}^{j} a_{j-k}\xi_k.$$

На практике удобно, во-первых, перенумеровать индексы, задав X_j в виде

$$X_j = \sum_{k=-j}^{\infty} a_{j+k}\zeta_k, \tag{33}$$

$\zeta_k = \xi_{-k}$.

Во-вторых, необходимо исключить бесконечное суммирование — ограничить число слагаемых, то есть вместо (33) задать новые случайные величины X_j^* в виде

$$X_j^* = \sum_{k=-j}^{M} a_{j+k}\zeta_k, \tag{34}$$

где M — конечное число, выбираемое в зависимости от n и H.

Обозначим $S_n^* = X_1^* + X_n^*$. Докажем лемму, позволяющую выбрать константу M так, чтобы относительное изменение дисперсии при замене (33) на (34) было невелико.

Лемма 3.1 *Пусть* $H \neq 1/2$, $\varepsilon > 0$,

$$M = \left[n \left(\frac{(H-1/2)^2}{L_H^2 \varepsilon (2-2H)} \right)^{\frac{1}{2-2H}} \right].$$

Тогда

$$\lim_{n\to\infty} \frac{\mathbf{D}(S_n - S_n^*)}{\mathbf{D}S_n} \leq \varepsilon.$$

Доказательство.

Заметим, что

$$S_n - S_n^* = \sum_{j=1}^{n} \sum_{k=M+1}^{\infty} a_{j+k}\zeta_k =$$

$$= \sum_{k=M+1}^{\infty} \sum_{j=1}^{n} a_{j+k}\zeta_k.$$

Перестановка порядка суммирования здесь возможна в силу сходимости с вероятностью 1 внутренних рядов в соответствии с теоремой о трех рядах.

Так как

$$\sum j = 1^n a_{j+k} = \frac{1}{L_H}\left((n+k+1)^{H-1/2} - (k+1)^{H-1/2}\right),$$

$$\mathbf{D}\zeta_k = 1,$$

то

$$\mathbf{D}(S_n - S_n^*) = \frac{1}{L_H^2} \sum_{k=M+1}^{\infty} \left((n+k+1)^{H-1/2} - (k+1)^{H-1/2}\right)^2.$$

Пусть $M+1 = cn$. Тогда

$$\lim_{n\to\infty} \frac{\mathbf{D}(S_n - S_n^*)}{\mathbf{D}S_n} =$$

$$= \lim_{n\to\infty} \frac{\sum_{k=cn}^{\infty} \left((n+k+1)^{H-1/2} - (k+1)^{H-1/2}\right)^2}{L_H^2 n^{2H}} =$$

$$= \frac{1}{L_H^2} \lim_{n\to\infty} \sum_{k=cn}^{\infty} \frac{1}{n} \left((1 + \frac{k}{n} + \frac{1}{n})^{H-1/2} - (\frac{k}{n} + \frac{1}{n})^{H-1/2}\right)^2 =$$

$$= \frac{1}{L_H^2} \int_c^{\infty} \left((1+x)^{H-1/2} - x^{H-1/2}\right)^2 dx.$$

В силу неравенства

$$\left((1+x)^{H-1/2} - x^{H-1/2}\right)^2 \le \left((H-1/2)x^{H-3/2}\right)^2,$$

получаем

$$\lim_{n\to\infty}\frac{\mathbf{D}(S_n-S_n^*)}{\mathbf{D}S_n}\le\frac{(H-1/2)^2}{L_H^2}\int_c^\infty x^{2H-3}dx=$$

$$=\frac{(H-1/2)^2c^{2H-2}}{L_H^2(2H-2)}.$$

Для выполнения условия

$$\lim_{n\to\infty}\frac{\mathbf{D}(S_n-S_n^*)}{\mathbf{D}S_n}\le\varepsilon$$

достаточно

$$c\ge\left(\frac{(H-1/2)^2}{L_H^2\varepsilon(2-2H)}\right)^{\frac{1}{2-2H}}.$$

Следовательно, можно взять

$$M=\left[n\left(\frac{(H-1/2)^2}{L_H^2\varepsilon(2-2H)}\right)^{\frac{1}{2-2H}}\right].$$

Доказательство завершено.

Для целей моделирования примем $\varepsilon=0,01$.

Алгоритм моделирования
фрактального броуновского движения

Алгоритм 3.3 *1. Для заданных $H\neq 1/2$, $n>1$ вычислим константы*

$$L_H=\sqrt{\frac{2H\Gamma(3/2-H)}{\Gamma(H+1/2)\Gamma(2-2H)}},$$

$$M=\left[n\left(\frac{(H-1/2)^2}{L_H^2\varepsilon(2-2H)}\right)^{\frac{1}{2-2H}}\right].$$

2. Генерируем случайные величины ζ_k, $k=-n,\dots,M$. Эти случайные величины принимают значения 1 и -1 с вероятностями по $1/2$ и независимы в совокупности.

3. Найдем значения $X_j^* = \sum_{k=-j}^{M} a_{j+k}\zeta_k, \;\; j = 1, \ldots, n.$

4. Просуммируем найденные значения: $S_n^* = X_1^* + \ldots + X_n^*$.

В качестве приложения процедуры моделирования проверим соответствие среднеквадратического отклонения оценки, получаемой элементарным знаковым методом, результатам моделирования.

Теоретическое значение стандартного отклонения

$$\sigma = \frac{\sigma_{1/2}}{\sqrt{n}} = \frac{2\pi}{\sqrt{n}\ln 2}.$$

Результаты для модельных данных подсчитаны по 10 000 независимым экспериментам. Получаем следующую таблицу, из которой видно согласие оценки с теоретическим значением.

Табл. 3.1. Теоретические и выборочные значения стандартных отклонений оценки параметра H по методу знаков при $H = 0,5$.

n	σ	$\tilde{\sigma}$
16384	0,00885	0,00889
8192	0,01251	0,01254
4096	0,01770	0,01783
2048	0,02503	0,02528
1024	0,03540	0,03599
512	0,05007	0,05080

3.5 Критерии разладки фрактального шума

Построим класс критериев для проверки гипотезы о том, что исследуемые данные — фрактальный гауссовский шум, против альтернативной гипотезы о том, что происходит разладка (изменение параметров) фрактального гауссовского шума. Отметим, что здесь

результаты предыдущего параграфа уже не годятся, так как при соновной гипотезе $H \neq 1/2$.

Поэтому критерии будут основываться на разностях оценок параметра H, полученных разными методами. Напомним, что оценки методом нормированного размаха $\tilde{H}_1$ и методом дисперсии $\tilde{H}_2$ ответственны за глобальное поведение процесса, а оценка $\tilde{H}_3$ — за локальную характеристику — частоту перемены знака приращениями процесса. Используя результаты моделирования и разработанные процедуры оценивания параметров, получаем средние значения и стандартные отклонения разностей оценок параметра H разными методами. На основании этих величин построим критерии проверки гипотезы о соответствии выборки модели фрактального гауссовского шума.

Каждая из приведенных ниже оценок вычислена по 10 000 независимых экспериментов, в ходе каждого из которых моделировалась последовательность длины $2^{14} = 16384$. Элементы последовательности суммировались по τ штук. Результаты получены совместно с Н. С. Закревской и опубликованы в работе [178].

Приведенные ниже таблицы будут использоваться критерием разладки фрактального шума по разности оценок.

Табл. 3.2. Выборочные разности оценок параметра H при $H = 0,1$.

$H = 0,1$	$\tilde{H}_2 - \tilde{H}_3$	$\tilde{H}_1 - \tilde{H}_2$	$\tilde{H}_1 - \tilde{H}_3$
$\tau = 1$			
среднее	-0,17703422	0,17923102	0,0021968
станд. откл.	0,036489233	0,032609995	0,014616065
$\tau = 2$			
среднее	-0,14864335	0,19210632	0,04346297
станд. откл.	0,042055216	0,037852771	0,021427846
$\tau = 4$			
среднее	-0,14864335	0,19210632	0,04346297
станд. откл.	0,042055216	0,037852771	0,021427846
$\tau = 8$			
среднее	-0,09906138	0,22349412	0,12443274
станд. откл.	0,062656741	0,051872978	0,033636814
$\tau = 16$			
среднее	-0,08035762	0,24383414	0,16347652
станд. откл.	0,071254242	0,060207322	0,049254307
$\tau = 32$			
среднее	-0,07919516	0,26429596	0,1851008
станд. откл.	0,097946017	0,067610298	0,071115158

Табл. 3.3. Выборочные разности оценок параметра H при $H = 0,2$.

$H = 0,2$	$\tilde{H}_2 - \tilde{H}_3$	$\tilde{H}_1 - \tilde{H}_2$	$\tilde{H}_1 - \tilde{H}_3$
$\tau = 1$			
среднее	-0,14237918	0,15727741	0,01489823
станд. откл.	0,035246687	0,030459726	0,016603262
$\tau = 2$			
среднее	-0,11895032	0,1689961	0,05004578
станд. откл.	0,041892942	0,035137924	0,021686848
$\tau = 4$			
среднее	-0,11895032	0,1689961	0,05004578
станд. откл.	0,041892942	0,035137924	0,021686848
$\tau = 8$			
среднее	-0,08218572	0,19852188	0,11633616
станд. откл.	0,061262276	0,048596221	0,033206582
$\tau = 16$			
среднее	-0,07887883	0,21794118	0,13906235
станд. откл.	0,08624373	0,058539551	0,049795107
$\tau = 32$			
среднее	-0,07418146	0,24037828	0,16619682
станд. откл.	0,109714767	0,071501567	0,067484294

Табл. 3.4. Выборочные разности оценок параметра H при $H = 0,3$.

$H = 0,3$	$\tilde{H}_2 - \tilde{H}_3$	$\tilde{H}_1 - \tilde{H}_2$	$\tilde{H}_1 - \tilde{H}_3$
$\tau = 1$			
среднее	-0,1014374	0,12876271	0,02732531
станд. откл.	0,036165723	0,030692642	0,019416744
$\tau = 2$			
среднее	-0,08433383	0,13870339	0,05436956
станд. откл.	0,042525024	0,03525273	0,024635365
$\tau = 4$			
среднее	-0,08433383	0,13870339	0,05436956
станд. откл.	0,042525024	0,03525273	0,024635365
$\tau = 8$			
среднее	-0,06173435	0,16498069	0,10324634
станд. откл.	0,062533829	0,047858462	0,03905053
$\tau = 16$			
среднее	-0,05050501	0,18161748	0,13111247
станд. откл.	0,071071564	0,056407248	0,049454837
$\tau = 32$			
среднее	-0,05364633	0,20210082	0,14845449
станд. откл.	0,097307689	0,067983405	0,060773401

Табл. 3.5. Выборочные разности оценок параметра H при $H = 0,4$.

$H = 0,4$	$\tilde{H}_2 - \tilde{H}_3$	$\tilde{H}_1 - \tilde{H}_2$	$\tilde{H}_1 - \tilde{H}_3$
$\tau = 1$			
среднее	-0,063201	0,10956313	0,04636213
станд. откл.	0,035557578	0,026836393	0,021746884
$\tau = 2$			
среднее	-0,05736817	0,11898748	0,06161931
станд. откл.	0,041322667	0,030786063	0,025036643
$\tau = 4$			
среднее	-0,05736817	0,11898748	0,06161931
станд. откл.	0,041322667	0,030786063	0,025036643
$\tau = 8$			
среднее	-0,05579668	0,14387893	0,08808225
станд. откл.	0,053989735	0,041974402	0,036799524
$\tau = 16$			
среднее	-0,05119891	0,16014747	0,10894856
станд. откл.	0,068387095	0,050243817	0,047463248
$\tau = 32$			
среднее	-0,04947055	0,18048762	0,13101707
станд. откл.	0,094631286	0,061540765	0,062477004

Табл. 3.6. Выборочные разности оценок параметра H при $H = 0,5$.

$H = 0,5$	$\tilde{H}_2 - \tilde{H}_3$	$\tilde{H}_1 - \tilde{H}_2$	$\tilde{H}_1 - \tilde{H}_3$
$\tau = 1$			
среднее	-0.023032	0.084127	0.061095
станд. откл.	0.037376	0.026095	0.024822
$\tau = 2$			
среднее	-0.027875	0.092420	0.064545
станд. откл.	0.043872	0.030157	0.028647
$\tau = 4$			
среднее	-0.027875	0.092420	0.064545
станд. откл.	0.043872	0.030157	0.028647
$\tau = 8$			
среднее	-0.039474	0.114159	0.074686
станд. откл.	0.060682	0.041762	0.038782
$\tau = 16$			
среднее	-0.046658	0.128577	0.081919
станд. откл.	0.075210	0.050297	0.047281
$\tau = 32$			
среднее	-0.052893	0.146422	0.093529
станд. откл.	0.092467	0.061685	0.059766

Табл. 3.7. Выборочные разности оценок параметра H при $H = 0,6$.

$H = 0,6$	$\tilde{H}_2 - \tilde{H}_3$	$\tilde{H}_1 - \tilde{H}_2$	$\tilde{H}_1 - \tilde{H}_3$
$\tau = 1$			
среднее	0.010064	0.062017	0.072081
станд. откл.	0.038681	0.023529	0.027824
$\tau = 2$			
среднее	-0.008340	0.069665	0.061325
станд. откл.	0.043919	0.027160	0.030429
$\tau = 4$			
среднее	-0.008340	0.069665	0.061325
станд. откл.	0.043919	0.027160	0.030429
$\tau = 8$			
среднее	-0.036853	0.089048	0.052195
станд. откл.	0.060298	0.037700	0.040038
$\tau = 16$			
среднее	-0.048286	0.101701	0.053414
станд. откл.	0.072963	0.045552	0.047517
$\tau = 32$			
среднее	-0.059689	0.117374	0.057685
станд. откл.	0.089128	0.055897	0.058955

Табл. 3.8. Выборочные разности оценок параметра H при $H = 0,7$.

$H = 0,7$	$\tilde{H}_2 - \tilde{H}_3$	$\tilde{H}_1 - \tilde{H}_2$	$\tilde{H}_1 - \tilde{H}_3$
$\tau = 1$			
среднее	0.030767	0.042929	0.073696
станд. откл.	0.040552	0.022112	0.030947
$\tau = 2$			
среднее	-0.000573	0.049975	0.049402
станд. откл.	0.045783	0.025470	0.033807
$\tau = 4$			
среднее	-0.000573	0.049975	0.049402
станд. откл.	0.045783	0.025470	0.033807
$\tau = 8$			
среднее	-0.041834	0.066727	0.024893
станд. откл.	0.060457	0.035238	0.040928
$\tau = 16$			
среднее	-0.054999	0.077348	0.022349
станд. откл.	0.070078	0.042391	0.045965
$\tau = 32$			
среднее	-0.068660	0.090443	0.021783
станд. откл.	0.082082	0.051880	0.054137

Табл. 3.9. Выборочные разности оценок параметра H при $H = 0,8$.

$H = 0,8$	$\tilde{H}_2 - \tilde{H}_3$	$\tilde{H}_1 - \tilde{H}_2$	$\tilde{H}_1 - \tilde{H}_3$
$\tau = 1$			
среднее	0.022791	0.028008	0.050800
станд. откл.	0.036683	0.019734	0.029978
$\tau = 2$			
среднее	-0.018853	0.034917	0.016064
станд. откл.	0.042016	0.022705	0.032820
$\tau = 4$			
среднее	-0.018853	0.034917	0.016064
станд. откл.	0.042016	0.022705	0.032820
$\tau = 8$			
среднее	-0.066431	0.050392	-0.016040
станд. откл.	0.054396	0.031492	0.038700
$\tau = 16$			
среднее	-0.083921	0.060134	-0.023787
станд. откл.	0.065044	0.038085	0.044192
$\tau = 32$			
среднее	-0.097920	0.072261	-0.025659
станд. откл.	0.079647	0.046973	0.053042

Табл. 3.10. Выборочные разности оценок параметра H при $H = 0,9$.

$H = 0,9$	$\tilde{H}_2 - \tilde{H}_3$	$\tilde{H}_1 - \tilde{H}_2$	$\tilde{H}_1 - \tilde{H}_3$
$\tau = 1$			
среднее	-0.012699	0.015234	0.002535
станд. откл.	0.040094	0.018159	0.035454
$\tau = 2$			
среднее	-0.051151	0.021121	-0.030030
станд. откл.	0.043646	0.020793	0.037140
$\tau = 4$			
среднее	-0.051151	0.021121	-0.030030
станд. откл.	0.043646	0.020793	0.037140
$\tau = 8$			
среднее	-0.091008	0.033654	-0.057354
станд. откл.	0.055642	0.028571	0.043486
$\tau = 16$			
среднее	-0.105099	0.041602	-0.063497
станд. откл.	0.064605	0.034402	0.048124
$\tau = 32$			
среднее	-0.118214	0.051627	-0.066587
станд. откл.	0.076878	0.042212	0.055334

Для применения построенного ниже критерия требуется, чтобы разности оценок при верной основной гипотезе имели нормальное распределение. Отметим, что в настоящее время не существует теоретического доказательства этого факта, так как требуется совместная нормальность оценок разными методами. Поэтому нормальность разностей оценок проверялась экспериментально применением статистических процедур, описанных в [60].

Алгоритм проверки соответствия модели фрактального гауссовского шума состоит в следующем.

Алгоритм 3.4 *1. Выбрать две оценки параметра H и по имеющимся данным вычислить реализации этих оценок $\tilde{H}_i$ и $\tilde{H}_j$.*

2. По табл. 2.2 – 2.10 для соответствующего H и для соответствующего $tau = 16384/n$ найти оценку математического ожидания $\tilde{a}_{ij}$ и стандартного отклонения $\tilde{\sigma}_{ij}$ разности оценок.

3. Найти достигаемый уровень значимости на основании асимптотической нормальности разностей оценок:

$$\varepsilon^*_{ij} = 2(1 - \Phi(|\tilde{H}_i - \tilde{H}_j - \tilde{a}_{ij}|/\tilde{\sigma}_{ij})).$$

Этот критерий будет применяться для проверки соответствия данных модели фрактального броуновского движения.

3.6 Критерии проверки гипотез об однородности фрактальных гауссовских шумов и их обобщений

Пусть согласно основной гипотезе $(X_1, \ldots, X_{n+m})$ — стационарная гауссовская последовательность случайных величин с математическим ожиданием a дисперсией σ^2 и корреляциями $\mathrm{corr}(\xi_i,\ \xi_j) = r\,(i-j)$, где $r(\cdot)$ — некоторая корреляционная функция.

Отметим, что согласно (11) в случае фрактального гауссовского шума

$$r(t) = \frac{1}{2}\left(|t+1|^{2H} + |t-1|^{2H} - 2|t|^{2H}\right).$$

Рассмотрим статистику

$$J = \frac{\sum_{i=1}^{n} X_i}{\sum_{i=n+1}^{n+m} X_i} = \frac{A}{B}.$$

Справедлива следующая очевидная лемма.

Лемма 3.2 *Числовые характеристики числителя и знаменателя дроби* $J = A/B$ *равны*

$$EA = na,\ EB = ma,$$

$$DA = \sigma^2 \sum_{i=1}^{n} \sum_{j=1}^{n} r\,(i-j)\,,\ DB = \sigma^2 \sum_{i=1}^{m} \sum_{j=1}^{m} r\,(i-j)\,,$$

$$\operatorname{cov}(A,B) = \sigma^2 \sum_{i=1}^{n} \sum_{j=n+1}^{n+m} r\,(j-i)\,.$$

В следующей теореме найдено представление для плотности распределения отношения компонент нормального случайного вектора (в частности, для статистики J).

Теорема 3.4 *Если* (A,B) *— двумерный нормальный вектор, то плотность распределения случайной величины* A/B *имеет вид*

$$f(x) = \frac{\sigma_2}{\sigma_1} f_Z\left(\frac{\sigma_2}{\sigma_1}x\right),$$

где

$$f_Z(t) = \int_{-\infty}^{\infty} |z| \varphi_{a_4,\sigma_4^2}\left(z(t-\rho)\right) \varphi_{a_3,{\sigma_3}^2}(z) dz,$$

где $a_3 = a_2\frac{\sigma_1}{\sigma_2}$, $a_4 = a_1 - \rho a_3$, $a_1 = EA$, $a_2 = EB$, $\sigma_3^2 = \sigma_1^2 = DA$, $\sigma_4 = \sigma_1\sqrt{1-\rho^2}$, $\sigma_2^2 = DB$, $\rho = \frac{cov(A,B)}{\sigma_1\sigma_2}$.

В частности, для статистики J константы $EA, EB, DA, DB, cov(A,B)$ определены ранее в лемме.

Доказательство.

Можно задать A и B следующим образом:

$A = a_1 + \sigma_1(\rho\eta_1 + \sqrt{1-\rho^2}\eta_2)$;

$B = a_2 + \sigma_2\eta_1$,

где $\eta_1, \eta_2 \in \Phi_{0,1}$ — независимые.

Обозначим

$X = A,\ Y = \frac{\sigma_1}{\sigma_2}B,\ Z = X/Y.$

Тогда

$$Z = \frac{\sigma_2 A}{\sigma_1 B} = \frac{\sigma_2}{\sigma_1} J.$$

Следовательно, плотность распределения случайной величины J выражается через плотность распределения случайной величины Z равенством

$$f(x) = \frac{\sigma_2}{\sigma_1} f_Z\left(\frac{\sigma_2}{\sigma_1} x\right).$$

Так как

$X = a_1 + \sigma_1(\rho\eta_1 + \sqrt{1-\rho^2}\eta_2)$,

$Y = a_3 + \sigma_1\eta_1$,

то

$$F_Z(t) = P\{\frac{X}{Y} < t\} =$$

$$= P\{X < tY,\ Y > 0\} + P\{X > tY,\ Y < 0\} =$$

$$= P\{a_1 + \sigma_1(\rho\eta_1 + \sqrt{1-\rho^2}\eta_2) < t(a_3 + \sigma_1\eta_1), Y > 0\} +$$

$$+ P\{X > tY,\ Y < 0\} =$$

$$= P\{\eta_2 < \frac{t(a_3 + \sigma_1\eta_1) - a_1\sigma_1\rho\eta_1}{\sigma_1\sqrt{1-\rho^2}},\ Y > 0\} +$$

$$+ P\{X > tY,\ Y < 0\} =$$

$$= \int_{-\frac{a_2}{\sigma_1}}^{\infty} \Phi_{0,1}(\frac{t(a_3 + \sigma_1 t_1) - a_1 - \sigma_1\rho t_1}{\sigma_1\sqrt{1-\rho^2}})\varphi_{0,1}(t_1)dt_1 +$$

$$+ \int_{-\infty}^{-\frac{a_3}{\sigma_1}} (1 - \Phi_{0,1}(\frac{t(a_3 + \sigma_1 t_1) - a_1 - \sigma_1\rho t_1}{\sigma_1\sqrt{1-\rho^2}})\varphi_{0,1}(t_1))dt_1;$$

$$f_Z(t) = \int_{-\frac{a_3}{\sigma_1}}^{\infty} \frac{a_3 + \sigma_1 t_1}{\sigma_1\sqrt{1-\rho^2}}\varphi_{0,1}(\frac{t(a_3 + \sigma_1 t_1) - a_1 - \sigma_1\rho t_1}{\sigma_1\sqrt{1-\rho^2}})\varphi_{0,1}(t_1))dt_1 -$$

$$- \int_{-\infty}^{-\frac{a_3}{\sigma_1}} \frac{a_3 + \sigma_1 t_1}{\sigma_1\sqrt{1-\rho^2}}\varphi_{0,1}(\frac{t(a_3 + \sigma_1 t_1) - a_1 - \sigma_1\rho t_1}{\sigma_1\sqrt{1-\rho^2}})\varphi_{0,1}(t_1)dt_1;$$

$$f_Z(t) = \int_{-\infty}^{\infty} \frac{|a_3 + \sigma_1 t_1|}{\sigma_1\sqrt{1-\rho^2}} \varphi_{0,1}(\frac{t(a_3+\sigma_1 t_1) - a_1 - \sigma_1\rho t_1}{\sigma_1\sqrt{1-\rho^2}}) \varphi_{0,1}(t_1) dt_1.$$

$$f_Z(t) = \frac{1}{\sigma_3} \int_{-\infty}^{\infty} |a_3 + \sigma_1 t_1| \varphi_{0,1} \left(\frac{t(a_3+\sigma_1 t_1) - a_1 - \sigma_1\rho t_1}{\sigma_1\sqrt{1-\rho^2}} \right) \varphi_{0,1}(t_1) dt_1.$$

Сделаем замену $z = a_3 + \sigma_1 t_1$:

$$f_Z(t) = \frac{1}{\sigma_1\sigma_3} \int_{-\infty}^{\infty} |z| \varphi_{0,1} \left(\frac{tz - a_1 - \rho(z-a_3)}{\sigma_1\sqrt{1-\rho^2}} \right) \varphi_{0,1}(\frac{z-a_3}{\sigma_1}) dz =$$

$$= \frac{1}{\sigma_3} \int_{-\infty}^{\infty} |z| \varphi_{0,1} \left(\frac{tz - a_1 - \rho(z-a_3)}{\sigma_1\sqrt{1-\rho^2}} \right) \varphi_{a_3,{\sigma_1}^2}(z) dz.$$

Обозначим $a_4 = a_1 - \rho a_3$. Тогда

$$f_Z(t) = \int_{-\infty}^{\infty} |z| \varphi_{a_4,\sigma_3^2}\left(z(t-\rho)\right) \varphi_{a_3,{\sigma_1}^2}(z) dz.$$

Теорема доказана.

Получим представление плотности распределения в более явном виде.

Теорема 3.5 *Плотность распределения отношения A/B имеет вид*

$$f(x) = \frac{\sigma_2}{\sigma_1} f_Z\left(\frac{\sigma_2}{\sigma_1} x\right),$$

где

$$f_Z(t+\rho) = \frac{e^{-k_0 d_0}}{2\pi\sigma_4\sigma_3} \left(\frac{e^{-\frac{k_0 b_0^2}{4}}}{k_0} + b_0\sqrt{\frac{\pi}{k_0}} \Phi\left(b_0\sqrt{\frac{k_0}{2}}\right) - \frac{b_0}{2}\sqrt{\frac{\pi}{k_0}} \right),$$

$$k_0 = k_0(t) = \frac{\sigma_3^2 t^2 + \sigma_4^2}{2\sigma_4^2\sigma_3^2},\ b_0 = b_0(t) = -\frac{2a_4 t\sigma_3^2 + 2a_3\sigma_4^2}{\sigma_3^2 t^2 + \sigma_4^2},$$

$$c_0 = c_0(t) = \frac{a_4^2\sigma_3^2 + a_3^2\sigma_4^2}{\sigma_3^2 t^2 + \sigma_4^2},\ d_0 = d_0(t) = c_0(t) - \frac{b_0^2(t)}{4},$$

все константы определены в предыдущей теореме.

Доказательство

Используем ранее полученную формулу:

$$f_Z(t+\rho) = \int_{-\infty}^{\infty} |z| \varphi_{a_4,\sigma_4^2}(zt)) \, \varphi_{a_3,{\sigma_3}^2}(z) dz =$$

$$= \int_0^{\infty} z \varphi_{a_4,\sigma_4^2}(zt)) \, \varphi_{a_3,{\sigma_3}^2}(z) dz - \int_{-\infty}^{0} z \varphi_{a_4,\sigma_4^2}(zt)) \, \varphi_{a_3,{\sigma_3}^2}(z) dz.$$

Вычислим 1-й интеграл:

$$\int_0^{\infty} \frac{1}{\sigma_4\sqrt{2\pi}} \exp\left(-\frac{(zt-a_4)^2}{2\sigma_4^2}\right) \frac{1}{\sigma_3\sqrt{2\pi}} \exp\left(-\frac{(z-a_3)^2}{2\sigma_3^2}\right) z dz =$$

$$= \frac{1}{2\pi\sigma_4\sigma_3} \int_0^{\infty} \exp\left(-\frac{\sigma_3^2(zt-a_4)^2 + \sigma_4^2(z-a_3)^2}{2\sigma_4^2\sigma_3^2}\right) z dz.$$

Преобразуем показатель экспоненты:

$$\frac{\sigma_3^2(zt-a_4)^2 + \sigma_4^2(z-a_3)^2}{2\sigma_4^2\sigma_3^2} =$$

$$= \frac{\sigma_3^2(zt-a_4)^2 + \sigma_4^2(z-a_3)^2}{2\sigma_4^2\sigma_3^2} =$$

$$= \frac{\sigma_3^2(z^2t^2 - 2a_4tz + a_4^2) + \sigma_4^2(z^2 - 2a_3z + a^2)}{2\sigma_4^2\sigma_3^2} =$$

$$= \frac{z^2(\sigma_3^2t^2 + \sigma_4^2) + z(-2a_4t\sigma_3^2 - 2a_3\sigma_4^2) + a_4^2\sigma_3^2 + a_3^2\sigma_4^2}{2\sigma_4^2\sigma_3^2} =$$

$$= \frac{\sigma_3^2t^2 + \sigma_4^2}{2\sigma_4^2\sigma_3^2} \left(z^2 + z\left(-\frac{2a_4t\sigma_3^2 + 2a_3\sigma_4^2}{\sigma_3^2t^2 + \sigma_4^2}\right) + \frac{a_4^2\sigma_3^2 + a_3^2\sigma_4^2}{\sigma_3^2t^2 + \sigma_4^2}\right).$$

Для облегчения вычислений введем новые переменные:

$$k = \frac{\sigma_3^2t^2 + \sigma_4^2}{2\sigma_4^2\sigma_3^2}, b = -\frac{2a_4t\sigma_3^2 + 2a_3\sigma_4^2}{\sigma_3^2t^2 + \sigma_4^2}, c = \frac{a_4^2\sigma_3^2 + a_3^2\sigma_4^2}{\sigma_3^2t^2 + \sigma_4^2},$$

Первый интеграл принимает вид

$$\frac{1}{2\pi\sigma_4\sigma_3}\int_0^\infty e^{-k(z^2+bz+c)}zdz =$$
$$= \frac{1}{2\pi\sigma_4\sigma_3}\int_0^\infty e^{-k((z+\frac{b}{2})^2+d)}zdz.$$

Сделаем замены $z+\frac{b}{2}=u$, $d=c-\frac{b^2}{4}$. Тогда первый интеграл принимает вид

$$\frac{1}{2\pi\sigma_4\sigma_3}\int_{b/2}^\infty e^{-k(u^2+d)}(u-\frac{b}{2})du =$$

$$= \frac{e^{-kd}}{2\pi\sigma_4\sigma_3}(\int_{b/2}^\infty e^{-ku^2}udu - \frac{b}{2}\int_{b/2}^\infty e^{-ku^2}du).$$

Так как

$$\int_{b/2}^\infty e^{-ku^2}udu = \frac{e^{-\frac{kb^2}{4}}}{2k},$$

и $\int_{b/2}^\infty e^{-ku^2}du$ после замены $u\sqrt{2k}=v$ принимает вид

$$\int\limits_{b\sqrt{\frac{k}{2}}}^{\infty} e^{-\frac{v^2}{2}}\frac{dv}{\sqrt{2k}} =$$
$$= \sqrt{\frac{\pi}{k}}\left(1-\Phi\left(b\sqrt{\frac{k}{2}}\right)\right),$$

то 1-й интеграл равен

$$\frac{e^{-kd}}{2\pi\sigma_4\sigma_3}(\int_{b/2}^\infty e^{-ku^2}udu - \frac{b}{2}\int_{b/2}^\infty e^{-ku^2}du) =$$
$$= \frac{e^{-kd}}{2\pi\sigma_4\sigma_3}\left(\frac{e^{-\frac{kb^2}{4}}}{2k} - \frac{b}{2}\sqrt{\frac{\pi}{k}}\left(1-\Phi\left(b\sqrt{\frac{k}{2}}\right)\right)\right).$$

Теперь вычислим 2-й интеграл. С сохранением обозначений, введенных при вычислении первого интеграла, получаем

$$\int_{-\infty}^0 \frac{1}{\sigma_4\sqrt{2\pi}}e^{-\frac{(zt-a_4)^2}{2\sigma_4^2}}\frac{1}{\sigma_3\sqrt{2\pi}}e^{-\frac{(z-a_3)^2}{2\sigma_3^2}}zdz =$$

$$= \frac{e^{-kd}}{2\pi\sigma_4\sigma_3}\left(-\frac{e^{-\frac{kb^2}{4}}}{2k} - \frac{b}{2}\sqrt{\frac{\pi}{k}}\left(\Phi\left(b\sqrt{\frac{k}{2}}\right)\right)\right).$$

В итоге

$$f_Z(t+\rho) = \frac{e^{-kd}}{2\pi\sigma_4\sigma_3}\left(\frac{e^{-\frac{kb^2}{4}}}{2k} - \frac{b}{2}\sqrt{\frac{\pi}{k}}\left(1-\Phi\left(b\sqrt{\frac{k}{2}}\right)\right)\right) -$$

$$-\frac{e^{-kd}}{2\pi\sigma_4\sigma_3}\left(-\frac{e^{-\frac{kb^2}{4}}}{2k} - \frac{b}{2}\sqrt{\frac{\pi}{k}}\Phi\left(b\sqrt{\frac{k}{2}}\right)\right) =$$

$$= \frac{e^{-kd}}{2\pi\sigma_4\sigma_3}\left(\frac{e^{-\frac{kb^2}{4}}}{k} + b\sqrt{\frac{\pi}{k}}\Phi\left(b\sqrt{\frac{k}{2}}\right) - \frac{b}{2}\sqrt{\frac{\pi}{k}}\right).$$

Доказательство завершено.

Полученные формулы использовались в работах [181], [182], [191].

Если ковариационная функция известна не полностью (например, с точностью до параметров σ^2, H), то можно заменить параметры их состоятельными оценками. Критерий проверки гипотезы об однородности строится с помощью доверительного интервала по плотности распределения (асимптотического, если вместо параметров подставляются оценки).

3.7 Модель с зависимыми случайными величинами, распределенными по симметричному устойчивому закону

Результаты этого параграфа являются совместными с В. Е. Хиценко и В. С. Костиным и опубликованы в работе [167].

Моделируется стационарная случайная последовательность, элементы которой имеют симметричное абсолютно непрерывное устойчивое распределение. Совместное распределение элементов последовательности определяется тремя параметрами: параметром Херста,

регулирующим характер зависимости; параметром устойчивого закона, определяющим скорость сходимости плотности распределения к нулю с ростом аргумента; масштабным параметром. Обоснованы и реализованы сильно состоятельные методы оценивания этих параметров. Произведено сравнение различных методов оценивания параметра Херста.

Отметим, что фрактальное броуновское движение не является удовлетворительной моделью экономических временных рядов, приращения которых, как правило, отличаются от нормальных более «тяжелыми хвостами». На это указывает А. Н. Ширяев в работе [98] (гл.3).

Другой автомодельный процесс — процесс Леви, т. е. процесс с независимыми приращениями, распределенными по устойчивому закону, введенный в работах Леви и Хинчина (см. [137]). В этой модели нет зависимости приращений, проявляющей себя в экономических данных.

Обобщением обеих этих моделей в дискретном времени является линейный автомодельный процесс, введенный Такку. Его свойства систематически изучены в монографии Самородницкого и Такку [152]. Приращения этого процесса задаются как скользящие средние от независимых случайных величин, распределенных по устойчивому закону. Коэффициенты скользящего среднего выбираются специальным образом для обеспечения автомодельности процесса и оказываются убывающими по степенному закону с ростом лага.

Оцениванию параметров фрактального броуновского движения, а также параметров процессов Леви посвящено большое число статей. Автору не известны работы, в которых оценивались бы параметры линейного автомодельного процесса, однако построение процедур такого оценивания не является темой нашего рассмотрения.

В данной работе предложен способ генерации последовательности негауссовских случайных величин, имеющих устойчивый абсолютно непрерывный симметричный закон распределения. При этом случайные величины являются зависимыми, и возможно регулировать характер зависимости приращений по аналогии с показателем Херста фрактального броуновского движения. Частными случаями таких последовательностей являются фрактальный гауссовский шум и последовательность приращений процесса Леви.

Последовательность частичных сумм элементов таких последовательностей не является автомодельной (за исключением двух упомянутых выше частных случаев), но является полезной моделью временных рядов, встречающихся при экономическом анализе. Модель характеризуется тремя параметрами, для которых обоснованы и реализованы состоятельные процедуры оценивания. Наряду с оценками по методу максимального правдоподобия, трудоемкость вычисления которых очень велика, предложены и обоснованы сильно состоятельные процедуры оценивания параметров, имеющие малую трудоемкость вычисления, линейно зависящую от объема исследуемых данных. Для параметра автомодельного закона это метод моментов с логарифмической функцией, а для параметра Херста — бинарный знаковый метод. Моделирование показывает, что эти способы дают погрешность оценивания, близкую в среднеквадратическом смысле к погрешности метода максимального правдоподобия. В то же время трудоемкость этих способов гораздо ниже.

Изложим алгоритм моделирования последовательности с зависимыми приращениями, распределенными по устойчивому негауссовскому симметричному закону с абсолютно непрерывным распределением.

Алгоритм состоит в том, что сначала генерируется фрактальный

гауссовский шум, т.е. последовательность приращений фрактального броуновского движения с заданной зависимостью, определяемой показателем Херста. Затем эта нормальная последовательность преобразуется в последовательность с устойчивым симметричным законом распределения, но с более тяжелыми хвостами и с сохранением требуемой зависимости.

Алгоритм 3.5 *1. Генерирование вектора* $\mathbf{W} = (W_1, \ldots, W_n)^T$ *независимых случайных величин со стандартным нормальным распределением.*

2. Преобразование вектора $\mathbf{W}$ *во фрактальный гауссовский шум с параметром* H, $0 < H < 1$ *на основе разложения корреляционной матрицы [130]: матрицу*

$$R = (r(i-j))_{i,j=1}^{n},$$

где

$$r(k) = \mathbf{corr}(X_i;\ X_{i+k}) = \frac{1}{2}\left(|k+1|^{2H} + |k-1|^{2H} - 2|k|^{2H}\right),$$

подвергаем декомпозиции по Холецкому к виду $R = AA^T$, *где* A *— нижняя треугольная матрица. Далее формируем вектор фрактального гауссовского шума по схеме* $\mathbf{X} = A\mathbf{W}$.

3. Преобразование приращений к автомодельному закону по формуле

$$Y_i = {F_\alpha}^{-1}(\Phi(X_i)),$$

где Φ *— функция распределения стандартного нормального закона,* F_α *— функция распределения симметричного автомодельного закона.*

Вычисление функции распределения F_α симметричного абсолютно непрерывного автомодельного закона рассматривается ниже. Параметр α может принимать значения из полуинтервала $(1;\ 2]$.

Отметим, что Y_i распределены как приращения автомодельного процесса с независимыми приращениями, однако частичные суммы случайных величин Y_i не образуют автомодельного процесса (при одновременном выполнении неравенств $H \neq 1/2$ и $\alpha \neq 2$). Отсутствие автомодельности можно проверить непосредственно вычислением распределения суммы Y_1+Y_2. Параметрами процесса являются показатель Херста $H \in (0;\ 1)$ и коэффициент автомодельности приращений $\alpha \in (1;\ 2]$. При $H = 1/2$ случайные величины Y_i являются приращениями процесса Леви (в дискретном времени). При $\alpha = 2$ случайные величины Y_i образуют фрактальный гауссовский шум (в дискретном времени).

Следуя [149], мы будем использовать параметризацию (A) Золотарева [41]. В соответствии с этой параметризацией, характеристическая функция симметричного абсолютно непрерывного автомодельного закона имеет вид

$$\varphi(\lambda) = \exp\left(-\sigma^\alpha |\lambda|^\alpha\right).$$

В частности, при $\alpha = 2$ получаем центрированное нормальное распределение с дисперсией $2\sigma^2$.

Пусть F_α — функция распределения симметричного абсолютно непрерывного устойчивого закона. Соответствующая плотность распределения равна

$$f_\alpha(x) = \frac{1}{\pi}\int_0^\infty \cos(tx)\exp\left(-t^\alpha\right)dt.$$

Раскладывая косинус в ряд и интегрируя, можно прийти к формуле

$$F_\alpha(x) = \frac{1}{2} + \frac{1}{\pi}\sum_{n=0}^{\infty}\frac{(-1)^n x^{2n+1}}{\alpha(2n+1)!}\Gamma\left(\frac{2n+1}{\alpha}\right).$$

Здесь Γ — гамма-функция.

Ряд сходится для всех $x \in \mathbf{R}$ при $\alpha \in (1;\ +\infty)$.

В области сходимости ряд имеет смысл функции распределения устойчивого закона при $\alpha \leq 2$.

Скорость сходимости ряда зависит от значения параметра α и от значения переменной x. Чем меньше α и больше $|x|$, тем больше требуется слагаемых. Так, при $\alpha = 1,1$, $x = 10$ и заданной точности 10^{-6} требуется порядка 10^{10} слагаемых. При $\alpha = 1$ ряд сходится только для $|x| < 1$.

Таким образом, практические вычисления по этой формуле возможны для $\alpha > 1$, не слишком близких к 1.

Поэтому есть потребность в алгоритме, позволяющем вычислять F_α с высокой точностью для любых значений $\alpha \in (1;\ 2)$.

Мы будем использовать алгоритм Нолана, изложенный в [149]. Алгоритм основан на том, что интеграл, вычисляющий функцию распределения, приводится с помощью специальной замены к интегралу от ограниченной функции по конечному промежутку.

Для нашей задачи принимаем $\beta = 0$ (распределение не содержит дискретной компоненты), $\sigma = 1$ (параметр масштаба равен 1), $\mu = 0$ (смещение нулевое).

Введенную в [149] функцию $V(\theta; \alpha, \beta)$ при $\beta = 0$ обозначим $V_\alpha(\theta)$. Получаем

$$V_\alpha(\theta) = \left(\frac{\cos\theta}{\sin(\alpha\theta)}\right)^{\frac{\alpha}{\alpha-1}} \cdot \frac{\cos((\alpha-1)\theta)}{\cos\theta}.$$

Теорема 1 из [149] переформулируется следующим образом.

Для $x > 0$

$$f_\alpha(x) = \frac{\alpha x^{\frac{1}{\alpha-1}}}{\pi(\alpha-1)} \int_0^{\pi/2} V_\alpha(\theta) \exp\left(-x^{\frac{\alpha}{\alpha-1}} V_\alpha(\theta)\right) d\theta;$$

$$F_\alpha(x) = 1 - \frac{1}{\pi} \int_0^{\pi/2} \exp\left(-x^{\frac{\alpha}{\alpha-1}} V_\alpha(\theta)\right) d\theta;$$

$$f_\alpha(-x) = f_\alpha(x);\ F_\alpha(-x) = 1 - F_\alpha(x).$$

Кроме того, $f_\alpha(0) = \frac{\Gamma\left(1+\frac{1}{\alpha}\right)}{\pi};\ F_\alpha(0) = \frac{1}{2}$.

В [149] отмечено, что подынтегральная функция $\exp\left(-x^{\frac{\alpha}{\alpha-1}}V_\alpha(\theta)\right)$ ограничена, непрерывна и монотонно возрастает по θ на отрезке интегрирования. Отметим, что она равна 0 в точке 0, и равна 1 в точке $\pi/2$. Вычисление значений $F_\alpha(x)$ будем проводить численным интегрированием. Исключением являются случаи больших или маленьких значений x. При больших x подынтегральная функция очень близка к 0 всюду на отрезке, за исключением малой окрестности точки $\pi/2$. При близких к нулю x подынтегральная функция очень близка к 1 всюду на отрезке, за исключением малой окрестности точки 0.

В зависимости от требуемой точности $\Delta > 0$ и значения переменной x найдем нижнюю границу интервала интегрирования θ_- для больших значений x. Пусть $\theta_- = \frac{\pi}{2} - \varepsilon$. При $\varepsilon \to 0$ функция $V_\alpha(\frac{\pi}{2} - \varepsilon)$ одного порядка малости с $(\sin\varepsilon)^{\frac{\alpha}{\alpha-1}-1} \sim \varepsilon^{\frac{1}{\alpha-1}}$. Из равенства

$$\exp\left(-x^{\frac{\alpha}{\alpha-1}}\varepsilon^{\frac{1}{\alpha-1}}\right) = \Delta$$

получаем

$$\varepsilon = (-\ln\Delta)^{\alpha-1}x^{-\alpha};\ \theta_- = \frac{\pi}{2} - (-\ln\Delta)^{\alpha-1}x^{-\alpha}.$$

Этот интервал будем использовать при $\varepsilon < (-\ln\Delta)^{-1}$, то есть при $x > -\ln\Delta$.

Найдем верхнюю границу интервала интегрирования θ_+ для малых значений x в зависимости от $\Delta > 0$ и x. Пусть $\theta_- = \frac{\pi}{2} - \varepsilon$. При $\theta \to 0$ функция $V_\alpha(\theta)$ эквивалентна $(\alpha\theta)^{-\frac{\alpha}{\alpha-1}} \sim \varepsilon^{\frac{1}{\alpha-1}}$. Из эквивалентности

$$\exp\left(-x^{\frac{\alpha}{\alpha-1}}(\alpha\theta)^{-\frac{\alpha}{\alpha-1}}\right) \sim 1 - \Delta$$

получаем

$$\theta \sim \frac{x}{\alpha\Delta^{\frac{\alpha-1}{\alpha}}}.$$

Положим θ_+ равным правой части:

$$\theta_+ = \frac{x}{\alpha\Delta^{\frac{\alpha-1}{\alpha}}}.$$

Этот интервал будем использовать при $\theta_+ < 0,1$, то есть при $x < 0{,}1\alpha\Delta^{\frac{\alpha-1}{\alpha}}$.

Итак, получаем следующий алгоритм. Для данных $\alpha \in (1;\ 2)$, $\Delta > 0$ определяем интервал интегрирования:

если $x > -\ln\Delta$, то $\theta_- = \frac{\pi}{2} - (-\ln\Delta)^{\alpha-1}x^{-\alpha}$; иначе $\theta_- = 0$;

если $x < 0{,}1\alpha\Delta^{\frac{\alpha-1}{\alpha}}$, то $\theta_+ = \frac{x}{\alpha\Delta^{\frac{\alpha-1}{\alpha}}}$; иначе $\theta_+ = \frac{\pi}{2}$.

На интервале $(0;\ \theta_-)$ принимаем подынтегральную функцию равной 0;

на интервале $(\theta_+;\ \frac{\pi}{2})$ принимаем подынтегральную функцию равной 1;

на интервале $(\theta_-;\ \theta_+)$ вычисляем интеграл на основе квадратурной формулы Гаусса—Кронрода с 61 точкой ([79], §12.9).

Для отрицательных значений x используем формулу $F_\alpha(-x) = 1 - F_\alpha(x)$.

Сравнение результатов вычислений с результатами работы алгоритма Нолана показывает, что для $\alpha \in [1,1;\ 2]$ и $x \in [-10;\ 10]$ погрешность не превосходит 10^{-14}.

Оценивание параметров методом максимального правдоподобия по выборке $\mathbf{Y} = (Y_1, \ldots, Y_n)$ состоит в максимизации плотности совместного распределения

$$f(\mathbf{X}) = \frac{1}{(2\pi)^{n/2}\sigma^n\sqrt{\det R}}\exp(-\frac{1}{2\sigma^2}\mathbf{X}^T R^{-1}\mathbf{X}),$$

где $\mathbf{X} = (X_1, \ldots, X_n)$,

$$X_i = \Phi^{-1}(F_\alpha(Y_i)).$$

Максимизацию нужно проводить по совокупности параметров α, H, σ. Однако такая процедура оказывается вычислительно слож-

ной, и для выборок большого объема ее реализовать не удается. Поэтому будем использовать приближенный алгоритм: сначала найдем оценку параметра α в начальном приближении, затем ее уточним. Потом найдем оценки параметров σ и H.

В качестве начального приближения оценки параметра α будем использовать оценку по «хвосту» распределения. Эта оценка основана на том, что левый и правый «хвосты» симметричного устойчивого распределения с параметром α асимптотически пропорциональны $x^{-\alpha}$. Следовательно, для достаточно больших h плотность распределения случайной величины $Y_i\mathbf{I}\{|Y_i| \geq h\}$ приближенно равна

$$f(x) = K|x|^{-\alpha-1}\mathbf{I}\{|x| \geq h\}.$$

Найдем константу K из условий нормировки:

$$2\int_h^\infty f(x)dx = \frac{2Kh^{-\alpha}}{\alpha},$$

$$K = \frac{\alpha h^\alpha}{2}.$$

Найдем оценку параметра α:

$$\ln f(x) = \ln\alpha + \alpha\ln h - \ln 2 - (\alpha+1)\mathbf{I}\{|x| \geq h\}\ln|x|,$$

$$\frac{\partial}{\partial\alpha}\ln f(x) = \frac{1}{\alpha} + \ln h - \mathbf{I}\{|x| \geq h\}\ln|x|,$$

$$\sum_{i=1}^n \left(\frac{1}{\alpha} + \ln h - \mathbf{I}\{|Y_i| \geq h\}\ln|Y_i|\right) = 0,$$

$$\hat{\alpha} = \frac{1}{\overline{\mathbf{I}\{|Y| \geq h\}\ln|Y|} - \ln h}.$$

Проведенные расчеты показывают, что к наиболее точным результатам приводит выбор в качестве h 15% -й точки, то есть числа, которое по модулю превосходят 15% элементов выборки.

Теорема 3.6 *Оценка $\tilde{\alpha}$, максимизирующая*

$$\sum_{i=1}^n \ln f_\alpha(Y_i),$$

является сильно состоятельной оценкой параметра α.

Доказательство

Докажем существование математического ожидания

$$\mathbf{E} \ln f_\alpha(Y_1) = 2 \int_0^\infty f_\alpha(x) \ln f_\alpha(x) dx.$$

Так как $f_\alpha(+0) = f_\alpha(0) = \frac{1}{\pi}\Gamma(1 + \frac{1}{\alpha}) < \infty$, то

$$\int_0^1 f_\alpha(x) \ln f_\alpha(x) dx$$

сходится. Так как $f_\alpha(x) \sim Cx^{-\alpha-1}$ при $x \to +\infty$, то

$$\int_1^\infty f_\alpha(x) \ln f_\alpha(x) dx$$

сходится. Таким образом, $\mathbf{E} \ln f_\alpha(Y_1)$ существует.

Случайные величины $\ln f_\alpha(Y_i)$ — это функции от Z_i — компонент фрактального гауссовского шума. Так как последовательность $\{Z_i\}$ стационарна, и ее корреляционная функция стремится к нулю с ростом лага, то для последовательности $\{\ln f_\alpha(Y_i)\}$ выполнен усиленный закон больших чисел:

$$\frac{1}{n} \sum_{i=1}^{n} \ln f_\alpha(Y_i) \to \mathbf{E} \ln f_\alpha(Y_1)$$

с вероятностью 1 согласно эргодической теореме Биркгофа (см., напр., [82], с. 308).

В случае $H = 1/2$ оценка $\tilde{\alpha}$ является оценкой максимального правдоподобия, и ее сильная состоятельность следует из общих теорем об оценках максимального правдоподобия в силу существования $\mathbf{E} \ln f_\alpha(Y_1)$, а также ограниченности $f_\alpha(x)$ и ее непрерывности по параметру α при $\alpha \in [1;\ 2]$ (см. [9], с. 122, следствие 3).

В случае $H \neq 1/2$ оценка $\tilde{\alpha}$ не является оценкой максимального правдоподобия, но она отыскивается максимизацией той же суммы, что и в случае независимых компонент, и в силу сходимости является сильно состоятельной.

Доказательство завершено.

Для оценивания параметров σ^2 и H преобразуем данные к нормальному закону.

Нахождение оценки максимального правдоподобия и оценки бинарным знаковым методом осуществляется после этого преобразования так, как это было описано в соответствующих параграфах.

Приведем результаты для модельных данных.

Алгоритм моделирования и алгоритмы оценивания параметров были реализованы для каждого значения H от 0,1 до 0,9 с шагом 0,1. Шаг по α также равен 0,1, значения от 1,1 до 1,9. Результаты, приведенные в таблицах, получены усреднением 7218 вычислений для каждого значения α и H. Объем выборки n во всех случаях равнялся 1024. Через $\widehat{\alpha}$ и $\widehat{\sigma}_\alpha$ обозначены среднее выборочное значение и выборочное среднеквадратическое отклонение оценки параметра α.

Таблица 3.11. Результаты оценивания параметра α модифицированным методом максимального правдоподобия.

α	1,1	1,2	1,3	1,4	1,5	1,6	1,7	1,8	1,9
$\widehat{\alpha}$	1,101	1,201	1,303	1,402	1,503	1,601	1,700	1,803	1,903
$\widehat{\sigma}_\alpha$	0,023	0,026	0,028	0,029	0,031	0,032	0,032	0,032	0,029

В следующей таблице через $\widehat{H}$ и $\widehat{B}_H$ обозначены среднее выборочное значение и выборочное среднеквадратическое отклонение оценки параметра H модифицированным методом максимального правдоподобия.

Таблица 3.12. Результаты оценивания параметра H модифицированным методом максимального правдоподобия.

H	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
$\widehat{H}$	0,100	0,200	0,300	0,399	0,499	0,599	0,699	0,799	0,898
$\widehat{B}_H$	0,011	0,015	0,017	0,018	0,019	0,020	0,020	0,021	0,021

Среднеквадратическое отклонение оценки меньше при малых H.

В таблице 3.13 через $\tilde{H}$ и $\tilde{B}_H$ обозначены среднее выборочное значение и выборочное среднеквадратическое отклонение оценки параметра H бинарным знаковым методом.

Таблица 3.13. Результаты оценивания параметра H бинарным знаковым методом.

H	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
$\tilde{H}$	0,100	0,200	0,299	0,399	0,499	0,600	0,699	0,798	0,893
$\tilde{B}_H$	0,032	0,031	0,031	0,029	0,029	0,029	0,030	0,037	0,045

Оценка бинарным знаковым методом имеет отрицательное смещение, возрастающее по модулю с ростом H.

Среднеквадратическое отклонение оценки почти не зависит от параметра.

Сравнивая результаты вычислений, приведенные в таблицах 3.12 и 3.13, следует отметить, что отношение $\tilde{B}_H/\widehat{B}_H$ не превосходит 1,5 для $0,5 \leq H \leq 0,7$. Отметим, что такие значения параметра H типичны для ряда приложений. При этих значениях параметра H для достижения такого же среднеквадратического отклонения оценки модифицированным знаковым методом требуется примерно в 2,25 раза больше наблюдений, чем для оценки модифицированным методом максимального правдоподобия. Использование модифицированного знакового метода представляется оправданным ввиду его малой вычислительной сложности в ситуации, когда объем наблюдаемых данных велик.

3.8 Результаты главы 3

В главе 3 получены следующие основные результаты:

- Получен новый статистический критерий проверки нормальности малых выборок, основанный на отношении размаха выборки к минимальному спейсингу. Показано, что он лучше критерия Шапиро—Уилка при резко асимметричных альтернативах. Исправлены критические значения для критерия Шапиро—Уилка.
- Получен статистический критерий проверки гипотезы о том, что показатель Херста H фрактального гауссовского шума равен $1/2$ против альтернативной гипотезы $H \neq 1/2$. Критерий основан на использовании оценки параметра бинарным знаковым методом, найденной в предыдущей главе.
- Разработан алгоритм обнаружения разладки фрактального гауссовского шума. Статистический критерий отвергает гипотезу об отсутствии разладки, если разность между оценками параметра Херста превышает критическое значение. Критические значения вычисляются по результатам моделирования фрактального гауссовского шума. Показано, что наилучшим оказывается критерий, использующий оценки бинарным знаковым методом и методом дисперсии.
- Предложен критерий проверки гипотезы о стационарности гауссовской последовательности. Критерий основан на вычислении плотности распределения отношения сумм элементов последовательности при выполнении гипотезы.
- Предложена и изучена новая вероятностная модель — стационарная случайная последовательность, элементы которой имеют симметричное абсолютно непрерывное устойчивое распределе-

ние. Совместное распределение элементов последовательности определяется тремя параметрами: параметром Херста, регулирующим характер зависимости; параметром устойчивого закона, определяющим скорость сходимости плотности распределения к нулю с ростом аргумента; масштабным параметром. Обоснованы и реализованы сильно состоятельные методы оценивания этих параметров. Произведено сравнение различных методов оценивания параметра Херста.

ГЛАВА 4
Анализ однородности текстов

4.1 Вводные замечания

Интернет можно рассматривать как систему, структурными элементами которой являются тексты. Возникает задача анализа этих элементов. Тексты могут иметь внутреннюю структуру, в частности, быть составленными из разнородных частей, что выявляется методами обнаружения разладки. Как будет показано, модель выборки не является вполне адекватной для моделирования текстов, и для однородных текстов используется модель фрактального гауссовского шума, а для разнородных — модель разладки фрактального гауссовского шума. Для анализа текстов используется математический аппарат обнаружения разладки в модели выборки, обнаружения зависимости элементов (отличия коэффициента Херста от 1/2 в модели фрактального гауссовского шума), статистические критерии обнаружения разладки фрактального гауссовского шума, то есть изменения его параметров. Таким образом, применяется весь математический аппарат, разработанный в предыдущих главах.

В параграфе 4.2 рассмотрены однопараметрические вероятностные модели текста. Рассмотрены статистические свойства оценок параметров, отыскиваемых на основании статистики числа разных слов. Для последовательности количеств разных слов в модели Мандельброта выполняется функциональная центральная предельная теорема. Показана неадекватность модели выборки (систематическое изменение оценок параметров с ростом объема текста). В параграфе 4.3 разработан метод авторского инварианта — способ сопоставления тексту числовой последовательности, сохраняющей свои статистические свойства для фиксированного автора. Статистиче-

ские критерии обнаружения разладки применяются для анализа однородности текстов, а также комбинаций текстов разных авторов. Выработаны рекомендации к использованию этих методов для исследования авторства. Проанализированы собрания сочинений и показано, что модель выборки для текстов одного автора (и модель разладки для текстов двух авторов) становятся неадекватными при большом объеме текста. В параграфе 4.4 методы обнаружения фрактальности и методы обнаружения разладки фрактального гауссовского шума применяются к анализу текстов одного и двух авторов. В параграфе 4.5 приведены основные результаты главы 4.

4.2 Однопараметрические вероятностные модели статистик текста

Результаты этого параграфа являются совместными с Н. С. Закревской и М. Г. Чебуниным и опубликованы в работах [163], [171].

4.2.1 Оценки параметров и их состоятельность

Объектом изучения являются статистики текста на естественном языке. Исследуется адекватность моделирования этих статистик с помощью однопараметрических вероятностных распределений: Мандельброта с бесконечным носителем, Ципфа, геометрического. Текст моделируется последовательностью независимых одинаково распределенных случайных величин. Программно реализован подсчет статистик текста. Монотонная зависимость математического ожидания числа разных слов в тексте от параметра в каждой из исследуемых моделей позволяет строить оценки по методу подстановки. Реализованы алгоритмы отыскания оценок параметров и алгоритмы нахождения реально достигнутого уровня значимости. Проведен анализ ряда поэтических текстов на русском, английском,

немецком, французском языках. Выявлены зависимости параметров от языка и от года написания стихотворения.

Текст на естественном языке рассматривается как последовательность слов $X_1, \ldots, X_n$. Здесь величины X_i, $i = 1, \ldots, n$, принимают значения в множестве $\mathcal{S}$, которое мы будем называть словником языка. Исследуются статистики текста, которые не связаны непосредственно с именами слов, то есть измеримые относительно сигма-алгебры, порожденной событиями $\{X_i = X_j\}$, $1 \le i < j \le n$.

Изучается адекватность моделирования этих статистик с помощью однопараметрических вероятностных моделей текста: $X_1, \ldots, X_n$ независимы и одинаково распределены с распределением $\mathrm{P}\{X_1 = i\} = p_i$, $i = 1, 2, \ldots$

Обозначим $N^{(n)}$ — число различных слов в тексте:

$$N^{(n)} = 1 + \sum_{i=2}^{n} \mathbf{I}\{X_i \ne X_j,\ j = 1, \ldots, i-1\};$$

$N_k^{(n)}$ — число слов, встретившихся ровно k раз ($k = 1, \ldots, n$):

$$N_k^{(n)} = \sum_{1 \le i_1 < \ldots < i_k \le n} \mathbf{I}\{X_{i_1} = \ldots = X_{i_k} \ne X_j,\ j \in \{1, \ldots, n\} \backslash \{i_1, \ldots, i_k\}\}.$$

Заметим, что $N^{(n)} = \sum_{k=1}^{n} N_k^{(n)}$.

В работах [1–3] исследован вид *частотного спектра* $\{q_i,\ i = 1, 2, \ldots\}$ текста:

$$q_i = \frac{k}{n}, \quad i = N_n^{(n)} + \ldots + N_{k+1}^{(n)} + j, \quad j = 1, \ldots, N_k^{(n)}.$$

Дж. К. Ципф (см. [100]) показал, что частотный спектр близок к функции $q_i = S/i$, $i = 1, \ldots, M$. В работе Б. Мандельброта [63] приведен ряд теоретических соображений (простая вероятностная модель, информационная модель, марковская модель), приводящих

к закону $q_i = C \cdot i^{-\alpha}$, $i = 1, \ldots, M$, $\alpha > 1$. В статье Ю. А. Шрейдера [99] тот же закон получен для вероятностной модели наибольшей степени общности. Исследование частотного спектра сопряжено с рядом трудностей. Так, если предполагать, что существуют вероятности $p_1 \geq p_2 \geq \ldots$ употребления слов данным автором, то $q_i \to p_i$ почти наверное, но для любого конечного текста (за исключением вырожденных искусственных примеров) число слов, встретившихся 1 раз, положительно. Для этих слов $q_i = 1/n$ и плохо аппроксимируется приведенными законами. Кроме того, для малых значений i также могут иметь место отклонения от приведенных законов, обусловленные особенностями авторского стиля [99].

Поэтому в настоящей работе использован следующий подход: параметр распределения оценивается по значению статистики $N^{(n)}$, а адекватность модели проверяется с помощью статистики $N_1^{(n)}$.

Рассматриваются 3 модели:

1) $p_i = C(\alpha) i^{-\alpha}$, $i = 1, 2, \ldots$; $\alpha > 1$. Здесь

$$C(\alpha) = \left(\sum_{i=1}^{\infty} i^{-\alpha} \right)^{-1}$$

(распределение Мандельброта с бесконечным носителем).

2) $p_i = S(M) i^{-1}$, $i = 1, \ldots, M$; M — целое. Здесь

$$S(M) = \left(\sum_{i=1}^{M} i^{-1} \right)^{-1}$$

(распределение Ципфа).

3) $p_i = p(1-p)^{i-1}$, $i = 1, 2, \ldots$; $0 < p < 1$ (геометрическое распределение).

В [133] получен ряд предельных теорем для статистики $N_1^{(n)}$. В частности, показано, что если $\lim_{i \to \infty} p_{i+1}/p_i = 1$ (как в модели 1), то $N_1^{(n)} \to_p \infty$ при $n \to \infty$ (теорема 3(1)). Если $\limsup_{i \to \infty} p_{i+1}/p_i <$

1 (как в модели 3), то $\mathbf{E}N_1^{(n)}$ равномерно ограничено (теорема 2, теорема 3(4)).

Следующая элементарная лемма часто используется в дальнейшем.

Лемма 4.1 *Для любого* $n = 1,\ 2,\ \ldots,$ *выполнено:*
$\mathbf{E}N^{(n)} = \Sigma_{i=1}^{\infty}(1-(1-p_i)^n)$.
$\mathbf{E}N_k^{(n)} = c_n^k \Sigma_{i=1}^{\infty} {p_i}^k(1-p_i)^{n-k},\ k = 1, \ldots, n.$

Доказательство.

$$\mathbf{E}N^{(n)} = \sum_{i=1}^{\infty} \mathbf{P}(\{X_1 = i\} \cup \ldots \cup \{X_n = i\})$$

$$= \sum_{i=1}^{\infty}(1 - \mathbf{P}(\{X_1 \neq i\} \cap \ldots \cap \{X_n \neq i\})) = \sum_{i=1}^{\infty}(1-(1-p_i)^n);$$

$$\mathbf{E}N_k^{(n)} = \sum_{i=1}^{\infty} c_n^k \mathbf{P}\{X_1 = \ldots = X_k = i,\ X_{k+1} \neq i,\ \ldots\ ,\ X_n \neq i\}$$

$$= c_n^k \sum_{i=1}^{\infty} {p_i}^k (1-p_i)^{n-k}.$$

Доказательство завершено.

Для обоснования алгоритма нахождения оценок параметров нам потребуется утверждение о монотонной зависимости математического ожидания от параметра при замене p_k их значениями в моделях 1, 2, 3.

Теорема 4.1 *Для любого фиксированного* $n > 1$ *математическое ожидание числа различных слов* $\mathbf{E}N^{(n)}$ *строго монотонно по* α, M, p *в моделях 1, 2, 3.*

При доказательстве используем следующую техническую лемму.

Лемма 4.2 *Пусть* $\{a_i\}$, $\{b_i\}$, $i = 1, 2, \ldots$ — *числовые последовательности,* $0 < a_i < 1$, *последовательность* $\{a_i\}$ *строго возрас-*

имеет, $b_i < 0$ *при* $i < i_0$, $b_i \geq 0$ *при* $i \geq i_0$, $\sum_{i=1}^{\infty} b_i = 0$. *Тогда* $\sum_{i=1}^{\infty} a_i b_i > 0$.

Доказательство.

$$\sum_{i=1}^{\infty} a_i b_i = \sum_{i=1}^{i_0-1} a_i b_i + \sum_{i=i_0}^{\infty} a_i b_i > a_{i_0} \cdot \left(\sum_{i=1}^{i_0-1} b_i + \sum_{i=i_0}^{\infty} b_i \right) = 0.$$

Доказательство леммы завершено.

Доказательство теоремы.

1) В модели 1

$$\mathbf{E}N^{(n)} = \sum_{i=1}^{\infty} \left(1 - \left(1 - \frac{i^{-\alpha}}{\sum_{j=1}^{\infty} j^{-\alpha}} \right)^n \right).$$

$$\frac{d}{d\alpha} \mathbf{E}N^{(n)} =$$

$$= \sum_{i=1}^{\infty} n \left(1 - \frac{i^{-\alpha}}{\sum_{j=1}^{\infty} j^{-\alpha}} \right)^{n-1} \frac{-i^{-\alpha} \ln i \sum_{j=1}^{\infty} j^{-\alpha} + i^{-\alpha} \sum_{j=1}^{\infty} j^{-\alpha} \ln j}{\left(\sum_{j=1}^{\infty} j^{-\alpha} \right)^2}.$$

Положим

$$a_i = \left(1 - \frac{i^{-\alpha}}{\sum_{j=1}^{\infty} j^{-\alpha}} \right)^{n-1};$$

$$b_i = i^{-\alpha} \ln i \sum_{j=1}^{\infty} j^{-\alpha} - i^{-\alpha} \sum_{j=1}^{\infty} j^{-\alpha} \ln j.$$

Тогда выполнены условия леммы при

$$i_0 = \exp \left(\frac{\sum_{j=1}^{\infty} j^{-\alpha} \ln j}{\sum_{j=1}^{\infty} j^{-\alpha}} \right).$$

Следовательно, $\sum_{i=1}^{\infty} a_i b_i > 0$, $\frac{d}{d\alpha} \mathbf{E}N^{(n)} < 0$.

2) Исследуем монотонность по M в модели 2. Докажем, что для всех $M \geq 1$ выполнено неравенство

$$\sum_{i=1}^{M+1} \left(1 - \left(1 - \frac{i^{-1}}{\sum_{j=1}^{M+1} j^{-1}} \right)^n \right) > \sum_{i=1}^{M} \left(1 - \left(1 - \frac{i^{-1}}{\sum_{j=1}^{M} j^{-1}} \right)^n \right),$$

то есть

$$1-\left(1-\frac{(M+1)^{-1}}{\sum_{j=1}^{M+1} j^{-1}}\right)^n > \sum_{i=1}^{M}\left(\left(1-\frac{i^{-1}}{\sum_{j=1}^{M+1} j^{-1}}\right)^n-\left(1-\frac{i^{-1}}{\sum_{j=1}^{M} j^{-1}}\right)^n\right). \tag{35}$$

Так как

$$a^n-b^n=(a-b)\sum_{k=0}^{n-1} a^{n-1-k}b^k \tag{36}$$

для любого $n\geq 2$, то $a^n-b^n\leq n(a-b)a^{n-1}$ при $a\geq b\geq 0$, и правая часть неравенства (35) допускает оценку

$$\sum_{i=1}^{M}\left(\left(1-\frac{i^{-1}}{\sum_{j=1}^{M+1} j^{-1}}\right)^n-\left(1-\frac{i^{-1}}{\sum_{j=1}^{M} j^{-1}}\right)^n\right)\leq$$

$$\leq\sum_{i=1}^{M} n\left(\frac{i^{-1}}{\sum_{j=1}^{M} j^{-1}}-\frac{i^{-1}}{\sum_{j=1}^{M+1} j^{-1}}\right)\left(1-\frac{M^{-1}}{\sum_{j=1}^{M+1} j^{-1}}\right)^{n-1}=$$

$$=n\left(1-\frac{\sum_{i=1}^{M} i^{-1}}{\sum_{i=1}^{M+1} i^{-1}}\right)\left(1-\frac{M^{-1}}{\sum_{j=1}^{M+1} j^{-1}}\right)^{n-1}=$$

$$=\frac{n(M+1)^{-1}}{\sum_{i=1}^{M+1} i^{-1}}\left(1-\frac{M^{-1}}{\sum_{i=1}^{M+1} i^{-1}}\right)^{n-1}.$$

Преобразуем левую часть неравенства (35) в соответствии с формулой (36) и получим искомое неравенство:

$$1-\left(1-\frac{(M+1)^{-1}}{\sum_{j=1}^{M+1} j^{-1}}\right)^n=\frac{(M+1)^{-1}}{\sum_{i=1}^{M+1} i^{-1}}\sum_{j=0}^{n-1}\left(1-\frac{(M+1)^{-1}}{\sum_{i=1}^{M+1} i^{-1}}\right)^j>$$

$$>\frac{n(M+1)^{-1}}{\sum_{i=1}^{M+1} i^{-1}}\left(1-\frac{(M+1)^{-1}}{\sum_{i=1}^{M+1} i^{-1}}\right)^{n-1}>\frac{n(M+1)^{-1}}{\sum_{i=1}^{M+1} i^{-1}}\left(1-\frac{M^{-1}}{\sum_{i=1}^{M+1} i^{-1}}\right)^{n-1}.$$

3) Докажем монотонное возрастание $\mathbf{E}N^{(n)}$ при $p\to 0$ в модели 3.

$$\frac{d}{dp}\mathbf{E}N^{(n)}=\frac{d}{dp}\sum_{i=1}^{\infty}(1-(1-p(1-p)^{i-1})^n)=$$

$$=\sum_{i=1}^{\infty}(-n)(1-p(1-p)^{i-1})^{n-1}(-(1-p)^{i-1}+p(i-1)(1-p)^{i-2})=$$

$$=-n\sum_{i=1}^{\infty}(1-p(1-p)^{i-1})^{n-1}(1-p)^{i-2}(ip-1).$$

Так как последовательность $a_i = (1-p(1-p)^{i-1})^{n-1}$ строго возрастает при $i \to \infty$, полагая $b_i = (1-p)^{i-2}(ip-1)$, получаем, что

$$\sum_{i=1}^{\infty} b_i = (1-p)^{-1}\left(p\sum_{i=1}^{\infty} i(1-p)^{i-1} - \sum_{i=1}^{\infty}(1-p)^{i-1}\right)$$
$$= (1-p)^{-1}\left(-p\cdot\frac{d}{dp}\left(\frac{1-p}{p}\right) - \frac{1}{p}\right) = 0.$$

$b_i < 0$ при $i < p^{-1}$; $b_i \geq 0$ при $i \geq p^{-1}$. Следовательно, $\frac{d}{dp}\mathbf{E}N^{(n)} < 0$ при $n > 1$, $0 < p < 1$.

Доказательство завершено.

В силу теоремы, существует единственный корень уравнения $\mathbf{E}N^{(n)} = N$ в моделях 1, 2, 3.

Докажем состоятельность оценок, полученных подстановкой $m_n(\tilde{\theta}) = N^{(n)}$, где $m_n(\theta) = \mathbf{E}N^{(n)}$, с подстановкой α, M, p вместо θ в моделях 1, 2, 3.

Сначала докажем следующую лемму.

Лемма 4.3 *Имеет место неравенство* $\mathbf{D}N^{(n)} \leq \mathbf{E}N^{(n)}$.

Доказательство.

Заметим, что

$$\mathbf{E}\left(N^{(n)}\right)^2 = \mathbf{E}\left(\sum_{i=1}^{\infty}(1-\mathbf{I}\{X_1 \neq i,\ldots,X_n \neq i\})\right)^2$$
$$= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}(1-\mathbf{P}\{X_1 \neq i,\ldots,X_n \neq i\} - \mathbf{P}\{X_1 \neq j,\ldots,X_n \neq j\}$$
$$+\mathbf{P}\{X_1 \neq i,\ldots,X_n \neq i, X_1 \neq j,\ldots,X_n \neq j\})$$
$$= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\mathbf{I}\{i \neq j\}(1-(1-p_i)^n-(1-p_j)^n+(1-p_i-p_j)^n) + \mathbf{E}N^{(n)}$$
$$\leq \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}(1-(1-p_i)^n-(1-p_j)^n+(1-p_i-p_j+p_ip_j)^n) + \mathbf{E}N^{(n)}$$
$$= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}(1-(1-p_i)^n)(1-(1-p_j)^n) + \mathbf{E}N^{(n)} = \left(\mathbf{E}N^{(n)}\right)^2 + \mathbf{E}N^{(n)}.$$

Отсюда $\mathbf{D}N^{(n)} \leq \mathbf{E}N^{(n)}$.

Доказательство завершено.

Докажем, что $\mathbf{E}N^{(n)} \to \infty$ в моделях 1 и 3, более того, получим нижние оценки скорости роста.

Лемма 4.4 $\mathbf{E}N^{(n)} > K_1 \cdot [n^{1/\alpha}]$ *в модели 1;* $\mathbf{E}N^{(n)} > K_2 \cdot [\ln n]$ *в модели 3. Здесь* $[\cdot]$ *— целая часть числа. Можно полагать* $K_1 = 1 - e^{-C(\alpha)}$*;* $K_2 = (1 - e^{-p}) \cdot \ln^{-1}\left(\frac{1}{1-p}\right)$.

Доказательство.

1) В модели 1

$$\mathbf{E}N^{(n)} = \sum_{i=1}^{\infty} \left(1 - \left(1 - C(\alpha) i^{-\alpha}\right)^n\right).$$

Срежем сумму уровнем i_* таким, что $p_{i_*} \approx C(\alpha)/n$. Так как $p_i = C(\alpha) i^{-\alpha}$, то $i_* = \left[n^{1/\alpha}\right]$.

$$\mathbf{E}N^{(n)} \;>\; \sum_{i=1}^{\left[n^{1/\alpha}\right]} \left(1 - \left(1 - C(\alpha) i^{-\alpha}\right)^n\right)$$

$$> \sum_{i=1}^{\left[n^{1/\alpha}\right]} \left(1 - \left(1 - \frac{C(\alpha)}{n}\right)^n\right) \;>\; [n^{1/\alpha}] \left(1 - e^{-C(\alpha)}\right).$$

2) В модели 3

$$\mathbf{E}N^{(n)} = \sum_{i=1}^{\infty} (1 - (1 - p(1-p)^{i-1})^n).$$

Срежем сумму уровнем i_* таким, что $p_{i_*} \approx p/n$. Так как $p_i = p(1-p)^{i-1}$, то $i_* = \left[\ln n \cdot \ln^{-1}\left(\frac{1}{1-p}\right)\right] + 1$.

$$\mathbf{E}N^{(n)} \;>\; \sum_{i=1}^{\left[\ln n \cdot \ln^{-1}\left(\frac{1}{1-p}\right)\right]+1} \left(1 - \left(1 - \frac{p}{n}\right)^n\right)$$

$$>\; [\ln n] \cdot \ln^{-1}\left(\frac{1}{1-p}\right) \cdot (1 - e^{-p}).$$

Доказательство завершено.

Получим верхние оценки скорости роста $\mathbf{E}N^{(n)}$.

Лемма 4.5 *Для любого $\delta > 1$ существуют K_1^*, K_2^* такие, что*
$\mathbf{E}N^{(n)} \le n^{\delta/(\alpha-1)} + K_1^*$ *в модели 1;*
$\mathbf{E}N^{(n)} \le \ln n \cdot \delta \cdot \ln^{-1}(1-p)^{-1} + K_2^*$ *в модели 3.*

Доказательство.

1) В модели 1 положим $i_+ = [n^{\delta/(\alpha-1)}]$. Тогда $n < i^{(\alpha-1)/\delta}$ для $i > i_+$,

$$\mathbf{E}N^{(n)} = \sum_{i=1}^{i_+} \left(1 - \left(1 - C(\alpha) i^{-\alpha}\right)^n\right) + \sum_{i=i_+ +1}^{\infty} \left(1 - \left(1 - C(\alpha) i^{-\alpha}\right)^n\right)$$

$$\le i_+ + \sum_{i=1}^{\infty} \left(1 - \left(1 - C(\alpha) i^{-\alpha}\right)^{i^{(\alpha-1)/\delta}}\right).$$

Так как

$$\left(1 - \left(1 - C(\alpha) i^{-\alpha}\right)^{i^{(\alpha-1)/\delta}}\right) \sim C(\alpha) i^{-\alpha+(\alpha-1)/\delta}$$

при $i \to \infty$, и $\alpha - (\alpha-1)/\delta > 1$ при $\delta > 1$, то по предельному признаку сравнения ряд

$$\sum_{i=1}^{\infty} \left(1 - \left(1 - C(\alpha) i^{-\alpha}\right)^{i^{(\alpha-1)/\delta}}\right)$$

сходится.

2) В модели 3 положим $i_+ = [\delta \ln n \cdot \ln^{-1}(1-p)^{-1}] + 1$. Тогда $n < (1-p)^{-i_+/\delta}$ для $i > i_+$,

$$\mathbf{E}N^{(n)} < i_+ + \sum_{i=1}^{\infty} \left(1 - \left(1 - p(1-p)^i\right)^{(1-p)^{-i/\delta}}\right).$$

Так как

$$1 - \left(1 - p(1-p)^i\right)^{(1-p)^{-i/\delta}} \sim p(1-p)^i - i/\delta$$

при $i \to \infty$, то ряд

$$\sum_{i=1}^{\infty} \left(1 - \left(1 - p(1-p)^i\right)(1-p)^{-i/\delta}\right)$$

сходится.

Доказательство завершено.

Следующий элементарный факт имеет место для любого распределения (в случае конечного носителя доказательство тривиально). Для простоты формулировок мы докажем его при условии, что все $p_i > 0$.

Лемма 4.6 *Пусть* $p_i > 0$, $i = 1,\ 2,\ \ldots$ *Тогда* $\dfrac{\mathbf{E}N^{(n)}}{n} \to 0$ *при* $n \to \infty$.

Доказательство.

Для любого $\varepsilon > 0$ выберем i_0 так, что $\sum_{i=i_0+1}^{\infty} p_i < \varepsilon$. Обозначим $A = \{(\forall i \le i_0)(\exists j \le \sqrt{n}) X_j = i\}$. Тогда

$$\limsup_{n\to\infty} \frac{\mathbf{E}N^{(n)}}{n} \le \lim_{n\to\infty} \frac{[\sqrt{n}\,] + \varepsilon(n - [\sqrt{n}\,])\mathbf{P}(A) + n\mathbf{P}(\bar{A})}{n} = \varepsilon.$$

Доказательство завершено.

Теперь все готово для того, чтобы доказать состоятельность оценок в моделях 1 и 3 (для модели 2 будет отдельно доказан более общий факт).

Обозначим $\mathbf{E}N^{(n)} = m_n(\theta)$, $\theta \in \Theta$. Для модели 1 полагаем $\theta = \alpha$, $\Theta = (1;\ \infty)$. Для модели 2 полагаем $\theta = M$, $\Theta = \{1,\ 2,\ \ldots\}$. Для модели 3 полагаем $\theta = p$, $\Theta = (0;\ 1)$.

Так как $m_n(\theta)$ — строго монотонная функция при $\theta \in \Theta$, то определена оценка $\tilde{\theta}_n$ равенством $N^{(n)} = m_n(\tilde{\theta}_n)$.

Теорема 4.2 *Оценка* $\tilde{\theta}_n$ *состоятельна в моделях 1 и 3.*

Доказательство.

Так как (в силу леммы 4.3) $\mathbf{D}m_n(\tilde{\theta}_n) \le \mathbf{E}m_n(\tilde{\theta}_n) = m_n(\theta)$, а $m_n(\theta) = o(n)$ при $n \to \infty$ в моделях 1 и 3 в силу леммы 4.6, то для любой функции g такой, что $g(m_n(\theta)) \to 0$ при $m_n(\theta) \to \infty$, по неравенству Чебышева

$$\mathbf{P}\left\{\sqrt{m_n(\theta)g(m_n(\theta))}\left|\frac{m_n(\tilde{\theta_n})}{m_n(\theta)} - 1\right| > \varepsilon\right\} \le \frac{g(m_n(\theta)}{\varepsilon^2} \to 0$$

в моделях 1, 3, так как $m_n(\theta) \to \infty$ в силу леммы 4.4.

Отсюда следует сходимость $\tilde{\theta}_n \to_p \theta$, если для любых $\theta \ne \theta{+}h \in \Theta$ существуют n_0 и ε такие, что для всех $n \ge n_0$ выполнено неравенство

$$m_n(\theta) - m_n(\theta + h) \ge \varepsilon\sqrt{m_n(\theta)/g(m_n(\theta))}.$$

1) Для модели 1 положим $0 < L_1 < L_2 < \infty$.

$$\begin{aligned} & m_n(\alpha) - m_n(\alpha + h) \\ > & \sum_{i=[L_1 n^{1/\alpha}]+1}^{[L_2 n^{1/\alpha}]} \left((1 - C(\alpha + h)i^{-\alpha-h})^n - (1 - C(\alpha)i^{-\alpha})^n\right) \\ > & \sum_{i=[L_1 n^{1/\alpha}]+1}^{[L_2 n^{1/\alpha}]} \left(\left(1 - \frac{C(\alpha + h)}{\left(L_2 n^{1/\alpha}\right)^{\alpha+h}}\right)^n - \left(1 - \frac{C(\alpha)}{{L_1}^{\alpha} n}\right)^n\right) \\ \sim & \ (L_2 - L_1)n^{1/\alpha}\left(1 - e^{-C(\alpha)/L_1}\right). \end{aligned}$$

Следовательно, можно выбрать $\varepsilon > 0$ так, что для всех достаточно больших n

$$m_n(\alpha) - m_n(\alpha + h) > \varepsilon n^{1/\alpha}.$$

Покажем, что можно выбрать g так, что

$$n^{1/\alpha} \ge \sqrt{m_n(\alpha)/g(m_n(\alpha))}.$$

При $\alpha \in (1;\ 2]$ достаточно положить $g(m_n(\alpha)) = m_n(\alpha)/n$.

При $\alpha > 2$ положим $g(t) = t^{-q}$. Найдем $q > 0$ из условия

$$n^{1/\alpha} \ge \sqrt{m_n(\alpha)/g(m_n(\alpha))} = {m_n}^{(1+q)/2}.$$

Так как $m_n(\alpha) \le n^{\delta/(\alpha-1)} + K_1^*$ в силу леммы 4.5, то из уравнения

$$\frac{1}{\alpha} = \frac{1+q}{2} \cdot \frac{\delta}{\alpha - 1}$$

получаем $q = \dfrac{2\alpha - 2 - \alpha\delta}{\alpha\delta}$, причем выбираем $\delta \in (1;\ 2(1-1/\alpha))$ для выполнения условия $q > 0$.

2) В модели 3 положим

$$L^- = -\frac{\ln n}{\ln(1-p)}, \qquad L^+ = -\frac{\ln n}{\ln(1-p-h/2)}.$$

Тогда

$$m_n(p) - m_n(p+h) >$$
$$\sum_{L^- < i < L^+} \left(\left(1 - (p+h)(1-p-h)^i\right)^n - \left(1 - p(1-p)^i\right)^n\right).$$

Так как

$$\lim_{n\to\infty} \inf_{L^- < i < L^+} \left(1 - (p+h)(1-p-h)^i\right)^n = 1,$$
$$\lim_{n\to\infty} \sup_{L^- < i < L^+} \left(1 - p(1-p)^i\right)^n \le e^{-1},$$

то для любого $h > 0$ существуют n_0', $\Delta > 0$ такие, что для всех $n > n_0'$ выполнено $m_n(p) - m_n(p+h) > (L^+ - L^-)\Delta > K_0(h)\ln n$, где $K_0(h) > 0$ при $h > 0$.

Следовательно, для любых $h > 0$, $\delta > 1$, $\beta \in (1/2;\ 1)$ существуют $\varepsilon > 0$, $n_0 \ge n_0'$ такие, что для всех $n > n_0$

$$m_n(p) - m_n(p+h) > \varepsilon\left(\ln n \cdot \delta \cdot \ln^{-1}(1-p)^{-1} + K_2^*\right)^\beta \ge \varepsilon m_n^\beta(p)$$

согласно лемме 4.5. Поэтому достаточно положить

$$g(m_n(p)) = m_n^{1-2\beta}(p).$$

Доказательство завершено.

В модели 2 состоятельность оценки $\tilde{M}_n$ очевидна. Мы докажем несколько более общий факт для схемы серий: $M = M_n$ является функцией от n. Доказательство наследует ряд идей из учебника Феллера ([88], том 1, гл. 4, §2).

Теорема 4.3 *Пусть* $M = M_n$, $n \to \infty$, $\varepsilon > 0$. *Если*

$$(1+\varepsilon)M_n \ln^2 M_n \le n,$$

то оценка $\tilde{M}_n$ *состоятельна, то есть*

$$\tilde{M}_n - M_n \to_p 0.$$

Доказательство.

Требуется доказать, что $\mathbf{P}\{N^{(n)} = M_n\} \to 1$ и $\tilde{M}_n - N^{(n)} \to 0$ почти наверное. Справедливо следующее неравенство:

$$\begin{aligned}
\mathbf{P}\{N^{(n)} = M_n\} \;\ge\; & \big(1 - \mathbf{P}\{X_1 \ne 1,\ \dots\ ,\ X_{[n/M_n]} \ne 1\}\big) \\
& \cdot \big(1 - \mathbf{P}\{X_{[n/M_n]+1} \ne 2,\ \dots\ ,\ X_{2[n/M_n]} \ne 2\}\big) \\
\cdot \dots \cdot & \big(1 - \mathbf{P}\{X_{(M_n-1)[n/M_n]+1} \ne M_n,\ \dots\ ,\ X_{M_n[n/M_n]} \ne M_n\}\big) \\
= \;& \Big(1 - (1 - S(M_n))^{[n/M_n]}\Big) \cdot \dots \cdot \left(1 - \left(1 - \frac{S(M_n)}{M_n}\right)^{[n/M_n]}\right) \\
\ge \;& \Big(1 - (1 - S(M_n))^{n/M_n}\Big)^{M_n} \;\sim\; \left(1 - \left(1 - \frac{1}{\ln M_n}\right)^{(1+\varepsilon)\ln^2 M_n}\right)^{M_n} \\
\sim \;& \big(1 - e^{-(1+\varepsilon)\ln M_n}\big)^{M_n} \;=\; \big(1 - M_n^{-(1+\varepsilon)}\big)^{M_n} \;\sim\; e^{-M_n^{-\varepsilon}} \;\to\; 1.
\end{aligned}$$

Так как $1 \le N^{(n)} \le M_n$ п. н., то $|N^{(n)} - \tilde{M}_n| \to 0$ п. н., если

$$\max_{1 \le k \le M_n} |\sum_{i=1}^{k} (1 - (1-p_i)^n) - k| \to 0.$$

Достаточно доказать, что

$$\sum_{i=1}^{M_n} (1 - (1-p_i)^n) - M_n \to 0.$$

Данное выражение неположительно. Получим нижнюю оценку.

$$\begin{aligned}
\sum_{i=1}^{M_n} (1 - (1-p_i)^n) - M_n \;\ge\; & \sum_{i=1}^{M_n} \left(1 - \left(1 - \frac{S(M_n)}{M_n}\right)^n\right) - M_n \\
\sim \; M_n \left(1 - \exp\left(-\frac{n}{M_n \ln M_n}\right)\right) - M_n \;\sim\; & \exp\left(\ln M_n - \frac{n}{M_n \ln M_n}\right) \to 0.
\end{aligned}$$

Доказательство завершено.

4.2.2 Функциональная центральная предельная теорема

Обозначим

$$\alpha(x) = \max\{j|\ p_j \geq 1/x\}.$$

Следуя [132], будем предполагать, что функция $\alpha(x)$ правильно меняется на бесконечности, т. е. $\alpha(x) = x^\theta L(x)$, и $\theta \in (0,1)$, где $L(x)$ — медленно меняющаяся функция при $x \to \infty$ (см. обсуждение этого условия в [132]).

Это предположение включает в себя модель Мандельброта с бесконечным носителем.

Опишем поведение всей траектории числа разных слов и числа слов, встретившихся фиксированное число раз, с помощью функциональной центральной предельной теоремы.

Обозначим $N_k^{*(n)} = \sum_{j=k}^{\infty} N_k^{(n)}$ — число слов, встретившихся не менее чем k раз.

Для любых $t \in [0,1],\ k \geq 1$ обозначим

$$Y_{n,k}^*(t) = \frac{N_k^{*([nt])} - \mathbf{E}N_k^{*([nt])}}{(\alpha(n))^{1/2}}, \qquad Y_{n,k}(t) = \frac{N_k^{([nt])} - \mathbf{E}N_k^{([nt])}}{(\alpha(n))^{1/2}}.$$

Обозначим $K_{k,\theta} = \theta\Gamma(k-\theta)$ для $k \geq 0$. В частности, $K_{0,\theta} = -\Gamma(1-\theta)$.

В [171] доказана следующая теорема.

Теорема 4.4 *Пусть $\theta \in (0,1)$, $\nu \geq 1$ — целое. Тогда процесс $\left(Y_{n,1}^*(t), \ldots, Y_{n,\nu}^*(t),\ 0 \leq t \leq 1\right)$ сходится слабо в равномерной метрике в $D(0,1)$ к ν-мерному гауссовскому процессу с нулевым математическим ожиданием и ковариационной функцией $(c_{ij}^*(\tau,t))_{i,j=1}^{\nu}$: for $\tau \leq t$, $i,j \in \{1,\ldots,\nu\}$ (полагая $0^0 = 1$)*

$$c_{ij}^*(\tau,t) = \begin{cases} \sum\limits_{s=0}^{i-1} \sum\limits_{m=0}^{j-s-1} \frac{\tau^s (t-\tau)^m K_{m+s,\theta}}{t^{m+s-\theta} s! m!} - \sum\limits_{s=0}^{i-1} \sum\limits_{m=0}^{j-1} \frac{\tau^s t^m K_{m+s,\theta}}{(t+\tau)^{m+s-\theta} s! m!}, & i < j; \\ t^\theta \sum\limits_{m=0}^{j-1} \frac{K_{m,\theta}}{m!} - \sum\limits_{s=0}^{i-1} \sum\limits_{m=0}^{j-1} \frac{\tau^s t^m K_{m+s,\theta}}{(t+\tau)^{m+s-\theta} s! m!}, & i \geq j; \end{cases}$$

$c^*_{ij}(\tau, t) = c^*_{ji}(t, \tau)$.

4.2.3 Анализ соответствия текстов моделям

Если вероятностная модель адекватна реальным данным, а оценка параметра в данной модели состоятельна, то оценка параметра с ростом объема выборки должна сходиться по вероятности к константе — истинному значению параметра. Так как для рассматриваемых моделей состоятельность доказана теоремами 4.2, 4.3, то отсутствие такой сходимости свидетельствует о неадекватности модели.

Вычисление параметров реализовано в виде пакета программ на языке C++.

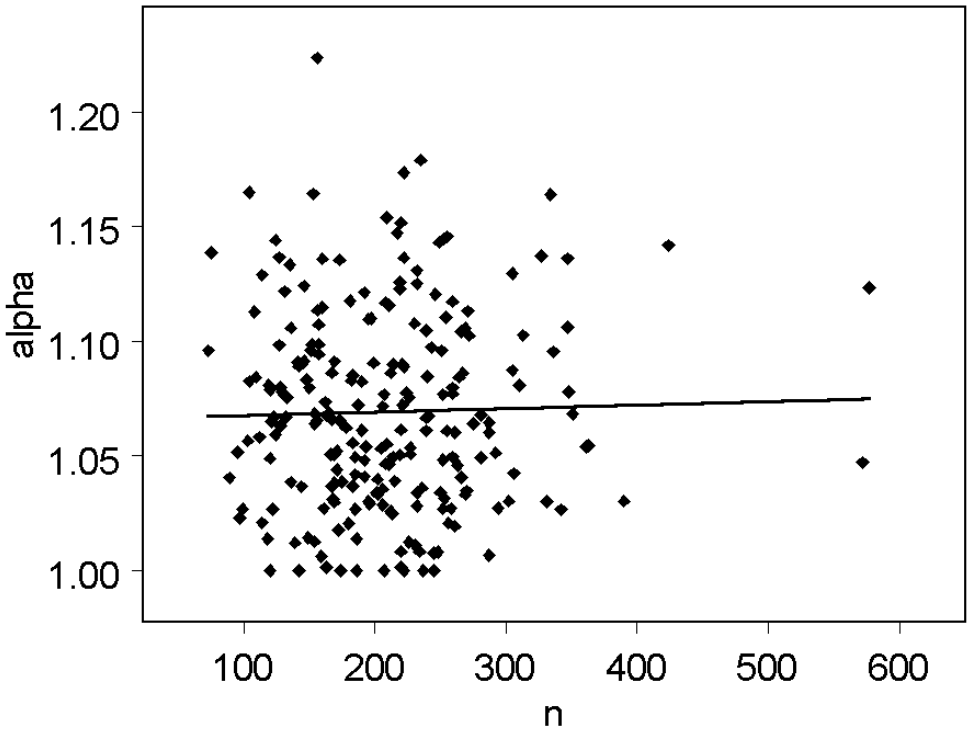

Рис. 4.1.

На рис. 1–3 показаны графики зависимости оценок для параметров α, $\lg M$, p от объема текста в моделях 1–3 соответственно, полученные по выборке из 223 текстов поэта М. Щербакова [102].

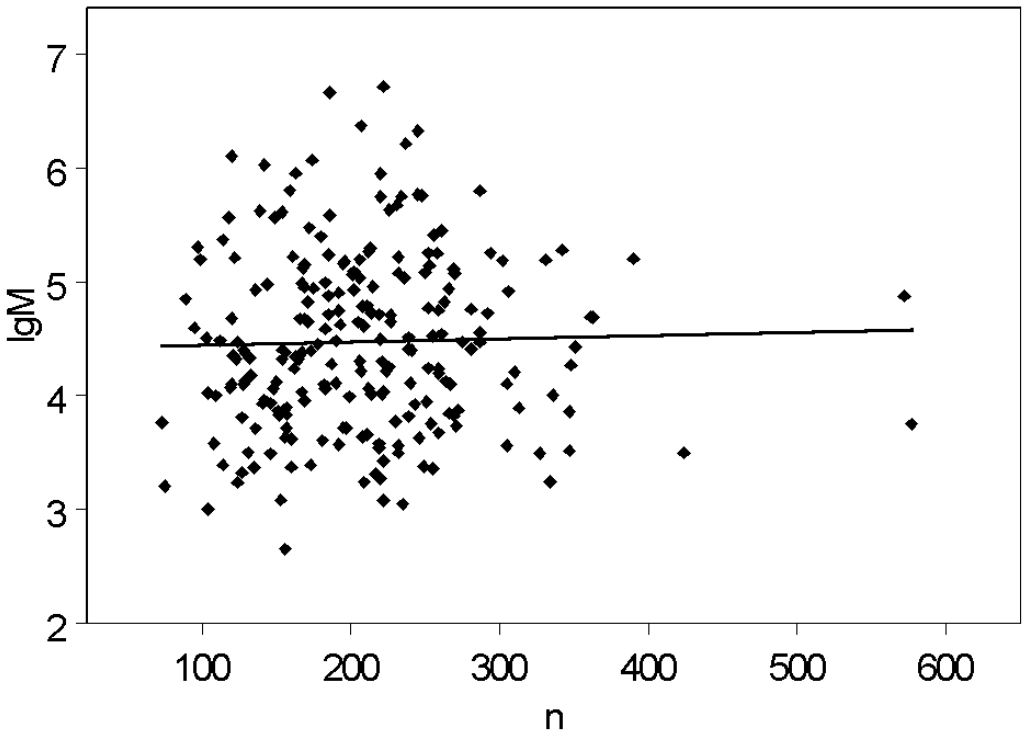

Рис. 4.2.

На графиках показаны также прямые линейной регрессии параметров на объем текста n. Для адекватной модели эта прямая должна быть почти параллельной оси абсцисс.

Лучше всего это условие выполняется для модели 2, почти столь же мал наклон к оси абсцисс для модели 1, а модель 3 оказывается неадекватной: при $n > 550$ уравнение линейной регрессии дает даже недопустимые значения параметра p.

Другой способ проверки пригодности моделей состоит в нахожде-

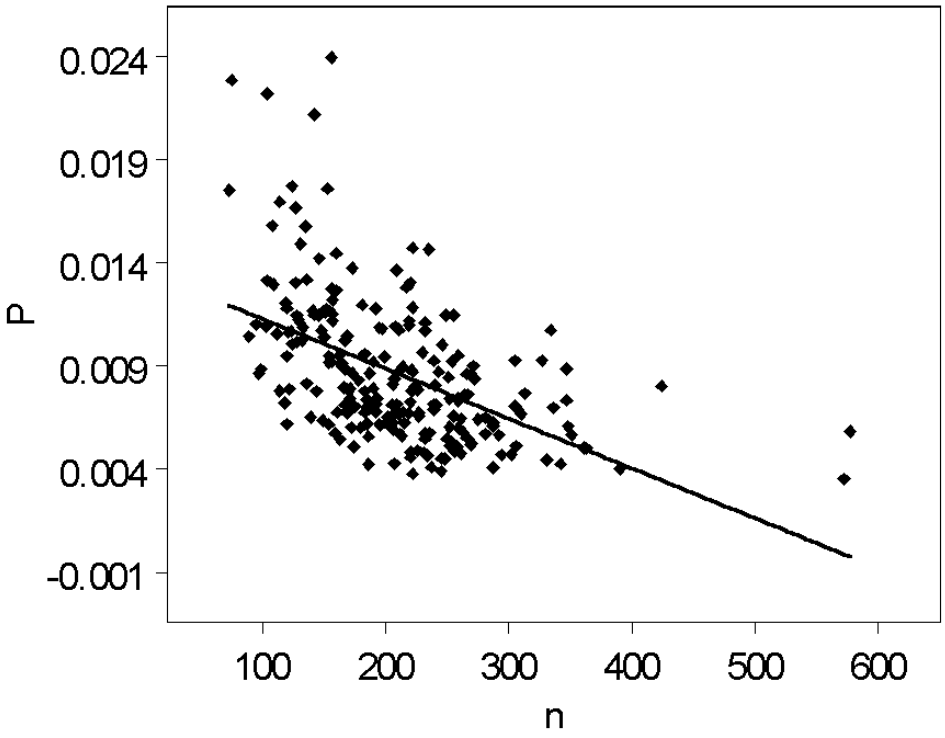

Рис. 4.3.

нии реально достигнутого уровня значимости (РДУЗ) R_θ. Гипотеза о том, что выборка удовлетворяет модели 1, 2, 3, проверяется с помощью статистики $N_1^{(n)}$ — числа слов, встретившихся в тексте ровно 1 раз. Обозначим через ξ случайную величину, имеющую распределение $N_1^{(n)}$ с параметром $\tilde{\theta}$. Пусть $\varepsilon \in (0;\ 1]$; C_1, C_2 — левая и правая квантили уровня $\varepsilon/2$ и $1-\varepsilon/2$ распределения ξ:

$$C_1 = \min\{i : \mathbf{P}\{\xi \le i\} \ge \varepsilon/2\}; \qquad C_2 = \max\{i : \mathbf{P}\{\xi \ge i\} \ge \varepsilon/2\}.$$

Рассматриваются гипотезы H_1: текст образован последовательностью независимых случайных величин, имеющих распределение в соответствии с моделью k (k может принимать значения 1, 2, 3); H_2 — отрицание гипотезы H_1. Используется следующий критерий:

если $N_1^{(n)} \in (C_1;\ C_2)$, то H_1 принимается;
если $N_1^{(n)} \in \{C_1;\ C_2\}$, то H_1 принимается с вероятностью $1/2$;
если $N_1^{(n)} \notin [C_1;\ C_2]$, то H_2 принимается.

Тогда $\mathbf{P}\{\xi < C_1\} + \frac{1}{2}\mathbf{P}\{\xi = C_1\}$ и $\mathbf{P}\{\xi > C_2\} + \frac{1}{2}\mathbf{P}\{\xi = C_2\}$ лежат в пределах $\frac{\varepsilon}{2} \pm \frac{1}{2}\mathbf{P}\{\xi = C_1\}$ и $\frac{\varepsilon}{2} \pm \frac{1}{2}\mathbf{P}\{\xi = C_2\}$ соответственно, а уровень критерия находится в пределах $\varepsilon \pm \frac{1}{2}\mathbf{P}\{\xi \in \{C_1;\ C_2\}\}$. Отметим, что критерий не имеет ни точного уровня ε, ни даже асимптотического, так как при $n \to \infty$ и фиксированном M во второй модели $\xi \to 0$ п.н., а в третьей модели $\mathbf{E}\xi$ равномерно ограничено ([133], теоремы 2, 3). Однако построенный критерий имеет уровень ε «в среднем» в том смысле, что полусумма верхней и нижней границ уровня значимости равняется ε. Полагая «в среднем»

$$\frac{\varepsilon}{2} = \mathbf{P}\{\xi < C_1\} + \frac{1}{2}\mathbf{P}\{\xi = C_1\};$$

$$\frac{\varepsilon}{2} = \mathbf{P}\{\xi > C_2\} + \frac{1}{2}\mathbf{P}\{\xi = C_2\};$$

получаем

$$\varepsilon = 2\min\left\{\mathbf{P}\{\xi < C_1\} + \frac{1}{2}\mathbf{P}\{\xi = C_1\};\quad \mathbf{P}\{\xi > C_2\} + \frac{1}{2}\mathbf{P}\{\xi = C_2\}\right\}.$$

Находим реально достигаемый уровень значимости $R_\theta = \varepsilon(N_{1,\cdot}^{(n)})$, заменяя C_1 и C_2 на вычисляемое по тексту значение $N_{1,\cdot}$. Получаем

$$R_\theta = 2\min\{\mathbf{P}\{\xi < N_{1,\cdot}\};\quad \mathbf{P}\{\xi > N_{1,\cdot}\}\} + \mathbf{P}\{\xi = N_{1,\cdot}\}. \qquad (37)$$

Отметим вновь, что критерий не имеет уровня ε и не является состоятельным, а статистики N_0 и N_1 зависимы. Кроме того, вычисления по формуле (37) проводятся моделированием случайной величины ξ (вероятность заменяется на частоту). Тем не менее, так введенный реально достигаемый уровень значимости позволяет оценивать адекватность моделей. Так, в среднем для рассматриваемой

выборки из [102] получен РДУЗ 0,53 для модели 1, 0,49 для модели 2, 0,12 для модели 3.

На основании проведенного анализа модель 3 признана нами неадекватной и в дальнейшем использоваться не будет.

Проведен анализ зависимости параметров моделей 1 и 2 от языка и года написания произведения.

Далее приведены данные по 2 монологам Гамлета и стихотворению Льюиса Кэрролла. 1-му монологу Гамлета в оригинале соответствуют $\tilde{\alpha} = 1.11$ и $\tilde{M} = 5221$, 2-му — $\tilde{\alpha} = 1.12$ и $\tilde{M} = 3627$. В переводе М. Лозинского $\tilde{\alpha} = 1.02$, $\tilde{M} = 240489$ и $\tilde{\alpha} = 1.00$, $\tilde{M} = 1108307$ для 1-го и 2-го монологов соответственно.

«Jabberwocky» Кэрролла переведено на многие языки, причем на русский язык его переводили разные авторы под совершенно разными названиями [183], [184]. В оригинале $\tilde{\alpha} = 1.14$, $\tilde{M} = 2021$. В переводе на французский (F. L. Warrin) $\tilde{\alpha} = 1.08$, $\tilde{M} = 11567$, в переводе на немецкий (R. Scott) $\tilde{\alpha} = 1.02$, $\tilde{M} = 169658$. В переводе на русский Д. Орловской стихотворение называется «Бармаглот», $\tilde{\alpha} = 1.06$, $\tilde{M} = 27721$; в переводе Вл. Орла это произведение называется «Умзара Зум», $\tilde{\alpha} = 1.04$, $\tilde{M} = 94258$.

Анализ показывает устойчивую зависимость параметров от языка: наибольшее значение α (и, соответственно, наименьшее значение M) для английского языка, затем следуют французский, русский и немецкий.

Зависимость от года написания различается у разных авторов. На рис. 4.4, 4.5 показаны графики для 30 стихотворений М. Цветаевой [93] и 183 стихотворений М. Щербакова [102].

Итак, утверждение «язык данного автора с годами становится проще (сложнее)» получает точное количественное выражение.

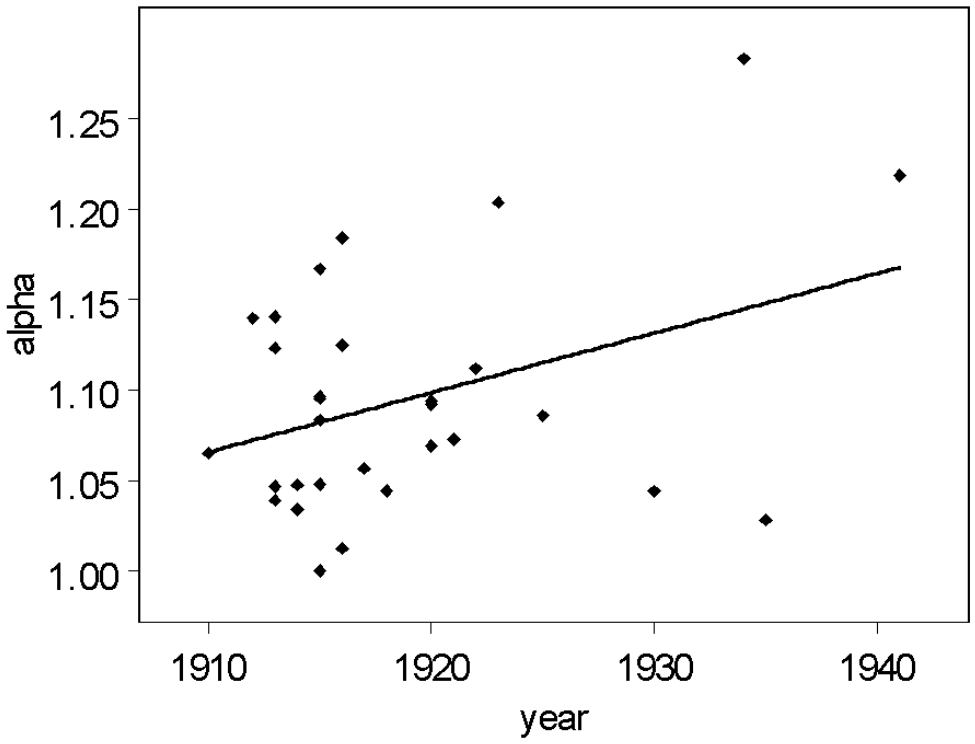

Рис. 4.4.

4.3 Применение статистического критерия к анализу однородности текста

Двигающим мотивом к этой работе послужила процедура написания рефератов студентами в век Интернета: зачастую реферат студента состоит просто в соединении двух или нескольких текстов, найденных с помощью поисковой системы. При такой процедуре «творчества» определить интеллектуальный вклад студента не представляется возможным. Поэтому важен алгоритм, позволяющий быстро выявить наличие разнородных фрагментов текста.

Для анализа однородности текста необходимо определить способ, с помощью которого тексту сопоставляется последовательность чи-

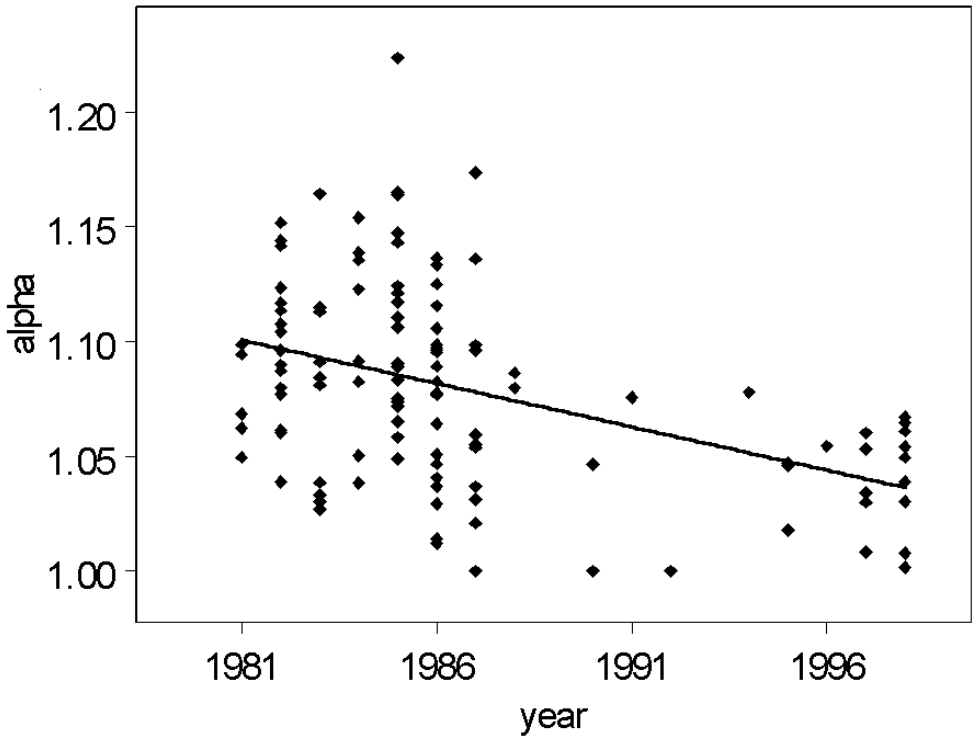

Рис. 4.5.

сел. Нами была разработана программа, которая сопоставляет тексту последовательность индикаторов вхождения слов текста в некоторый словарь.

Использовалось следующее техническое определение слова: слово русского языка — это последовательность русских букв и соединяющих их дефисов между любыми двумя символами, не являющимися русскими буквами. При этом словом не считается последовательность букв, начинающаяся или заканчивающаяся дефисом. Например, «сигма-алгебра» — это одно слово, «σ-алгебра» — вообще не слово, а «сигма - алгебра» — два слова (дефис, отделенный пробелами с двух сторон, считается за тире). Важным этапом исследова-

ния является выбор словаря. Были испытаны различные варианты словарей. Выяснилось, что наилучшие результаты различения разнородных и однородных текстов получаются при использовании в качестве словаря авторского инварианта, введенного В. П. Фоменко и Т. Г. Фоменко в [89]. Это набор служебных слов, состоящий из 3 частей:

1) предлоги: в, на, с, за, к, по, из, у, от, для, во, без, до, о, через, со, при, про, об, ко, над, из-за, из-под, под;

2) союзы: и, что, но, а, да, хотя, когда, чтобы, если, тоже, или, то есть, зато, будто;

3) частицы: не, как, же, даже, бы, ли, только, вот, то, ни, лишь, ведь, вон, то-есть, нибудь, уже, либо.

В связи с введенным нами техническим определением слова мы исключили из авторского инварианта союз «то есть».

Всего было взято для исследования 25 текстов:

1. Илья Ильф и Евгений Петров. Двенадцать стульев 12stul.txt
2. Алексей Толстой. Аэлита aelit.txt
3. Аркадий и Борис Стругацкие. Трудно быть богом be-god.txt
4. Виктор Пелевин. Чапаев и Пустота chapaev.txt
5. Михаил Булгаков. Собачье сердце dogheart.txt
6. Ник Перумов. Эльфийский клинок (часть 1) elf-klin.txt
7. Алексей Толстой. Гиперболоид инженера Гарина giper.txt
8. Аркадий и Борис Стругацкие. Град обреченный grad.txt
9. Борис Пастернак. Охранная грамота gramota.txt
10. Виктор Пелевин. Жизнь насекомых insec.txt

11. Михаил Веллер. Все о жизни life.txt
12. Владимир Набоков. Лолита lolita.txt
13. Владимир Набоков. Защита Лужина lughin.txt
14. Михаил Булгаков. Мастер и Маргарита master.txt
15. Николай Носов. Незнайка в Солнечном городе nezn-sun.txt
16. Николай Носов. Незнайка на Луне nezn-moon.txt
17. Иван Ефремов. На краю Ойкумены ojkum.txt
18. Виктор Пелевин. Поколение "П"pel-g.txt
19. Иван Ефремов. Туманность Андромеды tuman.txt
20. Александр Волков. Урфин Джюс и его деревянные солдаты urfin.txt
21. Александр Волков. Волшебник Изумрудного города volsh.txt
22. Андрей Сергеевич Некрасов. Приключения капитана Врунгеля vrungel.txt
23. Алексей Волков, Андрей Новиков. Мир спящего колдуна warlock.txt
24. Борис Пастернак. Доктор Живаго zhiwago.txt
25. Михаил Веллер. Приключения майора Звягина zwqgin.txt

Выбранные тексты исследовались поодиночке и в попарных комбинациях, полученных приписыванием одного текста к другому. Получилось $25 \times 24 = 600$ попарных комбинаций текстов, в том числе $3! + 9 \times 2 = 24$ комбинаций текстов одного автора и 576 комбинаций текстов разных авторов.

Для этих текстов были построены эмпирические мосты Z_n с помощью словаря авторского инварианта. Затем вычислялись максимальные по модулю отклонения эмпирического моста $|M_n| = \sup_{t\in[0;\ 1]} |Z_n(t)|$, и с помощью распределения Колмогорова отыскивался достигаемый уровень значимости

$$\varepsilon^* = 2\sum_{k=1}^{\infty}(-1)^{k+1}e^{-2k^2|M_n|^2}.$$

Задача состояла в том, чтобы научиться различать:

1) одно произведение одного автора от двух произведений разных авторов;

2) два произведения одного автора от двух произведений разных авторов.

Для этого нужно выбрать уровень значимости ε и соответствующее предельное значение M.

Отметим, что для произведения одного автора значения M не превосходят $3,4758$, что соответствует достигаемому уровню значимости $\varepsilon^* = 6,42 \cdot 10^{-11}$. Низкий достигаемый уровень значимости говорит о неполной адекватности модели выборки для описания появлений слов из выбранного словаря в тексте. Рассмотрим специфику текстов, давших наибольшие значения M.

Значение $M = 3,4758$ относится к произведению М. Веллера «Все о жизни». Это произведение лежит за рамками чисто литературного жанра и является в значительной мере философским, что и обуславливает существенную неоднородность. Как мы заметим ниже, такие значения отклонений эмпирического моста характерны для собраний сочинений одного автора.

Значения $M = 3,0324$ и $M = 2,5002$ относятся к произведениям «Двенадцать стульев» и «Град обреченный». Отметим, что каждый из этих романов написан двумя авторами, что, по-видимому, и яв-

ляется причиной неоднородности.

Значение $M = 2,5758$ достигается романом «Доктор Живаго»; неоднородность здесь обусловлена наличием стихов в конце романа, то есть, как и в случае «Все о жизни», выходом за границы жанра.

Для всех остальных произведений значения отклонений не превосходят $1,8658$, что соответствует достигаемому уровню значимости $0,0019$. Отметим, что наилучшая однородность достигается для произведений В. Набокова и сравнительно небольших повестей.

Приведем пример эмпирического моста для однородных данных (рис. 3.6).

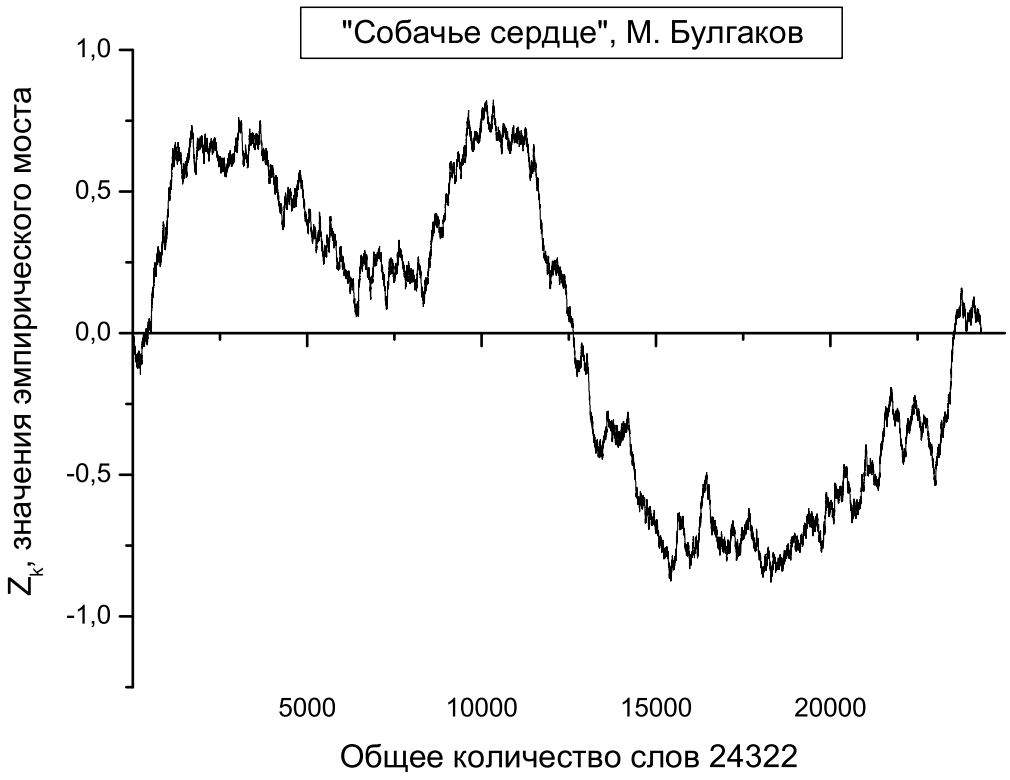

Рис. 4.6. График эмпирического моста для индикаторов служебных слов в повести М. Булгакова «Собачье сердце»

Табл. 4.1. Статистические характеристики разладки в текстах одного автора.

произведение	ν_n	T_n	M_n	n	ε^*
12stul	0,1757	17299	3,0324	81555	2,06013E-08
aelit	0,1616	24585	1,6169	40583	0,01072196
be-god	0,1793	19118	1,4080	48790	0,03793235
chapaev	0,1989	41149	1,3976	90712	0,040225523
dogheart	0,1874	18299	0,8796	24322	0,421550786
elf-klin	0,1948	42600	1,4521	135904	0,029485946
giper	0,1716	20425	1,2353	71760	0,094508689
grad	0,1915	35301	2,5002	107735	7,43849E-06
gramota	0,2163	14929	0,6578	26385	0,779863493
insec	0,2044	30248	1,5020	44423	0,021958702
life	0,2056	50536	3,4758	204368	6,41698E-11
lolita	0,2009	23144	0,7810	105601	0,575354971
luzhin	0,2018	19733	0,7643	51549	0,603220011
master	0,2169	93267	1,5886	112655	0,012853158
nezn-moon	0,2162	60006	1,2138	111824	0,105005271
nezn-sun	0,2173	51916	1,4898	63970	0,02361497
ojkum	0,1744	50047	1,6405	101940	0,009193475
pel-g	0,1935	31973	1,2439	66370	0,090576547
tuman	0,1698	51314	1,8658	85457	0,001893564
urfin	0,1892	31255	0,8909	37286	0,405414339
volsh	0,189	21264	1,6923	31800	0,00651071
vrungel	0,2051	23940	0,8794	34786	0,421761038
warlock	0,2065	16683	0,8966	32251	0,397389406
zhiwago	0,1994	62553	2,5758	152474	3,45377E-06
zwqgin	0,1833	47254	1,3930	112239	0,041259618

Анализ 576 попарных комбинаций текстов разных авторов дал следующие результаты. Только для 99 комбинаций текстов достигнутый уровень значимости превосходит 0,001 (эти комбинации приведены в таблице). Из них для 65 достигнутый уровень значимости больше 0,01, в том числе для 33 больше 0,1.

Табл. 4.2. Статистические характеристики разладки в текстах двух разных авторов.

произведения	ν_n	T_n	M_n	n	ε^*
warlock+vrungel	0,2058	40961	0,7024	67036	0,7073
dogheart+urfin	0,1885	18299	0,7631	61607	0,6052
vrungel+warlock	0,2058	23940	0,7779	67036	0,5804
lolita+chapaev	0,2	94656	0,7888	196312	0,5625
vrungel+lolita	0,2019	43563	0,8326	140386	0,4921
insec+lolita	0,2018	43445	0,8336	150023	0,4905
urfin+pel-g	0,192	37310	0,8416	103655	0,4781
insec+luzhin	0,203	64155	0,8588	95971	0,4521
luzhin+chapaev	0,1999	65359	0,8666	142260	0,4405
lolita+vrungel	0,2019	106445	0,8769	140386	0,4254
urfin+dogheart	0,1885	31255	0,9399	61607	0,3401
vrungel+luzhin	0,2031	54518	0,972	86334	0,3012
master+nezn-moon	0,2166	93267	0,9945	224478	0,2760
volsh+pel-g	0,1922	30647	0,9957	98169	0,2746
luzhin+vrungel	0,2031	42205	0,9985	86334	0,2716
zwqgin+dogheart	0,184	47254	1,0198	136560	0,2494
warlock+lolita	0,2022	27779	1,022	137851	0,2472
warlock+insec	0,2053	62498	1,0355	76673	0,2339
nezn-moon+gramota	0,2164	60006	1,0718	138208	0,2008
gramota+nezn-moon	0,2164	86390	1,0814	138208	0,1927
vrungel+insec	0,2047	65033	1,0854	79208	0,1894
be-god+zwqgin	0,1822	44592	1,0994	161028	0,1782
elf-klin+pel-g	0,1945	42600	1,1	202273	0,1777
warlock+luzhin	0,2036	51983	1,1062	83799	0,1729

произведения	ν_n	T_n	M_n	n	ε^*
nezn-moon+master	0,2166	205090	1,1515	224478	0,1410
zwqgin+volsh	0,1846	90503	1,1532	144038	0,1399
lolita+warlock	0,2022	110803	1,1583	137851	0,1366
giper+tuman	0,1707	123073	1,1628	157216	0,1338
dogheart+pel-g	0,1919	23023	1,165	90691	0,1324
luzhin+warlock	0,2036	42205	1,1806	83799	0,1231
dogheart+volsh	0,1883	45585	1,1909	56121	0,1173
chapaev+lolita	0,2	41149	1,2024	196312	0,1109
gramota+nezn-sun	0,2171	78300	1,2204	90354	0,1017
insec+vrungel	0,2047	30248	1,2286	79208	0,0977
be-god+ojkum	0,176	19308	1,2294	150729	0,0973
nezn-sun+master	0,2171	157236	1,2428	176624	0,0911
elf-klin+volsh	0,1936	123968	1,2624	167703	0,0825
zwqgin+urfin	0,1848	90503	1,269	149524	0,0798
vrungel+chapaev	0,2006	39252	1,3117	125497	0,0641
luzhin+insec	0,203	57470	1,3266	95971	0,0592
lolita+insec	0,2018	111522	1,3411	150023	0,0548
chapaev+elf-klin	0,1963	79231	1,3414	226615	0,0547
elf-klin+dogheart	0,1937	123968	1,3556	160225	0,0507
elf-klin+urfin	0,1935	123968	1,3575	173189	0,0502
master+nezn-sun	0,2171	93267	1,36	176624	0,0495
nezn-sun+gramota	0,2171	51916	1,3768	90354	0,0451
ojkum+giper	0,1733	122364	1,3816	173699	0,0440
pel-g+elf-klin	0,1945	108969	1,385	202273	0,0431
pel-g+urfin	0,192	31973	1,388	103655	0,0424
giper+12stul	0,1738	89058	1,3966	153314	0,0404

произведения	ν_n	T_n	M_n	n	ε^*
insec+warlock	0,2053	30248	1,3997	76673	0,0398
chapaev+luzhin	0,1999	41149	1,408	142260	0,0379
master+gramota	0,2168	93267	1,4094	139039	0,0376
pel-g+volsh	0,1922	31973	1,4149	98169	0,0365
volsh+dogheart	0,1883	21264	1,4448	56121	0,0308
gramota+master	0,2168	119651	1,4618	139039	0,0279
warlock+chapaev	0,2009	36717	1,4853	122962	0,0243
pel-g+dogheart	0,1919	31973	1,5068	90691	0,0213
gramota+warlock	0,211	24392	1,5262	58635	0,0190
zwqgin+be-god	0,1822	47254	1,5287	161028	0,0187
dogheart+zwqgin	0,184	71575	1,5713	136560	0,0143
elf-klin+chapaev	0,1963	59028	1,5714	226615	0,0143
pel-g+zhiwago	0,1978	66752	1,5732	218843	0,0142
be-god+dogheart	0,182	44592	1,5792	73111	0,0136
warlock+gramota	0,211	33692	1,6084	58635	0,0113
dogheart+zhiwago	0,1979	23023	1,6285	176795	0,0099
luzhin+elf-klin	0,1967	50430	1,6454	187452	0,0089
tuman+giper	0,1707	51314	1,648	157216	0,0087
elf-klin+zhiwago	0,1973	135217	1,7037	288377	0,0060
volsh+zhiwago	0,1977	30647	1,7331	184273	0,0049
master+warlock	0,2145	111875	1,7367	144905	0,0048
chapaev+vrungel	0,2006	68786	1,738	125497	0,0048
ojkum+12stul	0,175	119238	1,7541	183494	0,0043
volsh+zwqgin	0,1846	41566	1,7693	144038	0,0038
vrungel+elf-klin	0,1968	34780	1,7841	170689	0,0034
giper+ojkum	0,1733	93039	1,8007	173699	0,0031

произведения	ν_n	T_n	M_n	n	ε^*
warlock+nezn-moon	0,2142	91801	1,811	144074	0,0028
zhiwago+lolita	0,2	62553	1,8127	258074	0,0028
gramota+vrungel	0,21	34416	1,8214	61170	0,0026
elf-klin+luzhin	0,1967	59028	1,8243	187452	0,0026
giper+be-god	0,1748	76634	1,8332	120549	0,0024
nezn-moon+warlock	0,2142	107211	1,8333	144074	0,0024
pel-g+luzhin	0,1972	73393	1,8348	117918	0,0024
zhiwago+insec	0,2007	62553	1,8692	196896	0,0018
urfin+grad	0,191	124449	1,8709	145020	0,0018
chapaev+zhiwago	0,1993	153264	1,8713	243185	0,0018
zhiwago+warlock	0,2007	62553	1,8783	184724	0,0017
chapaev+dogheart	0,1964	79231	1,8856	115033	0,0016
chapaev+volsh	0,1963	79231	1,8879	122511	0,0016
urfin+zhiwago	0,1976	37668	1,8889	189759	0,0016
warlock+elf-klin	0,197	32247	1,8918	168154	0,0016
chapaev+warlock	0,2009	68786	1,8941	122962	0,0015
urfin+zwqgin	0,1848	47052	1,9002	149524	0,0015
dogheart+ojkum	0,177	24115	1,9044	126261	0,0014
dogheart+be-god	0,182	43439	1,9081	73111	0,0014
gramota+insec	0,2088	27472	1,9223	70807	0,0012
lolita+elf-klin	0,1975	95508	1,9381	241504	0,0011
elf-klin+vrungel	0,1968	59028	1,9398	170689	0,0011
volsh+grad	0,1909	67100	1,9453	139534	0,0010

Отметим, что здесь отклонение эмпирического моста от нуля может принимать большие значения, вплоть до 13.

При больших значениях M программа очень точно отыскивает

границу между текстами. Так, текст «Аэлита»+«Мастер и Маргарита» делится точкой экстремума эмпирического моста на два фрагмента, первый из которых содержит лишь несколько строк из первой главы «Мастера и Маргариты».

Рассмотрим несколько способов определения критического значения ε, используемого в критерии выбора между гипотезами «текст написан одним автором» и «текст является объединением частей, принадлежащих разным авторам», или, как мы будем обозначать это более кратко, «текст однороден» и «текст неоднороден». Большое значение при выборе критического значения M и ε играют частоты ошибок первого и второго рода. Частота α_i^* ошибки i-го рода — это частота случаев, когда верная i-ая гипотеза отвергается статистическим критерием. Таким образом, для рассматриваемого критерия α_1^* — это отношение числа случаев, когда для текста одного автора значение M_n оказалось не меньше критического, к общему числу текстов 25. Аналогично, α_2^* — это отношение числа комбинаций текстов двух разных авторов, для которых M_n оказалось меньше критического, к общему числу текстов двух разных авторов, то есть к 576.

Если выбрать $M_{max} = 3,4758$ — максимальное для текстов одного автора, то $\alpha_1^* = 0$, но, как показывают подсчеты в этом случае, число текстов двух авторов, для которых значение M_n меньше, равняется 265, и частота ошибки второго рода $\alpha_2^* = 265/576 = 0,460$. Такой выбор M полезен, когда задача состоит в наиболее достоверном выявлении существенной неоднородности текста.

Табл. 4.3. Статистические характеристики разладки в двух текстах одного автора.

произведения	ν_n	T_n	M_n	n	ε^*
aelit+giper	0,1680	35709	2,2519	112342	7,87431E-05
be-god+grad	0,1877	47346	2,7534	156524	5,19998E-07
chapaev+insec	0,2007	96633	1,7162	135134	0,005531114
chapaev+pel-g	0,1966	106004	1,6702	157081	0,007553569
dogheart+master	0,2117	33497	4,1050	136976	4,61903E-15
giper+aelit	0,1680	62746	2,5383	112342	5,06749E-06
grad+be-god	0,1877	35301	2,9651	156524	4,61764E-08
gramota+zhiwago	0,2019	82610	3,6979	178858	2,65059E-12
insec+chapaev	0,2007	48889	1,3334	135134	0,057121241
insec+pel-g	0,1979	59715	2,5414	110792	4,90759E-06
life+zwqgin	0,1978	201188	7,3365	316606	3,55062E-47
lolita+luzhin	0,2011	23144	0,6787	157149	0,74641719
luzhin+lolita	0,2011	19733	0,5288	157149	0,942772469
master+dogheart	0,2117	113799	4,0042	136976	2,36791E-14
nezn-moon+nezn-sun	0,2168	60006	1,0853	175793	0,18951463
nezn-sun+nezn-moon	0,2168	51916	1,1149	175793	0,166378215
ojkum+tuman	0,1723	101292	1,3096	187396	0,064751696
pel-g+chapaev	0,1966	107518	1,7988	157081	0,003094454
pel-g+insec	0,1979	72281	2,6936	110792	9,9729E-07
tuman+ojkum	0,1723	106736	2,1736	187396	0,000157496
urfin+volsh	0,1891	58549	1,1550	69085	0,138740603
volsh+urfin	0,1891	21264	1,1218	69085	0,161325718
zhiwago+gramota	0,2019	152983	2,2754	178858	6,36613E-05
zwqgin+life	0,1978	162651	8,1311	316606	7,48248E-58

Другой подход состоит в выборе такого M, чтобы сумма частот ошибок была минимальна. Это байесовский (с равными априорными вероятностями гипотез) подход к построению статистического критерия. Расчеты показывают, что для этого в рассматриваемой задаче достаточно принять $M_b = 1,8658$, в этом случае $\alpha_1^* = 4/24 = 0,167$, $\alpha_2^* = 83/576 = 0,144$, $\alpha_1^* + \alpha_2^* = 0,311$.

Это же значение $M_{mm} = M_b$ обеспечивает и минимаксный критерий, минимизирущий максимум из α_1^* и α_2^*.

Проанализируем табл. 4.3. Отметим, что из 24 сочетаний текстов одного автора уровень $M_b = 1,8658$ превышен в 13 случаях, то есть частота ошибки первого рода при проверке гипотезы об одном авторе двух текстов против гипотезы о двух разных авторах составляет $13/24 = 0,54$. В двух случаях достигнутые величины M чрезвычайно велики (и, соответственно, достигаемые уровни значимости чрезвычайно малы): при прямом и обратном соединении произведений М. Веллера. Отметим, что причина явления в том, что «Все о жизни» является не художественным произведением, а философским трактатом, и не подчиняется закону сохранения служебных слов (в частности, междометий), характерного для художественных текстов. В остальном достигаемые уровни M не превосходят 4,1. Соответствующие достигаемые уровни значимости не меньше 10^{-15}.

Рассмотрим теперь два романа М. Шолохова.

Табл. 4.4. Статистические характеристики разладки в романах «Поднятая целина» и «Тихий Дон».

произведения	n	ν_n	T_n	M_n	ε^*
Поднятая целина	204955	0,208737	97523	6,728881	9,40152E-40
Тихий Дон	421854	0,182457	210643	9,012559	5,60831E-71

Здесь каждый из романов дает очень большие значения отклонения эмпирического моста, не характерные ни для одного произведения, ни для пары произведений одного автора. Разладка на графиках оказывается чрезвычайно четкой и характерной (рис. 4.7, 4.8). Для того, чтобы более глубоко и последовательно исследовать это явление, мы составили собрания сочинений трех авторов.

Нами были составлены следующие три электронных собрания сочинений. Собрания сочинений получены из романов, повестей и рассказов автора, приписанных друг к другу в последовательности, соответствующей биографии автора.

Федор Михайлович Достоевский. «Униженные и оскорбленные», «Преступление и наказание», «Идиот», «Бесы», «Братья Карамазовы», «Подросток».

Лев Николаевич Толстой. «Детство». «Отрочество». «Юность». «Севастопольские рассказы». «Казаки». «Война и мир». «Анна Каренина». «Воскресение».

Михаил Алексеевич Шолохов. «Тихий Дон», книги 1 и 2. «Поднятая целина», часть 1. «Тихий Дон», книги 3,4. «Судьба человека». «Поднятая целина», часть 2.

Исследовались параметры разладки в каждом из собраний сочинений целиком, затем во фрагментах собраний сочинений, полученных разрезанием текста в точках, соответствующих максимальному отклонению эмпирического моста от оси абсцисс. Фрагментам текста присваивались имена приписыванием цифр 1 и 2 к названию исход-

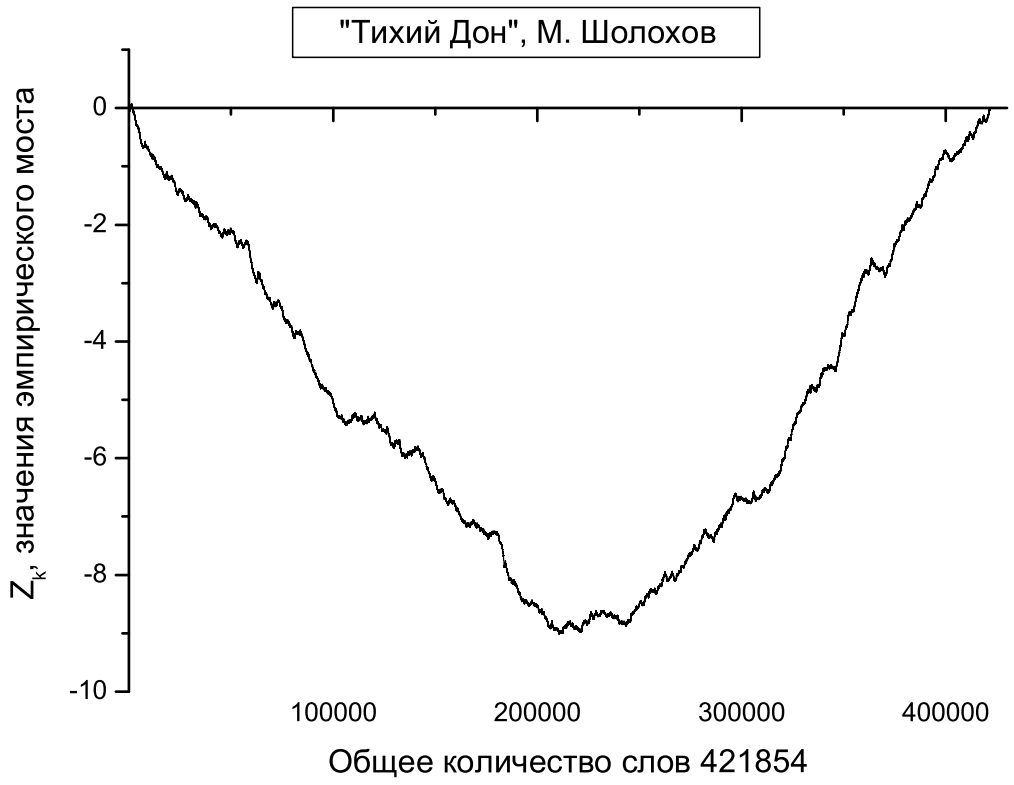

Рис. 4.7. График эмпирического моста для индикаторов служебных слов в романе «Тихий Дон»

ного текста, цифра 1 соответствовала первому фрагменту, цифра 2 — второму.

Табл. 4.5. Статистические характеристики разладки в собраниях сочинений и их фрагментах.

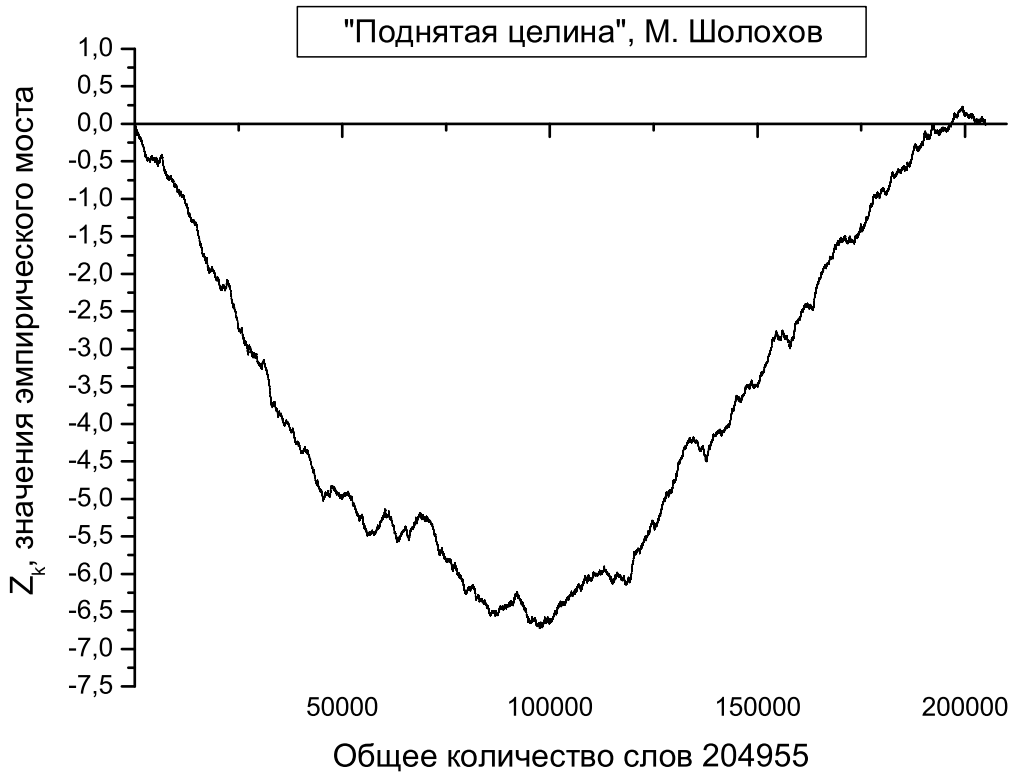

Рис. 4.8. График эмпирического моста для индикаторов служебных слов в романе «Поднятая целина»

произведения	n	ν_n	T_n	M_n	ε^*
dostoevski-all	1185242	0,238991	347125	3,480106	6,04546E-11
dostoevski-all1	347126	0,233202	95110	2,354132	3,0716E-05
dostoevski-all11	95111	0,227289	26457	1,90939	0,001362556
dostoevski-all12	252016	0,235564	71341	2,925917	7,32906E-08
dostoevski-all2	838117	0,240361	339021	2,486013	8,56888E-06
dostoevski-all21	339022	0,238375	164295	3,938874	6,6851E-14
dostoevski-all22	499096	0,241673	150422	2,671881	1,25956E-06
sholohov-all	637698	0,191609	407236	13,176486	3,1391E-151
sholohov-all1	407237	0,181841	187777	7,454041	1,09612E-48
sholohov-all11	187778	0,171735	102127	1,340005	0,055127833
sholohov-all111	102128	0,169599	58139	1,493484	0,023101834
sholohov-all112	85651	0,174329	39488	1,81604	0,002731634
sholohov-all12	219460	0,190289	91620	3,135968	5,74209E-09
sholohov-all2	230462	0,209812	109306	6,982154	9,05536E-43
sholohov-all21	109307	0,197459	46789	1,962837	0,000900725

Табл. 4.5. (окончание).

произведения	n	ν_n	T_n	M_n	ε^*
tolstoj-all	1058136	0,223217	127818	3,368991	2,76986E-10
tolstoj-all1	127819	0,23525	41641	3,218909	2,00103E-09
tolstoj-all11	41642	0,223572	25472	1,158508	0,136497001
tolstoj-all111	25473	0,227398	14108	0,776465	0,582863629
tolstoj-all112	16170	0,217506	10984	1,106582	0,172640324
tolstoj-all12	86178	0,240945	59255	1,182742	0,121865597
tolstoj-all121	59256	0,243449	35550	1,831481	0,002440632
tolstoj-all122	26923	0,235466	9035	1,510311	0,020880646
tolstoj-all2	930318	0,221937	252594	3,508171	4,08391E-11
tolstoj-all21	252595	0,216009	59322	1,856277	0,002032718
tolstoj-all211	59323	0,222403	38865	2,723605	7,20802E-07
tolstoj-all2111	38866	0,215343	21260	1,613999	0,010923504
tolstoj-all21111	21261	0,221507	5879	1,502185	0,021928378
tolstoj-all21112	17606	0,207899	9942	1,702325	0,006080468
tolstoj-all2112	20458	0,235845	15930	2,40669	1,8622E-05
tolstoj-all21121	15931	0,226745	7469	2,813355	2,66795E-07
tolstoj-all21122	4528	0,267946	1252	4,255296	3,74125E-16
tolstoj-all212	193273	0,213853	69503	3,055741	1,55072E-08
tolstoj-all22	677724	0,224477	531279	1,54569	0,016820304

Для собрания сочинений Достоевского значение $M = 3,48$ соответствует моменту биографии писателя, когда он начал играть в рулетку в Баден-Бадене. Это произошло во время работы над романом «Идиот». До этого момента характеристики его творчества более однородны ($M = 2,35$), особенно однороден фрагмент 1.1 (первые четыре главы «Униженных и оскорбленных»). После этого

момента на графике (рис. 3.9) видны чередования разных периодов. Наименьшую однородность имеет фрагмент 2.1 (конец «Идиота» и «Бесы»), наибольшую — фрагмент 2.2 («Братья Карамазовы» и «Подросток»). Как для всего собрания сочинений, так и для его фрагментов значение M не превосходит 4,26, достигаемый уровень значимости не меньше 10^{-14}.

В творчестве Толстого нет такой резкой смены частоты служебных слов, как в творчестве Достоевского. Для всего собрания сочинений $M = 3,37$, первый фрагмент включает в себя «Детство», «Отрочество», «Юность» и «Севастопольские рассказы». Дальнейшее разбиение позволяет выявить очень однородные фрагменты 1.1 («Детство») и 1.2. Разбиение фрагмента 2 позволяет выявить сравнительно однородные фрагменты 2.1 (до конца третьей части «Войны и мира») и 2.2. Отметим, что по мере дальнейшей фрагментации не происходит снижения отклонений эмпирического моста от оси абсцисс. Как для всего собрания сочинений, так и для его (даже довольно мелких) фрагментов значение M не превосходит 4,26, достигаемый уровень значимости не меньше 10^{-16}.

Эмпирический мост для собрания сочинений Шолохова разительно отличается от эмпирических мостов собраний сочинений Достоевского и Толстого. Здесь $M = 13,18$, что соответствует ничтожному уровню значимости порядка 10^{-151}. Визуально можно выделить три периода: первый (фрагмент 1.1) с частотой служебных слов около 0,17, второй (фрагменты 1.2 и 2.1) с частотой 0,19–0,20, третий (фрагмент 2.2) с частотой 0,22. Однородность каждого из фрагментов вполне соответствует тому, что наблюдалось для собраний сочинений Достоевского и Толстого, а неоднородность всего собрания сочинений в целом перешагивает все мыслимые границы. Отметим, что границы фрагментов 1.1 и 2.2 четко привязаны к соответствую-

Рис. 4.9. График эмпирического моста для индикаторов служебных слов в собрании сочинений Ф. Достоевского

щим произведениям. Конец фрагмента 1.1 привязан к концу второй книги «Тихого Дона». Из третьей книги во фрагмент вошло около 2 страниц текста, то есть с точностью до двух страниц отыскивается разладка в текстах, содержащих многие сотни страниц.

Аналогично и с фрагментом 2.2. Он включает целиком «Судьбу человека» и вторую часть «Поднятой целины», за исключением нескольких абзацев в начале «Судьбы человека», относящихся к фрагменту 2.1 и составляющих 3,5 страницы текста. Итак, проведенный анализ приводит к предположению, что под именем Шолохова скрываются два писателя: Шолохов-1 — автор двух первых книг «Тихого Дона», и Шолохов-2 — автор «Судьбы человека» и

Рис. 4.10. График эмпирического моста для индикаторов служебных слов в первом фрагменте собрания сочинений Ф. Достоевского

второй части «Поднятой целины». Остальные произведения представляются смесью фрагментов, принадлежащих этим двум авторам. В качестве альтернативы высказывается предположение о том, что Шолохов — уникальный автор, существенно изменивший частоту авторского инварианта в течение жизни.

Результаты анализа согласуются с выводами Г. Ермолаева [31], [32], Г. Кйецаа [47], О. В. Кукушкиной с соавт. [55], а также автора, известного под псевдонимом «Д» [22].

Другой пример анализа проведен на следующем материале.

Взяты следующие 10 произведений 5 авторов 20 века:

Рис. 4.11. График эмпирического моста для индикаторов служебных слов во втором фрагменте собрания сочинений Ф. Достоевского

А. Толстой. «Аэлита» и «Гиперболоид инженера Гарина»;

И. Ефремов. «На краю Ойкумены» и «Туманность Андромеды»;

А. Волков. «Волшебник Изумрудного города» и «Урфин Джюс и его деревянные солдаты»;

А. и Б. Стругацкие. «Трудно быть богом» и «Град обреченный»;

В. Пелевин. «Жизнь насекомых» и «Поколение “П” ».

Эти произведения имеют объем от 32 229 до 109 618 слов. Были проанализированы как эти 10 произведений по отдельности, так и все 90 возможных попарных комбинаций, полученных приписыва-

Рис. 4.12. График эмпирического моста для индикаторов служебных слов в собрании сочинений М. Шолохова

нием одного текста к другому.

Для этих текстов были построены эмпирические мосты Z_n с помощью как всего словаря авторского инварианта, так и отдельных его частей (предлогов, союзов, частиц). Затем вычислялись максимальные по модулю отклонения эмпирического моста $|M_n| = \sup_{t\in[0;\ 1]} |Z_n(t)|$, и с помощью распределения Колмогорова отыскивался достигаемый уровень значимости

$$\varepsilon^* = 2\sum_{k=1}^{\infty}(-1)^{k+1}e^{-2k^2|M_n|^2}.$$

Задача состояла в том, чтобы научиться различать:

Рис. 4.13. График эмпирического моста для индикаторов служебных слов во фрагменте 1.1 собрания сочинений М. Шолохова

1) одно произведение одного автора от двух произведений разных авторов;

2) два произведения одного автора от двух произведений разных авторов.

Для этого нужно выбрать уровень значимости ε и соответствующее предельное значение M.

Оказалось, что значения $|M_n|$ при использовании словаря предлогов получаются почти одинаковыми для текстов одного и двух авторов — это так называемое явление «слияния» всех авторов, описывающее не инвариант авторского стиля, а общие законы языка.

Рис. 4.14. График эмпирического моста для индикаторов служебных слов во фрагменте 1.2 собрания сочинений М. Шолохова

Напротив, при использовании словаря частиц значения $|M_n|$ меняются в широких пределах для разных классов текстов, что не позволяет разделять тексты разных авторов.

Только весь словарь авторского инварианта и словарь союзов в отдельности позволяют решать поставленные задачи.

Именно, для словаря авторского инварианта получилось, что $|M_n|$ у одного произведения одного автора колеблется в пределах от 0,71 («Жизнь насекомых») до 2,23 («Град обреченный»), соответствующие достигаемые уровни значимости 0,698 и $5 \cdot 10^{-5}$. Если исключать «Град обреченный» из рассмотрения как произведение, написанное

Рис. 4.15. График эмпирического моста для индикаторов служебных слов во фрагменте 2.1 собрания сочинений М. Шолохова

все же двумя авторами и существенно неоднородное, то следующим за ним будет «Поколение “П”» со значениями $|M_n| = 1,82$ и $\varepsilon^* = 0,0024$. В связи с этим выглядит логичным для анализа текстов полагать $\varepsilon = 0,001$ и соответственно $M = 1,97$. Действительно, тексты не могут обладать такой же однородностью, как последовательности независимых одинаково распределенных случайных величин, и использование стандартных уровней значимости 0,01–0,1 здесь никак не оправдано.

При рассмотрении 80 текстов, составленных из произведений разных авторов, оказалось, что значения $|M_n|$ колеблются в широких

Рис. 4.16. График эмпирического моста для индикаторов служебных слов во фрагменте 2.2 собрания сочинений М. Шолохова

пределах от 0,90 до 10,84, но только для 11 текстов оказываются меньше, чем 1,97, то есть частота ошибки второго рода составляет 0,138.

При рассмотрении 10 текстов, составленных из двух произведений одного автора, оказывается, что значения $|M_n|$ меняются в пределах от 0,88 до 3,70, при этом для 5 текстов значения больше 1,97. Отметим, что это тексты действительно существенно разнородные (для Стругацких значения 3,70 и 3,63 в прямом и обратном порядке следования текстов соответственно, для А. Толстого 2,47 и 2,84, для Пелевина 1,27 и 2,19). Итак, при использовании критического

уровня 0,001 удается обнаружить разнородные тексты, но при этом к разнородным критерий относит и тексты, написанные одним автором в разное время или в разных стилях.

Рассмотрим теперь, как изменятся результаты при использовании словаря, составленного только из союзов. Для текстов одного автора значения $|M_n|$ меняются от 0,91 («Урфин Джюс и его деревянные солдаты») до 2,37 («Аэлита»). После исключения рекордного значения 2,37 получаем пределы от 0,91 до 1,98 с достигаемыми уровнями значимости от 0,381 до 0,0007 соответственно. Выберем $\varepsilon = 0,0005$ и соответственно $M = 2,04$. В этом случае частота ошибки первого рода равна 0,1. При анализе текстов двух авторов оказывается, что $|M_n|$ меняется в пределах от 0,98 до 12,25, при этом для 11 авторов значения оказываются меньше, чем 2,04, то есть частота ошибки второго рода составляет 0,138, как и при использовании словаря авторского инварианта. При анализе текстов, составленных из двух произведений одного автора, оказывается, что значения супремума модуля эмпирического моста колеблются в пределах от 0,99 до 4,05, при этом только 3 значения превосходят уровень 2,04, то есть частота ошибки составляет 0,3.

Резюмируя результаты анализа, делаем вывод о том, что при использовании всего словаря авторского инварианта можно принять уровень значимости критерия равным 0,001, при использовании словаря союзов — равным 0,0005. При этом в обоих случаях один из 10 однородных текстов принимается за неоднородный, а 11 из 80 неоднородных текстов — за однородные, то есть при решении задачи различения одного текста одного автора и двух текстов двух разных авторов частоты ошибок составляют 0,1 и 0,138 соответственно.

При решении задачи различения двух произведений одного автора от двух произведений разных авторов в 5 случаях из 10 од-

нородные тексты принимаются за неоднородные при использовании словаря авторского инварианта, в 3 случаях из 10 — при использовании словаря союзов. В этом отношении использование словаря союзов представляется предпочтительным.

4.4 Проверка гипотез о фрактальности для текстов

Анализ, проведенный в предыдущем параграфе, показывает, что процессы, построенные по произведениям одного автора (табл. 4.1), собраниям сочинений или их фрагментам (табл. 4.5), не вполне соответствуют модели выборки — достигаемые уровни значимости часто близки к нулю. В качестве объяснения этого явления предлагается модель фрактального шума, которая исследуется на адекватность методами второй главы.

В таблице приведены

n — число слов в тексте;

τ — шаг масштабирования;

$\tilde{H}_1$, $\tilde{H}_2$, $\tilde{H}_3$ — оценки параметра методами нормированного размаха, дисперсии, знаков.

ε^*_{23}, ε^*_{12}, ε^*_{13} — достигаемые уровни значимости гипотезы о фрактальном гауссовском шуме для критериев, основанных на разностях соответствующих оценок;

$N = n/\tau$ — объем масштабированной выборки;

ε^*_0 — достигаемый уровень значимости гипотезы о том, что адекватной моделью является модель выборки.

Табл. 4.6. Оценки параметра Херста в собраниях сочинений и их фрагментах.

автор	n	τ	$\tilde{H}_1$	$\tilde{H}_2$	$\tilde{H}_3$	ε^*_{23}	ε^*_{12}	ε^*_{13}	N	ε^*_0
dostoevski	1185208	64	0,684	0,674	0,614	0,2008	0,0260	0,9175	18519	0,00E+00
dostoevski1	347114	64	0,697	0,642	0,612	0,3831	0,5825	0,4423	5424	2,98E-13
dostoevski2	838095	64	0,690	0,639	0,621	0,8377	0,6320	0,9044	13095	0,00E+00
sholohov	637697	64	0,840	0,832	0,681	0,0003	0,0235	0,0013	9964	0,00E+00
sholohov1	407237	64	0,792	0,759	0,640	0,0039	0,1888	0,0027	6363	0,00E+00
sholohov11	187778	64	0,654	0,558	0,613	0,7673	0,8615	0,7785	2934	7,33E-08
sholohov2	230461	64	0,813	0,763	0,626	0,0010	0,4868	0,0000	3601	2,29E-11
sholohov22	121155	64	0,649	0,610	0,598	0,4197	0,1838	0,9712	1893	1,67E-04
tolstoj	1058098	64	0,715	0,652	0,619	0,5512	0,9592	0,3834	16533	0,00E+00
tolstoj1	127802	64	0,740	0,673	0,624	0,1532	0,5509	0,1120	1997	1,07E-06
tolstoj11	41632	64	0,631	0,462	0,509	0,8821	0,3569	0,2723	651	8,45E-01
tolstoj12	86171	64	0,657	0,506	0,604	0,5003	0,2854	0,9914	1346	7,79E-04
tolstoj2	930297	64	0,685	0,662	0,614	0,3344	0,1007	0,9633	14536	0,00E+00
tolstoj21	252586	64	0,701	0,649	0,628	0,4943	0,5041	0,6964	3947	1,52E-12
tolstoj211	59323	64	0,694	0,639	0,650	0,6078	0,3099	0,8531	927	5,80E-05
tolstoj2111	38866	64	0,695	0,640	0,620	0,3702	0,2647	0,7662	607	9,25E-03
tolstoj2112	20458	64	0,770	0,591	0,735	0,3466	0,2700	0,7062	320	2,10E-04

Согласно табл. 4.6, уровни значимости гипотезы о фрактальном гауссовском шуме достаточно высоки, за исключением случаев, когда текст существенно неоднороден (собрание сочинений Шолохова и его фрагменты 1 и 2). При этом лучше всего диагностируют разладку разности оценок $\tilde{H}_2 - \tilde{H}_3$ или $\tilde{H}_1 - \tilde{H}_3$, одна из которых отвечает за глобальные, а другая за локальные характеристики. В качестве критического значения ε можно взять любое число от 0,01 до 0,1, при этом не будет ни одной ошибки обнаружения разладки.

Гипотеза об адекватности модели выборки не выдерживает проверки, достигаемые уровни значимости оказываются очень низкими, за единственным исключением («Детство» Л. Толстого, данные, содержащие после масштабирования всего 651 значение).

Итак, адекватной моделью процесса, описывающего появление

служебных слов в тексте, для одного автора является фрактальный шум, для склейки текстов двух авторов — разладка фрактального шума. В качестве алгоритма обнаружения разладки предлагаются алгоритмы, основанные на разностях оценок $\tilde{H}_2 - \tilde{H}_3$ или $\tilde{H}_1 - \tilde{H}_3$.

4.5 Результаты главы 4

В главе 4 методами оценивания параметров и проверки статистических гипотез изучались тексты на естественном языке.

- Предложены однопараметрические вероятностные модели текста. Изучены статистические свойства оценок параметров, отыскиваемых на основании статистики числа разных слов. Показана неадекватность модели выборки (систематическое изменение оценок параметров с ростом объема текста).
- Доказана функциональная центральная предельная теорема для числа разных слов.
- Разработан метод авторского инварианта — способ сопоставления тексту числовой последовательности, сохраняющей свои статистические свойства для фиксированного автора. Алгоритмы обнаружения разладки применяются для анализа однородности текстов, а также комбинаций текстов разных авторов. Выработаны рекомендации к использованию этих методов для исследования авторства. Проанализированы собрания сочинений и показано, что модель выборки для текстов одного автора (и модель разладки для текстов двух авторов) становятся неадекватными при большом объеме текста.
- Алгоритмы обнаружения фрактальности и алгоритмы обнаружения разладки фрактального гауссовского шума применены

к анализу текстов одного и двух авторов. Показано, что адекватной моделью процесса, описывающего появление служебных слов в тексте, для одного автора является фрактальный шум, для склейки текстов двух авторов — разладка фрактального шума.

ГЛАВА 5
Разладка в регрессионных моделях с циклическим трендом

5.1 Вводные замечания

Асимптотическое поведение сумм остатков регрессионной модели временного ряда впервые изучено в работе МакНила [139]. Предполагается, что $\{\varepsilon_i,\ i \geq 1\}$ — последовательность независимых одинаково распределенных случайных величин с нулевым математическим ожиданием и конечной ненулевой дисперсией σ^2. Пусть $\{g_j(\cdot),\ 1 \leq j \leq m\}$ — набор регрессионных функций, определенных на $[0,\ 1]$. Зададим треугольный массив $\{Y_{ni},\ 1 \leq i \leq n,\ n \geq 1\}$ зависимых переменных следующим образом:

$$Y_{ni} = \sum_{j=1}^{m} \theta_j g_j(i/n) + \varepsilon_i.$$

Итак, в этой модели все время наблюдений сжато на отрезок от нуля до единицы, и наблюдения отстоят друг от друга на равные интервалы времени. В матричной записи

$$\mathbf{Y}_n = \mathbf{X}_n\theta + \varepsilon_n,$$

где $\mathbf{X}_n$ — матрица регрессора размерности $n \times m$, в которой на j-м месте в i-й строке стоит элемент $g_j(i/n)$; θ — вектор-столбец длины m; ε_n и $\mathbf{Y}_n$ — вектор-столбцы длины n.

Обозначим через $\widehat{\theta}_n$ оценку Гаусса—Маркова векторного параметра θ. Она равна

$$\widehat{\theta}_n = (\mathbf{X}_n^T\mathbf{X}_n)^{-1}\mathbf{X}_n^T\mathbf{Y}_n,$$

если обратная матрица существует.

Если регрессионные функции интегрируемы с квадратом по Ри-

ману на [0, 1], то существует предел

$$\lim_{n\to\infty} n^{-1}\mathbf{X}_n^T\mathbf{X}_n = G$$

— матрица, компоненты которой равны $\int_0^1 g_i(t)g_j(t)dt$.

Обозначим через $\mathbf{g}(\cdot)$ вектор-столбец регрессионных функций.

Частичные суммы регрессионных остатков обозначим через $\widehat{\Delta}_{nk} = \sum_{i=1}^k \hat{\varepsilon}_{ni}$, где $\hat{\varepsilon}_{ni} = Y_{ni} - \widehat{Y}_{ni}$, $\widehat{\mathbf{Y}}_n = \mathbf{X}_n\hat{\theta}_n$.

Будем полагать $\widehat{\Delta}_{n0} = 0$.

Рассмотрим эмпирический мост регрессионных остатков Z_n. Согласно формуле (1), Z_n — это случайная ломаная с узлами в точках

$$\left(\frac{k}{n},\ \frac{\widehat{\Delta}_{nk} - \frac{k}{n}\widehat{\Delta}_{nn}}{\tilde{\sigma}_n\sqrt{n}}\right),$$

где

$$\tilde{\sigma}_n = \sqrt{\overline{\hat{\varepsilon}^2} - \left(\overline{\hat{\varepsilon}}\right)^2}.$$

Наряду с эмпирическим мостом рассмотрим случайную ломаную Z_n^σ, построенную по точкам

$$\left(\frac{k}{n},\ \frac{\widehat{\Delta}_{nk}}{\sigma\sqrt{n}}\right).$$

Теорема МакНила состоит в том, что *если регрессионные функции непрерывно дифференцируемы, и матрица G невырождена, то Z_n^σ слабо сходится в равномерной метрике в $C(0,1)$ к гауссовскому процессу $B = \{B(t),\ 0 \le t \le 1\}$ с нулевым математическим ожиданием и ковариационной функцией*

$$K(s,t) = \mathbf{E}B(s)B(t) = \min(s,t) - \int_0^1\int_0^1 g(x,y)\,dx\,dy,$$

где $g(x,y) = \mathbf{g}^T(x)G^{-1}\mathbf{g}(y)$.

Бишоф [112] ослабил предположения МакНила о регрессионных функциях, доказав, что теорема верна для непрерывных на [0, 1] функций (условие непрерывной дифференцируемости не требуется).

В частном случае, когда регрессионные функции — это 1, $\cos 2\pi kt$, $\sin 2\pi kt$, $k \in M$, где M — некоторое конечное множество, модель принимает вид

$$Y_{ni} = a_0 + \sum_{k \in M} (a_k \cos(2\pi ki/n) + b_k \sin(2\pi ki/n)) + \varepsilon_i. \qquad (38)$$

Известно, что любое периодическое колебание с траекториями из $D(0,1)$ может быть единственным образом представлено разложением в ряд Фурье по синусоидам. Рассматриваемая здесь регрессия — это модель с конечным числом взаимно ортогональных гармоник и аддитивным случайным шумом в дискретном времени. Эта модель известна в литературе [2], [136], [30] как модель линейной регрессии с циклическим трендом. Изучение статистических критериев обнаружения разладки в этой модели предпринято в связи с исследованием колебаний строительных конструкций [96]: ставится задача определения изменений прочностных характеристик конструкции по исследованию ее колебаний.

Предполагается, что моменты наблюдения равноотстоят друг от друга, и что период наблюдений состоит из целого числа периодов колебаний.

Макнил [139] в качестве следствия своей теоремы получил, что для регрессионной модели с циклическим трендом ковариационная функция предельного процесса равна

$$K(s,t) = \min(s,t) - st \qquad (39)$$
$$-\frac{1}{2\pi^2} \sum_{k \in M} \frac{1}{k^2} ((1 - \cos 2\pi ks)(1 - \cos 2\pi kt) + \sin 2\pi ks \sin 2\pi kt).$$

В §5.2 получены результаты для регрессии на порядковые статистики. В §5.3 доказана теорема о сходимости эмпирического моста остатков регрессии с циклическим трендом. Получены следствия из нее, пригодные для практического применения: найдены

предельные распределения статистик, основанных на значениях эмпирического моста в нескольких точках, и грубая (логарифмическая) асимптотика максимума модуля эмпирического моста. В §5.4 рассматриваются статистические критерии, основанные на изученных статистиках, и проводится их сравнение на численном примере. Краткое изложение полученных результатов содержится в §5.5.

5.2 Регрессия на порядковые статистики

Результаты этого параграфа являются совместными с Е. В. Шаталиным и опубликованы в работах [169], [170], [198].

Рассмотрим две регрессионные модели (одно и двухпараметрическую):

$$Y_i = \theta + \varepsilon_i, i = 1, \ldots, n,\ n \geq 1,$$

$$Y_i = a + bX_i + \varepsilon_i, i = 1, \ldots, n,\ n \geq 1,$$

где $\theta, a, b \in \mathbf{R}$ — неизвестные параметры регрессии, $\varepsilon_1, \ldots, \varepsilon_n$ (регрессионные ошибки) — независимые одинаково распределенные случайные величины с нулевым математическим ожиданием и конечной ненулевой дисперсией σ^2. Также предполагается, что $X_i = \xi_{i:n},\ i = 1, \ldots, n,$ — порядковые статистики, где случайные величины $\xi_1, \ldots, \xi_n$ независимы, одинаково распределены с функцией распределения F и не зависят от случайных величин $\varepsilon_1, \ldots, \varepsilon_n$.

Неизвестные параметры регрессии обычно оценивают по методу наименьших квадратов, получая оценки $\widehat{\theta}, \widehat{a}, \widehat{b}$. На основании регрессионной модели строятся прогнозные значения $\widehat{Y_i} = \widehat{\theta},\ \widehat{Y_i} = \widehat{a} + \widehat{b}X_i$. Остатками линейной регрессии называют случайные величины $\widehat{\varepsilon_i} = Y_i - \widehat{Y_i}$.

Приведем определение эмпирического моста. Эмпирический мост — это кусочно-линейная случайная ломаная $Z_n^0 = \{Z_n^0(t),\ 0 \leq t \leq$

1\} с узлами в точках

$$\left(\frac{k}{n},\ \frac{\widehat{\Delta}_k - \frac{k}{n}\widehat{\Delta}_n}{\sqrt{\widehat{\sigma^2}n}}\right),$$

где $\widehat{\Delta}_k = \hat{\varepsilon}_1 + \ldots + \hat{\varepsilon}_k,\ k = 1, \ldots, n,\ \widehat{\Delta}_0 = 0,\ \widehat{\sigma^2} = \overline{\hat{\varepsilon}^2} - (\overline{\hat{\varepsilon}})^2$.

Пусть $GL_F(t) = \int\limits_0^t F^{-1}(s)\,ds$ – *теоретическая обобщенная кривая Лоренца*, $F^{-1}(s) = \sup\{x : F(x) < s\}$ — квантильное преобразование (обобщенная обратная функция) функции распределения $F(x)$, $GL_F^0(t) = GL_F(t) - tGL_F(1)$.

Через $\Longrightarrow$ будем обозначать слабую сходимость (сходимость по распределению) в соответствующем пространстве. Так, в теоремах ниже через $\Longrightarrow$ обозначена слабая сходимость в пространстве $C(0,1)$, снабженном равномерной метрикой.

Из функциональной предельной теоремы (принципа инвариантности) следует теорема 1.

В [170] доказана следующая теорема.

Теорема 5.1 *Если* $Y_i = a + bX_i + \varepsilon_i,\ 0 < \mathbf{Var}\xi_1 < \infty$, *то* $Z_n^0 \Longrightarrow Z_F$, *где* Z_F *— центрированный гауссовский процесс с ковариационной функцией*

$$K_F(t,s) = \min\{t,s\} - ts - \frac{GL_F^0(t)GL_F^0(s)}{\mathbf{Var}\xi_1},\ t,s \in [0,1].$$

Следствие 5.1 *Если* $Y_i = a + bX_i + \varepsilon_i$, X_i *— порядковые статистики, построенные по выборке из нормального распределения, то*

$$\int\limits_0^1 (Z_n^0(t))^2\,dt \Longrightarrow \hat{\eta} = \int\limits_0^1 Z_\Phi^2(t)\,dt,$$

где $\hat{\eta}$ *имеет распределение* $\widehat{\omega}^2$, *которое представлено в табл. 3 на с. 65 в [65],* Z_Φ *— центрированный гауссовский процесс с ковариационной функцией*

$$\widehat{K}(t,s) = \min\{t,s\} - ts - \varphi(\Phi^{-1}(t))\varphi(\Phi^{-1}(s)),$$

где φ, Φ^{-1} — *плотность и квантильная функция стандартного нормального распределения соответственно.*

В [65] получены следующие выражения:

$$F_{\widehat{\eta}}(x) = 1 + \frac{1}{\pi} \sum_{k=1}^{\infty} \int\limits_{\lambda_{2k-1}}^{\lambda_{2k}} \frac{e^{-x\lambda/2}}{(-D(\lambda))^{1/2}} \frac{d\lambda}{\lambda}$$

(с. 56), где λ_k — собственные числа ядра $\widehat{K}$, $\lambda_1 \dots \lambda_{10}$ приведены на с. 37 в [65], для остальных λ_k справедлива эквивалентность $\lambda_k \sim ((2k-1)\pi)^2$, $k > 10$;

$$D(\lambda) = \frac{2}{\sqrt{\lambda}} A(\sqrt{\lambda}) \sin \frac{\sqrt{\lambda}}{2};$$

$$A(\mu) = \Big(1 + \frac{\mu^2\sqrt{3}}{2\pi}\Big) \cos \frac{\mu}{2} + 4\mu^3 I_1(\mu);$$

$$I_1(\mu) = \int\limits_0^{1/2} \kappa_1(x) \sin \mu x \, dx \int\limits_x^{1/2} \kappa_1(x) \cos \mu(1/2 - t) \, dt;$$

$$\kappa_1(x) = \varphi(\Phi^{-1}(x)).$$

(с. 34–37 в [65]). С помощью данных формул можно вычислить $F_{\widehat{\eta}}(x)$, результат приведен в табл. 3 в [65].

5.3 Регрессия с циклическим трендом

Теорема 5.2 *Эмпирический мост* Z_n *остатков регрессии для модели (38) слабо сходится в равномерной метрике в* $C(0,1)$ *к гауссовскому процессу* $B = \{B(t),\ 0 \le t \le 1\}$ *с нулевым математическим ожиданием и ковариационной функцией (39), которую можно записать также в виде*

$$K(s,t) = \min(s,t) - st - \frac{2}{\pi^2} \sum_{k \in M} \frac{1}{k^2} \sin \pi k s \sin \pi k t \cos \pi k (s-t). \quad (40)$$

Доказательство.

Эквивалентность формул (39) и (40) устанавливается непосредственно с помощью формул тригонометрии.

Отметим, что $\widehat{\Delta}_{nn} = 0$ в случае, когда процесс задан формулой (38).

Вместе с тем $\tilde{\sigma}_n$ является состоятельной оценкой параметра σ в силу состоятельности оценок коэффициентов регрессии, ограниченности косинуса и синуса, сходимостей $\sum_{i=1}^{n} \varepsilon_i/n \to 0$, $\sum_{i=1}^{n} \varepsilon_i^2/n \to \sigma^2$ с вероятностью единица. Поэтому предельная теорема для Z_n^σ верна также для эмпирического моста.

Доказательство завершено.

Модель (38) будем считать основной гипотезой, а в качестве альтернативной предполагается, что в некоторый момент времени значение a_0 в модели (38) заменяется на некоторое отличное от него значение b_0. При альтернативной гипотезе предполагается, что это изменение происходит только один раз за весь период наблюдений.

Задача состоит в том, чтобы построить класс статистических критериев, различающих основную и альтернативную гипотезы. После построения статистических критериев необходимо их сравнить и выбрать более мощный. Естественным подходом к обнаружению разладки, т. е. изменения параметров модели в процессе наблюдения, является анализ регрессионных остатков.

Простейшим критерием обнаружения разладки является критерий, основанный на сравнении средних значений остатков, подсчитанных по первой и второй половинам наблюдений. Ясно, что разность средних должна нормироваться среднеквадратическим отклонением или его выборочным аналогом. В результате получаем статистический критерий, близкий к критерию Стьюдента: большие по модулю значения нормированной разности средних свидетельству-

ют против основной гипотезы. Этот критерий, основанный на разности средних по первой и второй половине наблюдений, в терминах эмпирического моста основан на статистике $Z_n(1/2)$. Действительно, если число наблюдений n четно, то, так как сумма остатков регрессии в модели (38) равна нулю, разность между суммой первых $n/2$ и последних $n/2$ остатков равна $2\widehat{\Delta}_{n,n/2}$. При делении на $\tilde{\sigma}_n\sqrt{n}$ получаем $2Z_n(1/2)$. При нечетном n логично использовать ту же статистику как обеспечивающую симметрию между первой и второй половинами наблюдений.

Обозначим

$$J_1 = Z_n^2(1/2)/\mathbf{D}B(1/2) = Z_n^2(1/2)/\left(\frac{1}{4} - \frac{2}{\pi^2}\sum_{k\in M}\frac{1}{k^2}\sin^2\frac{\pi k}{2}\right).$$

При верной основной гипотезе статистика J_1 сходится слабо к распределению χ_1^2. Отметим, что для конечного множества M всегда $\mathbf{D}B(1/2) > 0$, а если бы множество M включало в себя все положительные нечетные числа, то предельный процесс B в точке $1/2$ имел бы вырожденное в нуле распределение.

Предложенный критерий можно обобщить следующим образом. Вместо одной точки $1/2$ взять d точек: $\frac{1}{d+1}, \ldots, \frac{d}{d+1}$, и рассмотреть декоррелированные и нормализованные (в соответствии с корреляционной функцией (40)) значения эмпирического моста Z_n в этих точках. Тогда сумма квадратов этих значений будет иметь в пределе хи-квадрат распределение с d степенями свободы в силу того, что сумма квадратов линейных функций от значений эмпирического процесса в фиксированных точках является непрерывным (в равномерной метрике) функционалом от эмпирического моста. И полученный таким образом критерий, и статистику, на которой он основан, будем обозначать J_d. Обозначим

$$\mathbf{z}_d = (z_{1,d}, \ldots, z_{d,d})^T = \left(Z_n\left(\frac{1}{d+1}\right), \ldots, Z_n\left(\frac{d}{d+1}\right)\right)^T.$$

Через C_d обозначим ковариационную матрицу вектора $\mathbf{b}_d = \left(B\left(\frac{1}{d+1}\right), \ldots, B\left(\frac{d}{d+1}\right)\right)^T$:

$$C_d = \mathbf{E}\mathbf{b}_d^T\mathbf{b}_d.$$

Статистика J_d вычисляется по формуле

$$J_d = \mathbf{z}_d^T C_d^{-1}\mathbf{z}_d.$$

Опишем подробнее критерии J_2 и J_3. Для построения критерия J_2 используем значения $Z_n(1/3)$ и $Z_n(2/3)$. Согласно (40) предельное значение $B(1/3)$ имеет дисперсию

$$c_{11} = K(1/3, 1/3) = \frac{2}{9} - \frac{2}{\pi^2}\sum_{k\in M}\frac{1}{k^2}\sin^2\frac{\pi k}{3}.$$

В силу формул приведения, дисперсия предельного значения $B(1/3)$ такая же: $c_{22} = c_{11}$. Ковариация случайных величин $B(1/3)$ и $B(2/3)$ равна $c_{12} = c_{11}/2$.

Следовательно, статистика

$$J_2 = \frac{4}{3b_{11}}\left(Z_n^2(1/3) - Z_n(1/3)Z_n(2/3) + Z_n^2(2/3)\right)$$

сходится слабо к распределению χ_2^2. Для построения критерия J_3 выполняем те же действия со значениями $Z_n(1/4)$, $Z_n(1/2)$ и $Z_n(3/4)$. Особенно простой вид имеет статистика J_3 в случае, когда M состоит из k, кратных 4. В этом случае матрица ковариаций равна

$$C_3 = \frac{1}{16}\begin{pmatrix} 3 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 3 \end{pmatrix},$$

и статистика

$$J_3 = 8(Z_n^2(1/4) - Z_n(1/4)Z_n(1/2) + Z_n^2(1/2)$$

$$-Z_n(1/2)Z_n(3/4) + Z_n^2(3/4))$$

сходится слабо к распределению χ^2_3.

Наряду с критериями J_1, J_2, J_3 будем использовать критерий, основанный на статистике $J = \sup_{t\in[0,1]} |Z_n(t)|$. Как показано в главе 1, этот критерий является асимптотически наиболее мощным в широком классе критериев в предположении, что $M = \emptyset$, т. е. в модели случайной выборки. В общем случае (при $M \neq \emptyset$) для предельного закона статистики нет аналитического описания, из общей теории [131] известна лишь грубая асимптотика больших уклонений

$$\ln \mathbf{P}\{\sup_{t\in[0,1]} |B(t)| \geq x\} \sim -\frac{x^2}{2\sup_{t\in[0,1]} K(t,t)}$$

при $x \to \infty$. Если M состоит только из четных чисел, то $\ln \mathbf{P}\{\sup_{t\in[0,1]} |B(t)| \geq x\} \sim -2x^2$.

5.4 Сравнение критериев

Смоделируем процесс (38) и сравним критерии, основанные на статистиках J_1, J_2, J_3 и J. Положим $n = 1024$, $M = \{4; 16\}$, $b_1 = b_2 = \sigma = 1$, $a_0 = a_1 = a_2 = 0$. В случае разладки будем полагать $b_0 = 0,2$, а момент разладки τ равномерно распределенным на целых числах от 1 до n. Графики траектории процесса изображены на рис. 5.1. Отметим, что периоды синусоид можно определить на основе быстрого преобразования Фурье (рис. 5.2). Для процесса с разладкой он имеет очень похожий вид. После оценивания параметров и вычитания прогнозных значений получаем остатки регрессии (рис. 6.3). Непосредственно на основании изучения рисунка затруднительно сделать вывод о том, произошла ли разладка в каждом конкретном случае. Поэтому вычислим значения статистик J_1, J_2, J_3 и J и достигнутые ими уровни значимости. Результаты приведены в табл. 5.1.

Таблица 5.1. Значения статистик критериев и достигнутые уровни

значимости для примера моделирования процесса.

Критерий	Без разладки		С разладкой	
	Значение статистики	Достигнутый уровень значимости	Значение статистики	Достигнутый уровень значимости
J_1	0.65	0.42	2.42	0.12
J_2	1.75	0.42	6.34	0.04
J_3	2.91	0.41	5.91	0.12
J	0.82	0.51	1.32	0.06

В приведенном примере каждый из рассмотренных критериев не дает оснований предполагать разладку в случае, когда разладки на самом деле нет. В случае, когда разладка есть, критерии J_1 и J_3 отвергают основную гипотезу на уровне 0,12, критерий J на уровне 0,06, а критерий J_2 на уровне 0,04. Отметим, что уровень для критерия J требует коррекции, так как точное предельное распределение его статистики неизвестно.

Для того, чтобы скорректировать критерий J, а также проверить соответствие критических уровней для остальных критериев, анализируем поведение критериев при отсутствии разладки. Эмпирические уровни значимости вычисляем по результатам 20000 моделирований процесса. Для критериев J_1, J_2, J_3 эмпирические уровни значимости отличаются от выбранного теоретического уровня 0,05 не более чем на 0,002, а для критерия J эмпирический уровнень значимости принимает значительно более низкое значение 0,0389. Это происходит вследствие того, что для этого критерия неизвестно точное предельное распределение, а лишь его грубая (логарифмическая) асимптотика. Поэтому скорректируем критерий J, выбрав в качестве нового уровня $0,05^2/0,0389 \approx 0,0643$. Повторный численный эксперимент по 20000 результатам моделирования дает эмпирический уровень значимости 0,0510. Всюду в дальнейшем используется этот исправленный критерий J для уровня значимости 0,05. В случае разладки будем полагать момент разладки τ равномерно распределенным на целых числах от 1 до n, а значению математического ожидания после разладки будем последовательно придавать значения от 0,2 до 1 с шагом 0,2. Проведем вычисления с исправленным критерием J. Результаты вычислений приведены в табл. 5.2. Погрешность каждого вычисления не превосходит 0,01 с надежностью не менее 0,95.

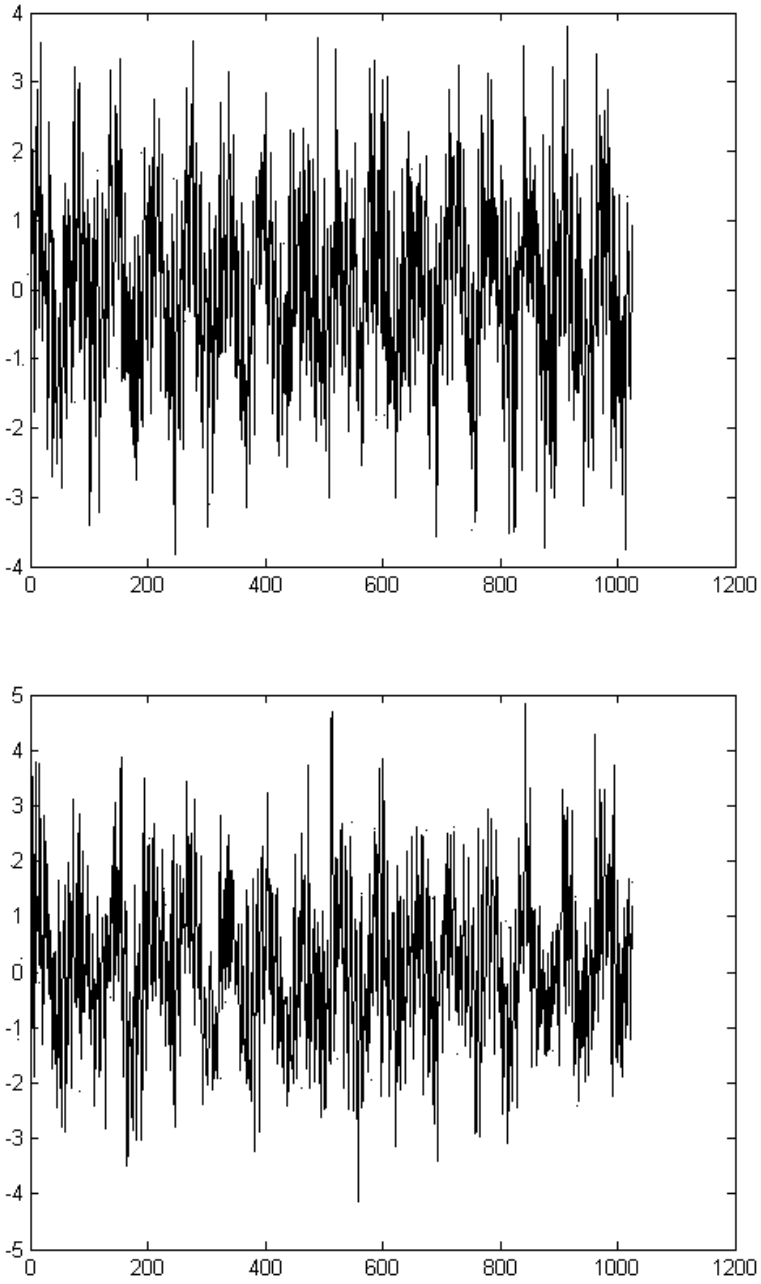

Рис. 5.1. График траектории процесса (вверху — без разладки; внизу — с разладкой)

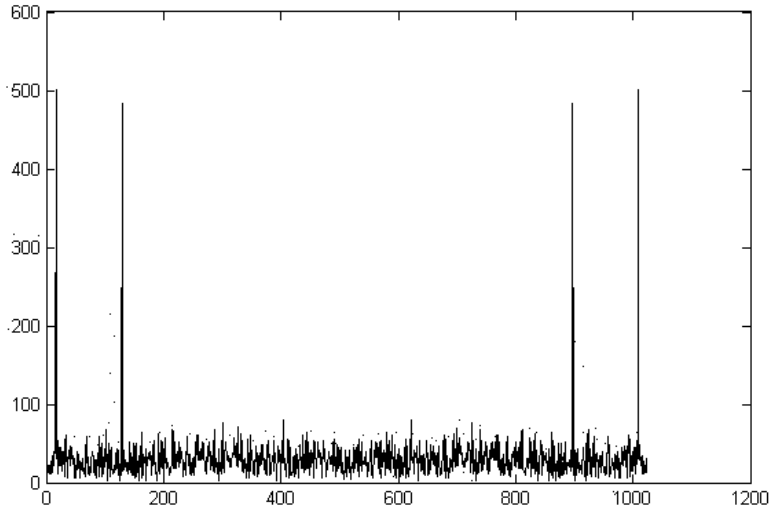

Рис. 5.2. График быстрого преобразования Фурье для процесса без разладки

Таблица 5.2. Эмпирические мощности критериев.

Используемый критерий	Разность математических ожиданий					
	0.0	0.2	0.4	0.6	0.8	1.0
J_1	0.05	0.40	0.69	0.80	0.85	0.87
J_2	0.05	0.43	0.74	0.82	0.86	0.88
J_3	0.05	0.43	0.79	0.86	0.90	0.91
J	0.05	0.50	0.80	0.88	0.91	0.93

Из таблицы видно, что исправленный критерий J, основанный на максимальном отклонении эмпирического моста, при рассматриваемой альтернативной гипотезе является более мощным, чем критерии J_1, J_2, J_3, основанные на декорреляции и нормировке значений эмпирического моста в фиксированных точках. В то же время с ростом d мощность критерия J_d все ближе к мощности критерия J.

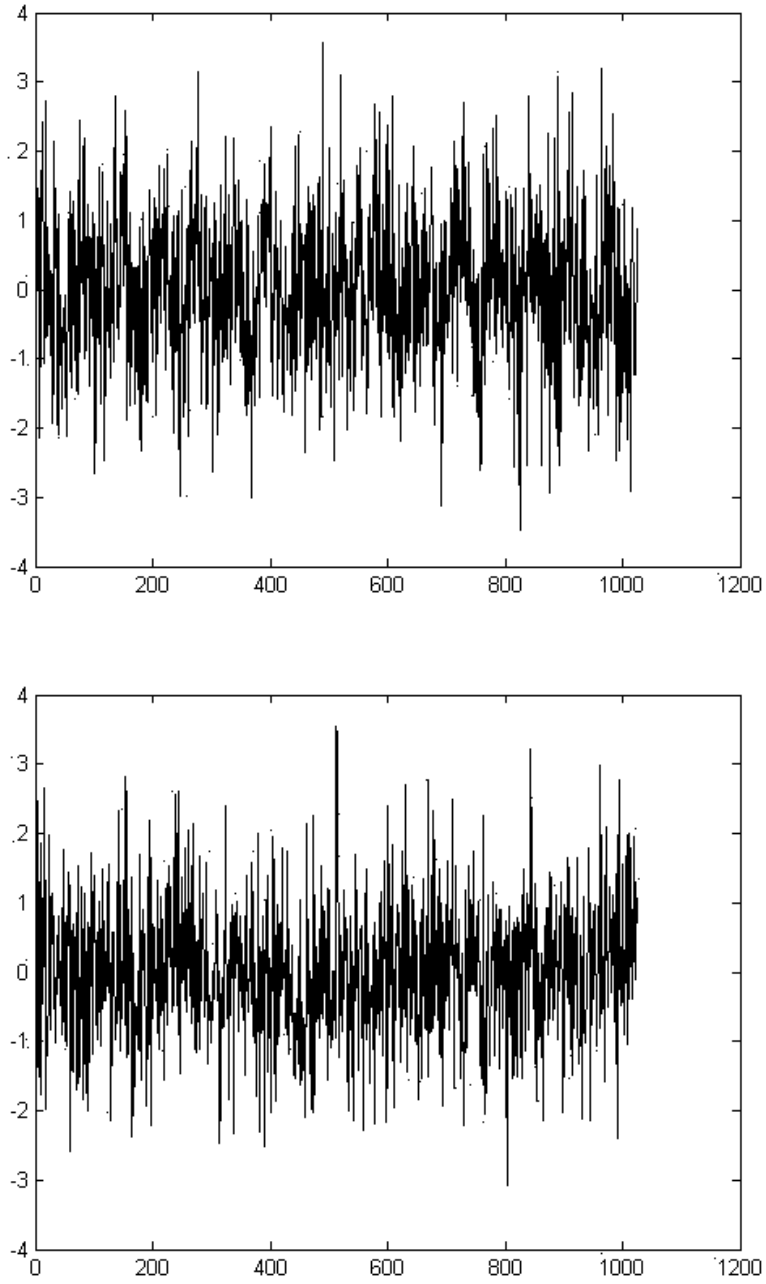

Рис. 5.3. Остатки регрессии (вверху — без разладки, внизу — с разладкой)

5.5 Результаты главы 5

В главе 5 построены статистические критерии обнаружения разладки регрессии на порядковые статистики и процесса гармонических колебаний со случайным шумом, пригодные для использования при компьютерном анализе изменений прочностных характеристик конструкции на основании записи ее колебаний. Класс таких критериев основан на декорреляции и нормировке значений эмпирического моста.

- Для регрессии на порядковые статистики получены результаты, описывающие асимптотику эмпирического моста, в том числе в случае, когда регрессор распределен по нормальному закону. Эти результаты позволяют строить алгоритмы проверки соответствия данных модели регрессии на порядковые статистики.

- Для обнаружения разладки модели с циклическим трендом вычисляется корреляционная матрица значений предельного гауссовского процесса в точках $\frac{1}{d+1}, \ldots, \frac{1}{d+1}$. Затем теоретические значения в этих точках ортогонализуются, нормируются, возводятся в квадрат и складываются. В результате получаем распределение хи-квадрат с d степенями свободы. Это распределение является предельным для статистики от эмпирического моста, построенной с использованием тех же преобразований.

- Более детально описаны критерии J_1, J_2, J_3. Для них статистики критериев приведены в явном виде (для критерия J_3 только в случае, когда число наблюдений кратно четырем).

- Наряду с ними рассматривается критерий, основанный на максимальном отклонении эмпирического моста от оси абсцисс. Этот критерий требует корректировки по результатам модели-

рования, так как для него известна только грубая асимптотика предельного распределения.

- Сравнение критериев проводится на численном примере и показывает преимущество критерия, основанного на максимальном отклонении эмпирического моста в случае, когда альтернатива состоит в однократном изменении математического ожидания в случайный момент времени, равномерно распределенный на интервале наблюдения. Однако критерии, основанные на измерениях в нескольких точках, приближаются к нему по мощности с ростом числа точек.

ГЛАВА 6
Анализ адекватности вероятностных моделей медицинских и экономических данных

5.6 Вводные замечания

В этой главе рассмотрены 4 массива данных, изученных при решении прикладных задач, и показаны решения, найденные автором с помощью разработанных в диссертации подходов. В параграфе 6.2 это данные медицинской статистики, в параграфах 6.3 и 6.4 экономические данные, а в параграфе 6.5 данные записей результатов моделирования колебаний строительной конструкции. Результаты параграфа 6.3 являются совместными с Н. С. Аркашовым и опубликованы в работе [168]. Результаты параграфа 6.5 являются совместными с А.М. Шахраманьяном и опубликованы в работе [201].

5.7 Модели зависимости концентрации от массы тела

Рассматриваемые данные — концентрации M_i (мг/л) маркеров FB MoM и PA MoM в крови пациента в зависимости от массы тела W_i (кг) для 4251 пациента.

Данные предоставлены фирмой «Typolog Software Ltd & Co.» (Германия), подрядчиком фирмы «SIEMENS Medical Solutions Diagnostics».

Задача состоит в выборе модели для корректировки значений MoM в целях устранения их зависимости от массы тела. Первая модель предложена сотрудниками фирмы и приводит к модели нелинейной регрессии. Вторая модель предложена диссертантом и очень проста в применении.

Значения MoM имеют большой размах и далеки от соответствия нормальному закону. Поэтому перейдем к их логарифмам. Значения

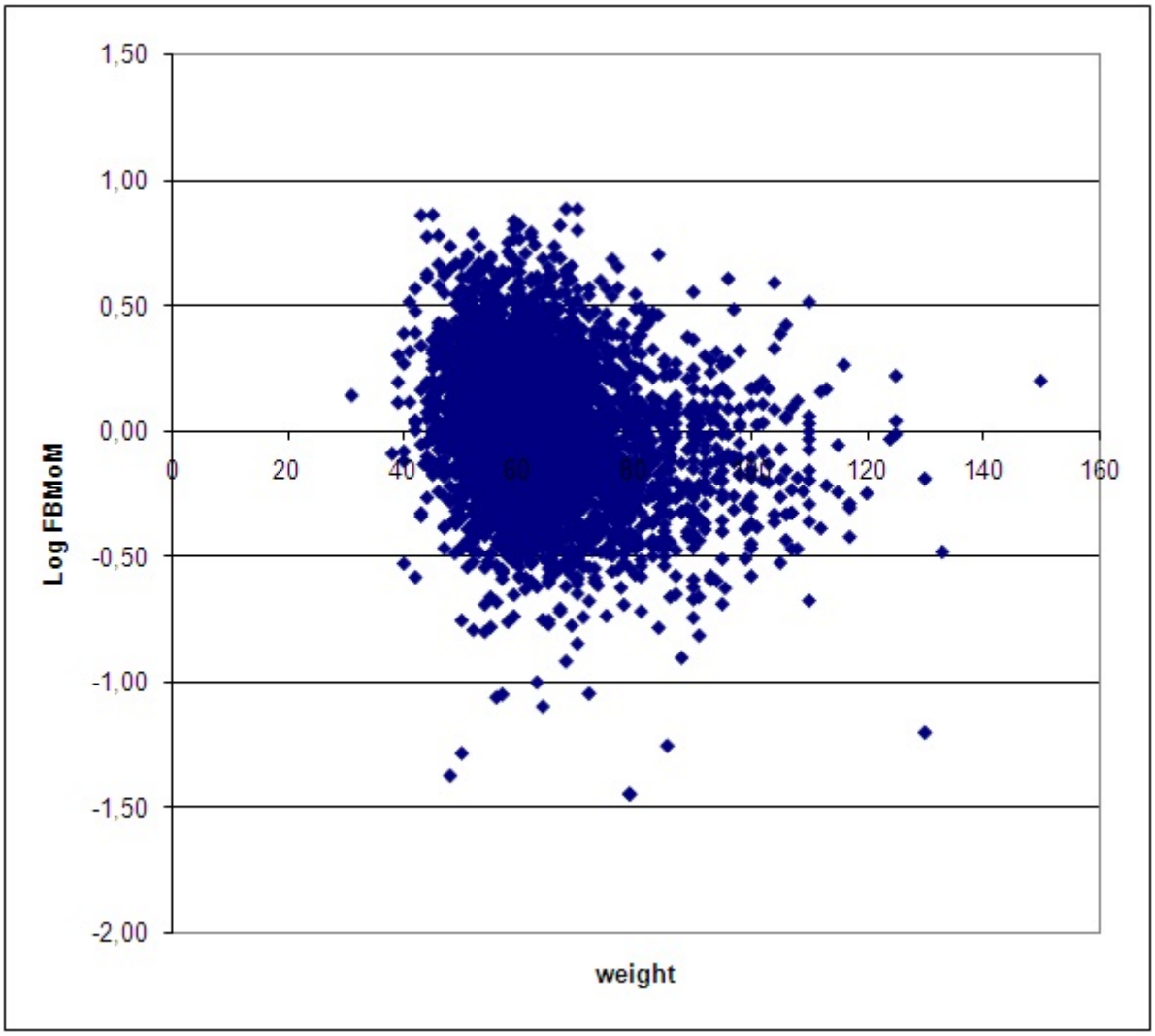

Рис. 6.1. Логарифмы FB MoM

Через $\widetilde{M_i}$ обозначим скорректированное значение MoM, через $\overline{M}$ и $\overline{W}$ соответствующие средние значения.

Предполагается, что $\lg M_i$ нормально распределены с математическим ожиданием c, и

$$\widetilde{M_i} = \frac{M_i}{a + (1-a)W^0/W_i}$$

одинаково распределены. Константа W^0 называется «нейтральной массой тела»: если $W_i = W^0$, то $\widetilde{M_i} = M_i$. В качестве W^0 выбирается геометрическое среднее, вычисления дают $W^0 \approx 63,27$ кг.

При известном a оценка константы c вычисляется элементарно методом наименьших квадратов, а оценивание константы a приводит

к задаче нелинейной оптимизации. Эта задача решается численно, $\hat{a} \approx 0,247$ для FB MoM, $\hat{a} \approx -0,334$ для PA MoM.

Для анализа адекватности модели построим эмпирический мост для скорректированных значений. На рис. 6.2 изображен эмпирический мост для FB MoM. Максимальное отклонение для него составляет 0,65, для PA MoM 0,79, что не дает оснований отвергать предложенную модель.

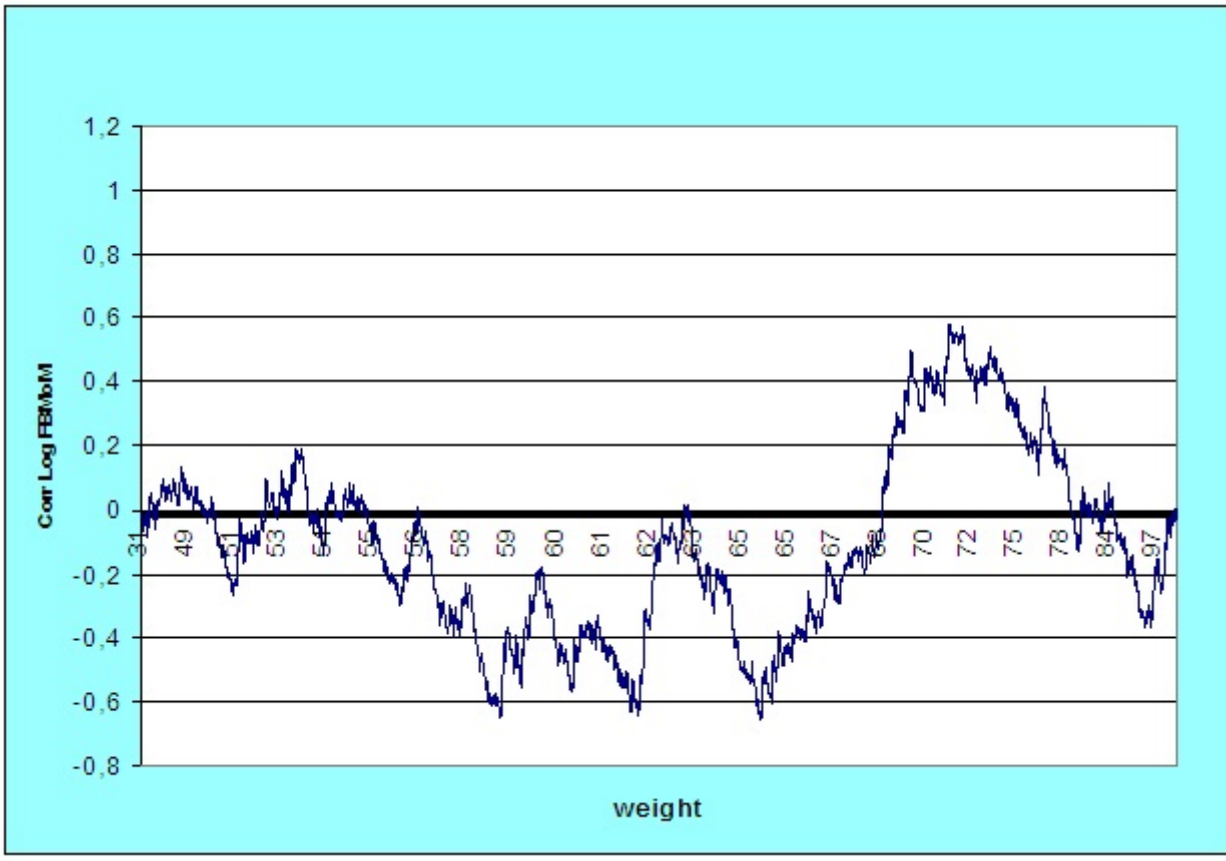

Рис. 6.2. Эмпирический мост для логарифмов FB MoM, исправленных в соответствии с первой моделью

Однако первая модель имеет два негативных качества: она приводит к сложным вычислениям, и результат вычисления оценки параметра a для PA MoM отрицателен ($\hat{a} \approx -0,334$) — нарушено

классическое (и имеющее естественную интерпретацию) условие положительности параметра в модели Михаэлиса—Ментен.

Поэтому перейдем ко второй модели:

$$\lg M_i = a + bW_i + \varepsilon_i.$$

Параметры оцениваются по методу наименьших квадратов: $\hat{a} \approx 0,308$, $\hat{b} \approx -0,0046$.

Логарифмы скорректированных значений вычисляются по формуле

$$\lg \widehat{M_i} = \lg M_i - \hat{b}(W_i - \overline{W}).$$

Здесь $\overline{W} \approx 64,29$ кг.

Для анализа адекватности модели построим эмпирический мост значений, скорректированных в соответствии с этой моделью. На рис. 6.3 изображен эмпирический мост для FB MoM (для PA MoM эмпирический мост имеет схожий вид). Максимальное его отклонение составляет 0,81 как для FB MoM, так и для PA MoM, что не дает оснований отвергать предложенную модель. Эта более простая модель была принята и использована при разработке программного обеспечения.

Подобный анализ для концентрации интерферона в крови был проведен в работе [180].

5.8 Анализ моделей цен на жилую недвижимость

Известно, что цена квадратного метра квартиры зависит от ряда характеристик жилья, из которых важнейшими являются удобство местоположения, число комнат и планировка квартиры, материал, из которого построен дом, внутренняя отделка, год сдачи дома в эксплуатацию и т.д. Часть этих характеристик может быть прямо или

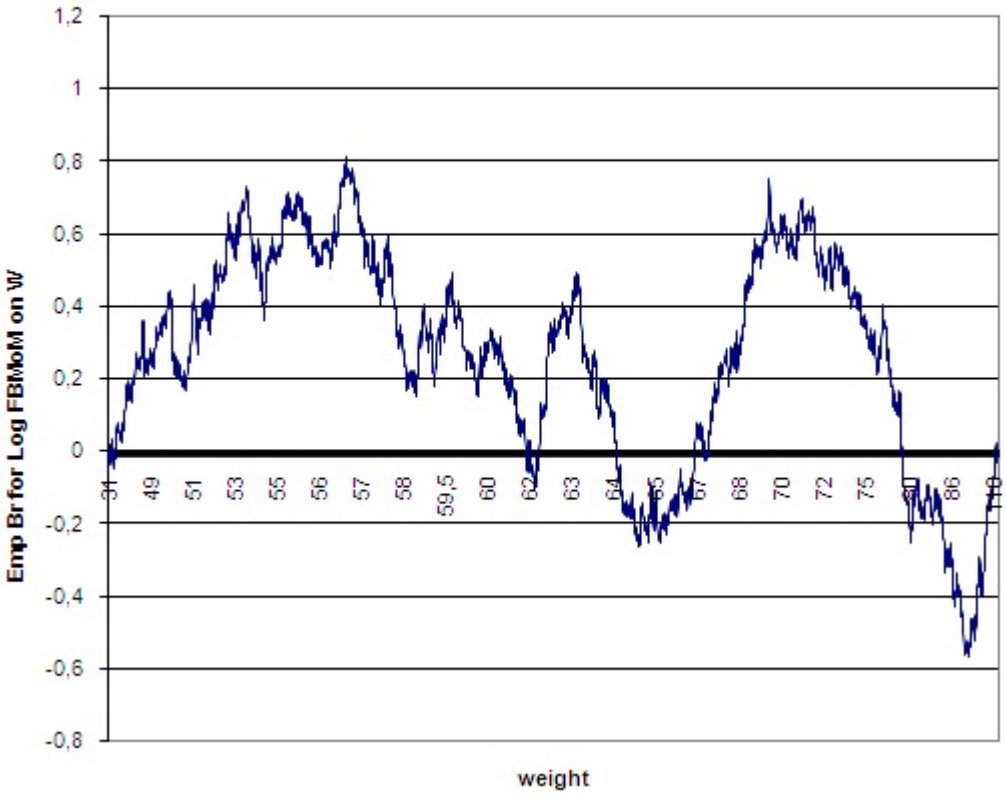

Рис. 6.3. Эмпирический мост для логарифмов FB MoM, исправленных в соответствии со второй моделью

косвенно получена из объявлений о продаже квартир. Кроме того, на цену предложения квартиры влияет уровень ожиданий продавца.

Мы построим ряд вероятностных моделей цены квадратного метра жилья, использующих известные характеристики квартир. Каждая модель по заданному набору характеристик выдает распределение вероятностей цены, позволяющее находить доверительные границы.

В качестве исходных данных, на которых проводится тестирование моделей, будем использовать базу данных новосибирского издания «Справочник по недвижимости» за апрель, май, июнь 2004 года.

В «Справочнике по недвижимости» приведены следующие характеристики квартир:

1) room — код числа комнат;
2) district — код района;
3) street — код улицы;
4) flat_floor — этаж квартиры;
5) house_floor — этажность дома;
6) material — код материала, из которого построен дом;
7) flat_plan — код планировки квартиры;
8) flat_type — код типа квартиры;
9) sanuzel — код типа санузла;
10) balkon — код наличия балкона/лоджии;
11) phone — код наличия телефона/линии;
12) sq_all — полная площадь (в кв. м.);
13) sq_useful — полезная площадь;
14) sq_kitchen — площадь кухни;
15) own — тип собственности;
16) price_all — цена.

Каждой из первых пятнадцати характеристик ставится в соответствие положительное число x_j, $j = 1, ..., 15$. Общую цену (16-ую характеристику) будем обозначать y. Числовые данные, соответствующие указанным характеристикам, упорядочены по возрастанию: во-первых, по коду числа комнат, затем по коду района и т.д., и в последнюю очередь по коду типа собственности.

Будем обозначать через y_t, $x_{t1}, ..., x_{t15}$ характеристики квартиры, находящейся в списке под номером t. Число строк этой матрицы (число квартир в списке) обозначим через n, стало быть t меняется в диапазоне от 1 до n.

Будем исходить из предположения о том, что цена квадратного метра получается умножением некоторой базовой цены на коэффициенты, зависящие от вышеперечисленных характеристик квартиры, а также на случайный коэффициент, соответствующий всем неучтенным свойствам товара и уровню ожиданий продавца.

После логарифмирования получаем модель

$$\ln y_t = \sum_{j=1}^{15} g_j(x_{tj}) + c + \eta_t. \tag{41}$$

Здесь c — константа, а $g_j(\cdot)$ — функции одной переменной, определяющие каждую конкретную модель. Предполагается, что случайные величины $\eta_1, \ldots, \eta_n$ независимы и одинаково распределены с нулевым математическим ожиданием и конечной ненулевой дисперсией.

Выбор $g_j \equiv 0$ означает, что модель не использует j-ую характеристику в качестве объясняющей переменной. 12-ая характеристика играет в нашей модели особую роль. Выбор $g_{12}(x) = \ln x$ соответствует предположению о том, что цена квадратного метра не зависит от общей площади квартиры.

В первой модели предполагается, что g_j — монотонные функции

специального вида, при этом в случае целочисленного аргумента значения аргумента заменяются для обеспечения монотонности. Во второй модели рассматриваются ступенчатые функции g_j целочисленных аргументов, а $g_{12}(x) = \ln x$. В третьей модели авторы отказываются от этого последнего предположения, подбирая функцию g_{12} специальным образом. Проверяется адекватность моделей в плане отсутствия систематических уклонений остатков.

Проведем предварительную подготовку данных — исключим строки, в которых не указана хотя бы одна из перечисленных выше шестнадцати характеристик.

Заметим, что непосредственное использование возрастающих функций g_i не адекватно рассматриваемым данным, для которых большинство характеристик являются кодами свойств квартир. От этих характеристик ожидаемая цена квартиры зависит немонотонным образом. Например, если названия улиц упорядочены в алфавитном порядке, то в этом случае, вообще говоря, неверно высказывание, что чем больше номер улицы, на которой находится квартира, тем больше ее цена, при прочих равных условиях. Поэтому для целочисленных характеристик проведем процедуру замены аргументов, состоящую в том, что каждое значение характеристики заменяется на соответствующую среднюю цену квадратного метра.

Через $w_t = y_t/x_{t,12}$ обозначим цену квадратного метра в квартире с номером t.

Обозначим через $a_{1j}, \ldots, a_{m_j j}$ все различные значения из множества $\{x_{1j} \ldots, x_{nj}\}$ аргументов характеристики с номером j. Здесь m_j — число различных значений этой характеристики. Пусть для определенности значения $a_{1j}, \ldots, a_{m_j j}$ упорядочены по возрастанию. Каждому значению a_{kj}, $k = 1, ..., m_j$ соответствует определенное множество значений w_t, для которых $x_{tj} = a_{kj}$. По этим множествам

вычислим средние значения $\overline{w_{1,j}}, ..., \overline{w_{m_j,j}}$.

Каждое число a_{ij} мы заменим на соответствующее ему значение $\overline{w_{i,j}}$: если $x_{tj} = a_{ij}$, то положим $\widehat{x}_{tj} = \overline{w_{i,j}}$. Тем самым мы вместо множества $\{x_{1j}, \ldots, x_{nj}\}$ получим новое множество $\{\widehat{x}_{1j}, \ldots, \widehat{x}_{nj}\}$ аргументов, соответствующих j-ой характеристике.

Для недискретной характеристики — площади квартиры — замена аргументов не производится: $\widehat{x}_{t,12} = x_{t,12}$, $t = 1, \ldots, n$.

Рассматривается следующая модель цены квартиры:

$$g_j(t) = c_j f(\alpha_j, \widehat{x}_{t,j}), \tag{42}$$

где

$$f(\alpha, x) = \begin{cases} \ln x, & \alpha = 0; \\ \frac{x^\alpha - 1}{\alpha}, & \alpha \neq 0, \end{cases}$$

где $c_1, c_2, ..., c_{15}, \alpha_1, ..., \alpha_{15}$ — неизвестные параметры. Мы исключаем из рассмотрения характеристики 13, 14, поскольку значения этих двух характеристик примерно пропорциональны общей площади. Также мы исключаем и код района, поскольку эта характеристика вполне определяется кодом улицы. Поэтому мы полагаем

$$g_2 \equiv 0, \ g_{13} \equiv 0, \ g_{14} \equiv 0,$$

или $c_2 = c_{13} = c_{14} = 0$.

Обратим внимание, что если $\alpha_1 = ... = \alpha_{15} = 0$, то в этом случае мы получим следующую модель цен:

$$y_t = \prod_{j=1}^{15} \widehat{x}_{t,j}^{c_j} \exp(c + \eta_t), \tag{43}$$

в этой модели цена меняется степенным образом в зависимости от известных значений характеристик $\widehat{x}_{t,j}$. Если же $\alpha_1 = ... = \alpha_{15} = 1$, мы получим такую модель:

$$\ln y_t = \sum_{j=1}^{15} c_j \widehat{x}_{t,j} + c + \eta_t, \tag{44}$$

в этом случае возникает обычная модель множественной регрессии с логарифмически сглаженными откликами (см. [??]).

Будем в дальнейшем предполагать, что $(\alpha_1, ..., \alpha_{15}) \in [0,1]^{15}$. Находить оценки параметров будем методом наименьших квадратов, минимизируя функцию

$$S(\alpha_1, ..., \alpha_{15}, c_1, ..., c_{15}, c) = \sum_{t=1}^{n} (\ln y_t - \sum_{j=1}^{15} c_j f(\alpha_j, \widehat{x}_{t,j}) - c)^2.$$

Отметим, что в силу сделанных предположений относительно $\eta_1, \ldots, \eta_n$, метод наименьших квадратов дает несмещенные оценки параметров с наименьшей дисперсией в линейной модели в классе всех линейных функций от наблюдений $\{\ln y_t\}_{t=1,\ldots,n}$.

Для каждого набора $(\alpha_1, ..., \alpha_{15})$ численно находим минимум функции $S(\alpha_1, ..., \alpha_{15}, c_1, ..., c_{15}, c)$ по параметрам $c_1, ..., c_{15}, c$, используя встроенную процедуру Excel. В итоге мы получаем

$$S^*(\alpha_1, ..., \alpha_{15}) = \min_{c_1,\ldots,c_{15},c} S(\alpha_1, ..., \alpha_{15}, c_1, ..., c_{15}, c).$$

Значения $(\alpha_1, ..., \alpha_{15}) \in [0,1]^{15}$, на которых достигает своего минимума функция $S^*(\alpha_1, ..., \alpha_{15})$, мы примем в качестве определяющих для модели (42).

Расчет производился по характеристикам 1, 3–12, 15. После вычислений были получены следующие значения:

j	1	3	4	5	6	7	8	9	10	11	12	15
$\tilde{\alpha}_j$	0	0	0	0	0	0	0	0	0	0	0,5	0
$\tilde{c}_j$	-0,38	0,89	0,12	0,32	0,37	0,53	-0,08	-0,14	-0,05	0,17	0,23	0,42

Здесь j — номер характеристики, $\tilde{\alpha}_j$, $\tilde{c}_j$ — найденные оценки параметров. Оценка константы $\tilde{c} = 9,56$. Выборочная дисперсия остатков $\widetilde{\sigma^2} = 0,021$. Обратим внимание, что вычисления проводились для данных за апрель, май и июнь, и для всех трех месяцев были получены значения параметров α_j, приведенные в таблице.

Поэтому в дальнейшем будем использовать значения $\alpha_1, ..., \alpha_{15}$ из таблицы. После фиксации этих коэффициентов модель описывает-

ся двенадцатью параметрами. Преимущество этой модели в сравнительно малом числе параметров и в реализации расчетов применением формальных процедур. Недостатки модели — отсутствие явной интерпретации оценок параметров и довольно большая выборочная дисперсия остатков. Кроме того, существенным недостатком модели является подверженность систематическим уклонениям остатков (см. рис. 6.4).

Рис. 6.4. Эмпирический мост для первой модели

Рассмотрим модель, в которой функции $g_j(\cdot)$ (см. (41)) целочисленных характеристик являются ступенчатыми, и число ступенек равно числу различных значений целочисленной характеристики. Для этих характеристик

$$g_j(x) = c_{ij}, \text{ если } x = a_{ij}, i = 1, \ldots, m_j \tag{45}$$

(обозначения a_{ij} и m_j были введены в описании предыдущей модели. Будем считать, что для каждого j значения a_{ij}, $i = 1, \ldots, m_j$,

упорядочены по возрастанию).

В этой модели необходимо принять решение об использовании нецелочисленных характеристик (x_{12} — полная площадь (в кв. м.); x_{13} — полезная площадь; x_{14} — площадь кухни).

Расссматривается модель $g_{12}(x) = \ln x$ (предполагая, что ожидаемая цена квадратного метра фактически не зависит от общей площади); $g_{13} \equiv 0$ (отказались использовать сведения о полезной площади, так как полезная площадь почти пропорциональна полной); вместо непрерывной характеристики x_{14} мы будем рассматривать дискретную характеристику $\mathbf{I}\{x_{14} \geq 10\}$ (ввели в рассмотрение индикатор большой площади кухни, надеясь по этой характеристике выявить квартиры в недавно построенных домах). Для дискретной характеристики $\mathbf{I}\{x_{14} \geq 10\}$ мы будем использовать в дальнейшем старое обозначение x_{14}, но теперь мы будем подразумевать под x_{14} дискретную характеристику. Характеристику «код района» мы исключаем из рассмотрения, ниже мы подробно разберем причины этого исключения. Поэтому мы полагаем, что $g_2 \equiv 0$.

Через $w_t = y_t/x_{t,12}$ мы будем по-прежнему обозначать цену квадратного метра в квартире с номером t, соответственно $z_t = \ln w_t$ — логарифм цены квадратного метра.

Итак, мы рассматриваем ровно 12 дискретных характеристик, номера этих характеристик: $\{1, 3 \ldots, 11, 14, 15\}$. Для удобства дальнейшего изложения перенумеруем эти характеристики: характеристике с номером 1 поставим в соответствие опять номер 1, характеристике с номером 3 поставим в соответствие новый номер 2, характеристике с номером 4 поставим в соответствие новый номер 3 и т.д., характеристике с номером 15 поставим в соответствие новый номер 12. Естественно, при этом изменятся индексы у функций g_j.

Далее для каждого элемента x_{tl} найдем число k_l такое что $a_{k_l l} =$

x_{tl}. Построим матрицу B размера $(n \times \sum_{j=1}^{12} m_j)$ следующим образом: каждый элемент матрицы B в t-ой строке и столбце с номером $\sum_{j=1}^{l-1} m_j + k_l,\ 1 \le k_l \le m_l,\ l = 1, \ldots, 12$ положим равным единице, а все остальные элементы положим равными нулю. Рассмотрим также вектор $\mathbf{d} = (c_{1,1}, c_{2,1}, ..., c_{m_1,1}, ..., c_{1,12}, ..., c_{m_{12},12})^T$. Отметим, что выполняется равенство:

$$\sum_{j=1}^{12} g_j(x_{tj}) = (B\mathbf{d})_t,$$

где $(B\mathbf{d})_t$ — t-ая строка матрицы $B\mathbf{d}$. Обратим внимание, что если для некоторой последовательности $l_1, ..., l_{12}$ такой, что $1 \le l_j \le m_j$, определить функции $r_j(\cdot) = g_j(\cdot) - c_{l_j j},\ j = 1, ..., 12$, в этом случае мы получим также ступенчатые функции, при этом $r_j(a_{l_j j}) = 0$. Заметим, что сумма $\sum_{j=1}^{12} g_j(x_{tj})$ будет отличаться от суммы $\sum_{j=1}^{12} r_j(x_{tj})$ лишь константой, поэтому в дальнейшем мы будем рассматривать функции g_j, в которых для некоторой последовательности $l_1, ..., l_{12}$ выполняется $c_{l_j j} = 0$ Последовательность $l_1, ..., l_{12}$ будем выбирать следующим специальным образом: рассмотрим ненулевые элементы s-ой строки матрицы B (мы брали $s = 20$, хотя можно взять любое значение от 1 до n). Пусть этим ненулевым элементам отвечают переменные $c_{l_j j},\ j = 1, ..., 12$, индексы l_j и дают нам необходимую последовательность. Поскольку мы полагаем $c_{l_j j},\ j = 1, ..., 12$ равными нулю, мы можем вычеркнуть столбцы матрицы B, в которых в s-ой строке находятся ненулевые элементы, при этом число столбцов новой матрицы B уменьшится ровно на 12. Кроме того, мы должны вычеркнуть и ровно 12 соответствующих компонент вектора $\mathbf{d}$. В итоге, равенство (41) мы можем переписать в виде:

$$z_t = (B\mathbf{d})_t + c + \eta_t. \tag{46}$$

Введем обозначения. Через $\mathbf{z}$ обозначим вектор $(z_1, ..., z_n)^T$, через

η обозначим вектор $(\eta_1, ..., \eta_n)^T$. Присоединим к матрице B вектор-столбец, составленный из единиц $(1, ..., 1)^T$ в качестве последнего столбца, получившуюся матрицу обозначим через D.

К вектору $\mathbf{d}$ присоединим константу c в качестве последней компоненты, полученный вектор обозначим через $\mathbf{e}$. Формулу (46) можно переписать так

$$\mathbf{z} = D\mathbf{e} + \eta. \tag{47}$$

Решение задачи методом наименьших квадратов $(\mathbf{z} - D\mathbf{e})^2 \to \min_{\mathbf{e}}$ приводит к системе линейных уравнений

$$D^t D\mathbf{e} = D^t\mathbf{z}, \tag{48}$$

причем размерность вектора $\mathbf{e}$ — число отличных от нуля параметров — равняется $\sum_{j=1}^{12} m_j - 11$.

Здесь сделаем одно очень важное замечание. Если бы мы не вычеркнули столбцы, соответствующие ненулевым элементам s-ой строки матрицы B, тогда бы вектор $(1, ..., 1)^T$ линейно выражался через вектор-столбцы матрицы B, что привело бы к неразрешимости системы (48), вычеркивание же упомянутых столбцов приводит к тому, что s-ая строка матрицы B будет состоять из нулей, поэтому вектор $(1, ..., 1)^T$ уже не может линейно выражаться через столбцы преобразованной матрицы B.

Отметим следующую особенность модели со ступенчатыми функциями. Для каждой функции g_j надо определить номер ступени l_j, которая имеет нулевую высоту: $c_{l_j j} = 0$, иначе система (48) не определена. Однако в некоторых ситуациях и этого не достаточно — требуется введение дополнительных нулевых значений. Для рассматриваемой модели такая ситуация возникает при рассмотрении характеристик «код улицы» и «код района». В типичном случае улица целиком содержится внутри района, но встречаются улицы, которые

относятся к двум районам. Таким образом, при совместном рассмотрении этих характеристик необходимо принять равной нулю высоту ровно одной ступени для улицы в группе районов, связанных общими улицами. Это очень неудобно в плане интерпретации результатов, и поэтому авторы отказались от использования характеристики «код района» в этой модели, поэтому полагаем, что $g_2 \equiv 0$.

По характеристикам квартир это число параметров распределилось следующим образом:

7 параметров соответствуют характеристике «число комнат» (при этом мы объединили характеристики 7 комнат и 8 комнат);

593 параметров — характеристике «код улицы»;

18 параметров — характеристике «этаж квартиры»;

22 параметров — характеристике «этажность дома»;

11 параметров — характеристике «код материала»;

4 параметра — характеристике «код планировки квартиры»;

8 параметров — характеристике «код типа квартиры»;

5 параметров — характеристике «код типа санузла»;

18 параметров — характеристике «код наличия балкона/лоджии»;

4 параметра — характеристике «код наличия телефона/линии»;

5 параметров — характеристика «код формы собственноси»;

1 параметр — «индикатор большой кухни» (того, что площадь кухни не менее 10 кв. м.);

1 параметр — константа (логарифм базовой цены квадратного метра).

Логарифм базовой цены квадратного метра вычисляется при значениях остальных характеристик, соответствующих нулевым значениям ступенчатых функций. Смысл каждого из остальных параметров — величина отклонения логарифма цены для соответствующего значения характеристики.

Отметим, что для вычисления матрицы D^tD вовсе не обязательно перемножать матрицы громадной размерности. Матрица формируется при однократном анализе строк исходных данных. Сначала матрица заполняется нулями. Если элементы с номерами j_1 и j_2 в строке с номером t принимают значения $a_{i_1j_1}$ и $a_{i_2j_2}$, то к соответствующему элементу матрицы D^tD прибавляется 1, затем происходит переход к следующей строке.

Полученное линейное уравнение относительно $\mathbf{e}$ решается методом Гаусса.

Оценка константы $\tilde{c} = 10,061$. Выборочная дисперсия остатков $\widetilde{\sigma^2} = 0,0189$. Эмпирический мост остатков приведен на рис. 6.5.

Рис. 6.5. Эмпирический мост для второй модели

Рассмотрим модель, в которой для целочисленных характеристик используются ступенчатые функции. Их, как и в предыдущей модели. Следуя нумерации характеристик, установленной для предыдущей модели, характеристике «общая площадь» будет соответство-

вать номер 13. Функцию $g_{13}(t)$ будем искать в виде

$$g_{13}(t) = \beta f(\alpha, x_{t,13}), \tag{49}$$

где

$$f(\alpha, x) = \begin{cases} \ln x, & \alpha = 0; \\ \frac{x^\alpha - 1}{\alpha}, & \alpha \neq 0, \end{cases}$$

β, α — неизвестные параметры.

Находить оценки параметров будем методом наименьших квадратов, минимизируя функцию

$$S(\beta, c_{11}, ..., c_{m_1 1}, c_{12}, ..., c_{m_2 2}, ..., c_{1,12}, ..., c_{m_{12}12}, c) = \sum_{t=1}^{n} (\ln y_t - \sum_{j=1}^{13} g_j(x_{tj}) - c)^2.$$

Итак, для всех $\beta \in [0, 1]$ численно находим минимум функции $S(\beta, c_{11}, ..., c_{m_1 1}, c_{12}, ..., c_{m_2 2}, ..., c_{1,12}, ..., c_{m_{12}12})$ по параметрам $c_{11}, ..., c_{m_1 1}, c_{12}, ..., c_{m_2 2}, ..., c_{1,12}, ..., c_{m_{12}12}$, исключая из рассмотрения переменные $c_{l_j j},\ j = 1, ..., 12$ — эти переменные для последовательности $\{l_j\}_{j=1..12}$ полагаются равными нулю. В итоге мы получаем

$$S^*(\beta) = \min_{c_{11}, ..., c_{m_{12}12}} S(\beta, c_{11}, ...c_{m_1 1}, c_{12}, ..., c_{m_2 2}, ..., c_{1,12}, ..., c_{m_{12}12}).$$

Значение $\beta \in [0, 1]$, на которых достигает своего минимума функция $S^*(\beta)$, мы примем в качестве определяющих для модели (49). В результате вычислений было получено значение β, равное $0,5$. Рассмотрим матрицу B, введенную в предыдущей модели. Присоединим к матрице B размера $(n \times \sum_{j=1}^{12} m_j - 12)$ вектор-столбцы $(\sqrt{x_{1,13}}, ..., \sqrt{x_{n,13}})^T$ и $(1, ..., 1)^T$ в качестве предпоследнего и последнего соответственно, получившуюся матрицу обозначим через D, к вектору $\mathbf{d}$ присоединим β и константу c в качестве предпоследней и последней компоненты соответственно, полученный вектор обозначим через $\mathbf{e}$. По аналогии с формулой (47) получим

$$\mathbf{z} = D\mathbf{e} + \eta. \tag{50}$$

Решение задачи методом наименьших квадратов $(\mathbf{z}-D\mathbf{e})^2 \to \min_{\mathbf{e}}$ приводит к системе линейных уравнений $D^T D\mathbf{e} = D^T\mathbf{z}$, причем размерность вектора $\mathbf{e}$ — число отличных от нуля параметров — равняется $\sum_{j=1}^{12} m_j - 10$.

Полученное линейное уравнение относительно $\mathbf{e}$ решается методом Гаусса.

Выборочная дисперсия остатков $\widetilde{\sigma^2} = 0,0176$.

Исследуем адекватность предложенных моделей. Эмпирическим мостом, построенным по значениям $\xi_1, \ldots, \xi_n$, называется непрерывная случайная ломаная с узлами в точках $(k/n,\ (nS_k - kS_n)/(n^{3/2}\tilde{\sigma}))$, $k = 0,\ 1, \ldots,\ n$. Здесь $S_0 = 0,\ \ S_k = \xi_1 + \ldots + \xi_k$, а $\tilde{\sigma}$ — выборочное среднеквадратическое отклонение: $\tilde{\sigma} = \sqrt{\overline{\xi^2} - \overline{\xi}^2}$. В случае, когда $\xi_1, \ldots, \xi_n$ — выборка из распределения с конечной ненулевой дисперсией, эмпирический мост C-сходится к стандартному броуновскому мосту при $n \to \infty$. В [164] проведено сравнение критериев обнаружения разладки в модели выборки. Критерий, основанный на максимуме модуля эмпирического моста, оказывается наилучшим в широком классе критериев с точки зрения относительной асимптотической эффективности по Питмену (см. главу 1).

Обозначим $J_n = \max_{0<k<n} |(nS_k - kS_n)/(n^{3/2}\tilde{\sigma})|$. Известно, что в модели выборки распределение J_n слабо сходится к распределению Колмогорова при $n \to \infty$.

Для остатков регрессии нет отмеченных сходимостей. Предельный процесс для эмпирического моста, построенного по регрессионным остаткам, устроен сложным образом, что не позволяет использовать распределение Колмогорова для вычисления достигнутого уровня значимости гипотезы об однородности остатков. Однако по виду эмпирического моста мы будем судить об адекватности модели, то есть об отсутствии или наличии систематических откло-

нений остатков. Формальный критерий будет состоять в подсчете эмпирического достигнутого уровня значимости: 500 раз переставляя случайным образом остатки регрессии и подсчитывая для них значения статистики J_n, вычислим эмпирический РДУЗ — частоту, с которой значение J_n для случайной перестановки превзойдет значение J_n, достигнутое исходной последовательностью остатков. Полученные частоты оказываются довольно близкими к значениям хвоста распределения Колмогорова, приведенными в последней строке таблицы.

Вычисления дают следующие значения.

модель	1	2	3
J_n	4,84	0,896	0,965
число случайных перестановок	500	500	500
число случаев превышения уровня J_n	0	195	144
эмпирический РДУЗ	0	0,390	0,288
хвост распределения Колмогорова	$9 \cdot 10^{-21}$	0,399	0,309

На основании проведенного анализа адекватности выяснено, что первая модель наиболее подвержена систематическим уклонениям остатков. Модель 2 в этом отношении оказывается несколько лучше модели 3.

Отметим, что используемый критерий эмпирического моста характеризует качество модели по отношению к фиксированному упорядочению. В нашем случае использовалось исходное упорядочение по возрастанию числа комнат, затем по названию района. Однако эксперименты с переупорядочиванием данных сначала по названию района, а затем по числу комнат дали близкие результаты, то есть и при другом переупорядочивании модель 2 не допускает больших систематических уклонений остатков.

5.9 Анализ моделей цен на автомобили на вторичном рынке

В этом параграфе анализируются объявления о продаже автомобилей «Тойота Королла» на сайте www.ngs.ru по данным на 02.06.2012 (всего 525 объявлений). Стандартный регрессионный анализ показывает, что едиственные значимые факторы — положение руля и год выпуска. Исследуем зависимость цены Y_i от года выпуска X_i.

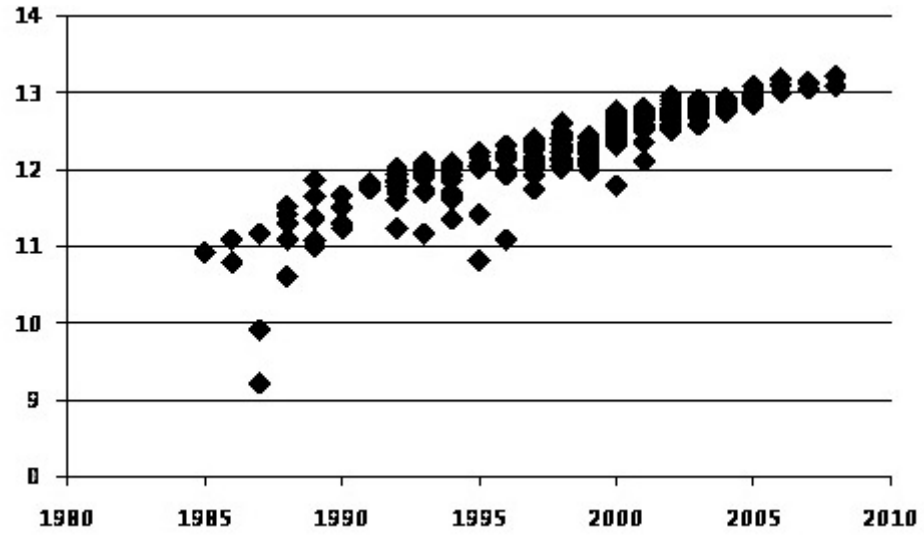

Рис. 6.6. Логарифмы цен в зависимости от года выпуска

Используется модель

$$\ln Y_i = aX_i + b + \varepsilon_i,\ i = 1, \ldots, n. \tag{51}$$

Оценки параметров (с точностью до 4 знаков)

$$\widehat{a} = 0.1089,\ \widehat{b} = -205.3.$$

Оценим Y_i и вычислим регрессионные остатки. Их выборочное стандартное отклонение $S = 0,2469$. Удаляем последовательно объ-

явления с остатками, выходящими за рамки 3 сигм. В результате осталось 364 объявлений,

$$\hat{a} = 0.09558,\ \hat{b} = -178.7,\ S = 0.1291.$$

Мы уменьшили S почти вдвое. Вычисляем эмпирический мост и обнаруживаем, что он имеет острые пики в точках 16 и 154. Не будем анализировать значения с 1 по 16 из-за малого числа точек. Остальные интервалы несколько скрректируем: с 17 до 148 точки соответствуют годам с 1991 до 1999, с 149 до 364 точки — годам с 2000 до 2008.

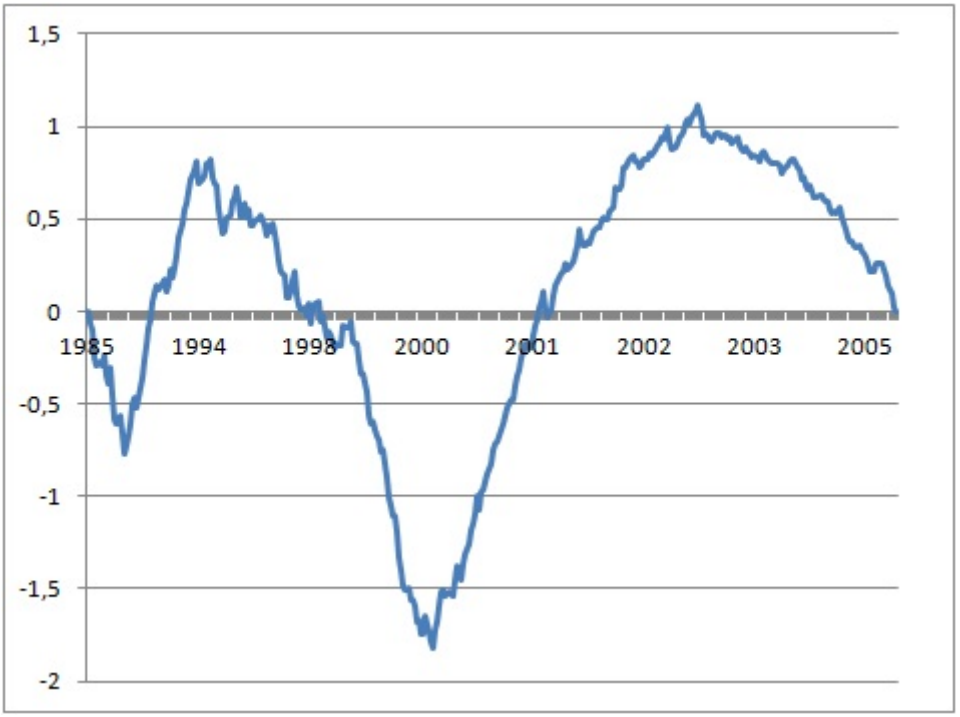

Рис. 6.7. Эмпирический мост остатков логарифмов цен на автомобили

Для первого интервала

$$\hat{a} = 0.06484,\ \hat{b} = -117.3,\ S = 0.1356.$$

Для второго

$$\hat{a} = 0.07144,\ \hat{b} = -130.3,\ S = 0.08699.$$

Протестируем предложенные модели.

Пусть $\widehat{Y_i} = \hat{a} + \hat{b}X_i$, $\hat{\varepsilon}_i = Y_i - \widehat{Y_i}$, $\widehat{\Delta}_i^0 = \hat{\varepsilon}_1 + \ldots + \hat{\varepsilon}_i$.

Обозначим $GL_F(t) = \int\limits_0^t F^{-1}(s)\,ds$, $GL_n(t) = \frac{1}{n}\sum_{i=1}^{[nt]} X_i$.

Используем теорему из [**?**]:

Theorem 6.1 *Пусть X_i — порядковые статистики из выборки $\{\varepsilon_i\}_{i=1}^n$ с ф. р. F, и $\{\varepsilon_i\}$ и $\{\xi_i\}$ независимы. Если $0 < \mathbf{D}\xi_1 < \infty$, то $\widehat{Z_n} \Longrightarrow Z_F^0$ Z_F^0 — гауссовский процесс (центрированный) со структурной функцией*

$$K_F^0(t,u) = \min\{t,u\} - tu - \frac{GL_F^0(t)GL_F^0(u)}{\mathbf{Var}\xi_1},\ t,u \in [0,1].$$

Пусть

$$K_n^0(t,u) = \min\{t,u\} - tu - \frac{GL_n^0(t)GL_n^0(u)}{S^2},\ t,u \in [0,1].$$

Тогда $K_n^0(t,u) \to K_F^0(t,u)$ равномерно при $n \to \infty$.

Пусть

$$\mathbf{a} = (a_1,\ldots,a_d) = \left(\frac{1}{d+1},\ldots,\frac{d}{d+1}\right),$$

$$G = \left(K_F^0(a_i,a_j)\right)_{i,j=1}^d,\quad G_n = \left(K_n^0(a_i,a_j)\right)_{i,j=1}^d,$$

$$q = (\widehat{Z_n}(a_1),\ldots,\widehat{Z_n}(a_d))^T.$$

Следствие 6.1 *Если выполнены условия теоремы 6.1, и матрица G^{-1} существует, то распределение статистики $q^T G_n^{-1} q$ сходится к χ^2-распределению с d степенями свободы.*

РДУЗ вычисляется по формуле $\alpha^* = 1 - F_{\chi_d^2}(q^T G_n^{-1} q)$.

Выберем $d = [n^{1/3}] + 1$. Для всей выборки $n = 364$, $d = 8$, для первой и второй части $n_1 = 132$, $d_1 = 6$, $n_2 = 216$, $d_2 = 7$ (годы 1991–1999 и 2000–2008).

Вычисления дают $\alpha^* << 10^{-4}$ для всей выборки, $\alpha_1^* = 0.1677$ для первой части, $\alpha_2^* = 0.07505$ для второй части. Поэтому модель

Year	1991	1992	1993	1994	1995	1996	1997	1998	1999
Price	130	139	148	158	169	180	192	205	219
Year	2000	2001	2002	2003	2004	2005	2006	2007	2008
Price	290	312	335	360	386	415	446	479	514

отвергается для всей выборки и принимается для интервалов 1991–1999 и 2000–2008.

Вычислим оценки цены в тысячах рублей.

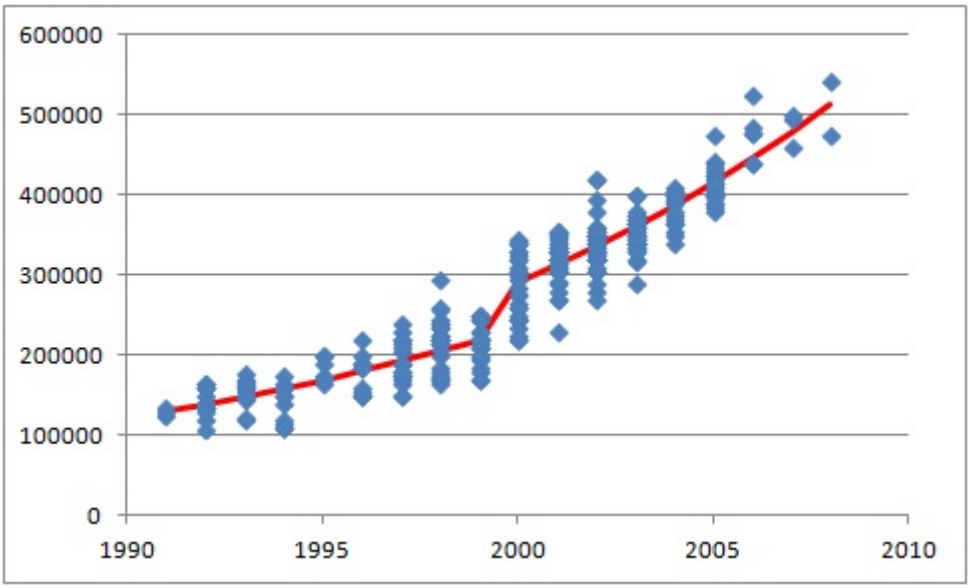

Рис. 6.8. Скорректированная модель цены автомобиля

5.10 Анализ дефектов строительных конструкций

Локализация дефекта строительной конструкции во времени и в пространстве имеет определяющее значение для его своевременного устранения и предотвращения разрушения. В настоящее время разработаны методы контроля колебаний строительных конструкций и их математического моделирования методом конечных элементов

[96]. В нашей работе исследуется статистический критерий, позволяющий по собранным дан-ным делать вывод о моменте образования дефекта, а также о его пространственной локализации. Этот критерий основан на изучении эмпирического моста — случайной ломаной, введенной в статье [164]. Основная гипотеза предполагает, что неоднородности нет. Возникновение неоднородности должно быть обнаружено внешними средствами — результатами математической обработки записей колебаний датчиков, расположенных на фиксированных этажах.

Исходные данные, на основании которых надо принять решение о наличии неоднородности — это результаты моделирования колебаний здания, которые физически описываются как мгновенные значения ускорений, а математически — как связанные между собой временные ряды колебаний с аддитивным шумом на выбранных этажах по каждой из осей координат. Исследуется модельная задача искусственного введения неоднородности путем утяжеления одного из этажей вдвое. Специфика изучаемых в работе процессов состоит в том, что на периодические колеба-ния накладывается случайный шум. В этой ситуации распределение эмпирического моста отличается от распределения для модели простой случайной выборки. Этот вопрос специально изучался в статье [199]. Как показано там, логарифмические асимптотики распределения максимального отклонения эмпирического моста от оси абсцисс в этих двух ситуациях совпадают, а при практическом вычислении вероятности следует вводить поправки, не превосходящие 29%. В связи с этим достигаемые уровни значимости вычисляются в данной работе на основании распределения Колмогорова с точностью до одной значащей цифры. Момент разладки определяется как абсцисса наибольшего отклонения эмпирического моста от оси абсцисс на основании работы [115].

Исходные данные — это результаты моделирования колебаний здания для уровней -2 (минус второго), 10, 20, 30 и 40 этажа. Используются два режима моделирования: без дефекта и с дефектом. В качестве дефекта рассматривается здание, в котором масса одного из этажей (а именно, 25-го этажа) в 2 раза больше, чем в исходной модели. Колебания (значения ускорений) измеряются по 3 координатам с временным шагом 0,01 с в течение 4 минут (всего 24000 отсчетов). В качестве колебаний на уровне -2 взяты результаты записи, осуществленные на реальном объекте и описанные в работе [96]. Колебания здания возникают из-за колебаний основания. Моделирование колебаний под этим воздействием начинается из состояния покоя. Оси координат привязаны к осям здания: ось x направлена вдоль фронтальной оси здания. Графики колебаний приведены в работе [96]. Для модели без дефекта характерны более быстрый рост и более медленное затухание колебаний, чем для модели с дефектом. Особенно велики отличия для колебаний вдоль вертикальной оси. Они сильно различаются для разных режимов, но изучение графиков не дает ответа на вопрос, какой этаж имеет массу, вдвое превышающую массу остальных. Через $x_d,\ y_d, z_d$ будем обозначать координаты колебаний на соответствующем уровне при наличии дефекта, а через $x_n,\ y_n,\ z_n$ — при отсутствии дефекта. Для каждого уровня от 10 до 40 вычислим квадраты амплитуд колебаний и найдем их разности $r = x_d^2 + y_d^2 + z_d^2 - (x_n^2 + y_n^2 + z_n^2)$. Для минус второго этажа эти вычисления не проводятся, так как для него нагруженный и ненагруженный случаи не отличаются — разности равны нулю. Вычислим кумулятивные суммы $r_L = cumsum(r)$. Разделим кумулятивные суммы на выборочное стандартное отклонение и на корень квадратный из числа отсчетов $n = 24000$. Получим нормированные кумулятивные суммы $r_N = (var(r)n)^{-1/2} r_L$.

Были изучены амплитуды спектров сигналов по каждой оси координат на каждом уровне от 10 до 40. Так как замеры произведены в течение 240 секунд с дискретностью 0,01 с, то дискретное преобразование Фурье дает частоты от 1/240 Гц до 50 Гц с шагом 1/240 Гц. Амплитуды вычисляются делением значений дискретного преобразования Фурье на число отсчетов (24000). Анализ показывает расщепление пиковой частоты, приближенно равной 5 Гц, для колебаний по оси Z в моделях с дефектом: между двумя максимумами амплитуд образуется минимум. В результате существенно понижается высота спектрального пика. Кроме того, происходит сдвиг (уменьшение) пиковой частоты. Так, на уровне 20-го этажа значение пиковой частоты изменилось с 5.07 Гц в модели без дефекта до 4.99 Гц в модели с дефектом. Эти отличия определяют возможность детектирования дефекта (нагруженного этажа) спектральным методом.

Анализ кумулятивных сумм на разных уровнях показывает рост зашумления с ростом высоты. Для того, чтобы анализировать зашумление, осуществим ортогонализацию кумулятивных сумм S по соответствующим координатам с кумулятивными суммами S_0 на уровне -2 по формуле $S_0 = S - < S, S_0 > S_0 / < S_0, S_0 >$.

От ортогонализованных сумм S_0 перейдем к их приращениям, то есть к исправленным значениям процесса x^0, y^0, z^0, а затем к приращениям процесса Δx^0, Δy^0, Δz^0. Просуммируем квадраты приращений по всем координатам и по всем уровням: $v(i) = \sum_j((\Delta x_j^0)^2 + (\Delta y_j^0)^2 + (\Delta z_j^0)^2)$. Здесь суммирование ведется по всем уровням $j > 0$. По этим суммам построим эмпирический мост: ломаную с узлами в точках $(k/n, (\sum_{i \le k} v(i) k \sum_{i \le n} v(i)/n)\sigma^{-1} n^{-1/2})$.

Здесь σ - выборочное среднеквадратическое отклонение для $\{v(i)\}_{i \le n}$. Для интервала стационарности характерно линейное по-

ведение эмпирического моста, что соответствует постоянному среднему значению (в данном случае квадратов приращений). Можно выделить 4 интервала линейного поведения эмпирического моста как для модели без дефекта, так и для модели с дефектом. Границы интервалов — это точки, в которых происходит излом эмпирического моста — одна аппроксимирующая линейная функция сменяется другой. Эти интервалы отыскиваются аналитически как моменты наибольшего отклонения эмпирического моста от горизонтальной оси. В эти моменты временной ряд разрезается на 2 части, и процедура выполняется снова. В результате этой процедуры найдены следующие границы интервалов: в модели без дефекта 1, 7304, 9392, 16816, 24000; в модели с дефектом 1, 7312, 9388, 16284, 24000. При этом величина отклонения на последнем интервале составляет 4.54 для модели без дефекта и 4.31 для модели с дефектом. Для дальнейшего анализа берется пересечение этих интервалов, то есть интервал от 16816 до 24000, содержащий 7185 значений. Этот интервал соответствует стационарному режиму, когда система, стартовавшая из неподвижного состояния, пришла в состояние стохастической стабильности. Построим эмпирические мосты по квадратам приращений на каждом положительном уровне по каждой координате. Проанализируем поведение эмпирического моста соответствующим статистическим критерием: достигнутый уровень значимости вычисляется как «хвост» распределения Колмогорова.

Этаж	Без дефекта		С дефектом	
	Значение статистики	Достигнутый уровень значимости	Значение статистики	Достигнутый уровень значимости
10	1,284	0,07	1,381	0,04
20	1,136	0,2	1,879	0,002
30	1,370	0,05	1,794	0,003
40	1,540	0,02	2,633	0,000002

Таким образом, критерий эмпирического моста (при выборе критического уровня 0,01) принимает гипотезу об однородности наблюдений для всех этажей в модели без дефекта, и отвергает для всех этажей, начиная с 20-го, в модели с дефектом. Наиболее сильное отличие зафиксировано для модели с дефектом на 40-м этаже. Итак, возможно использовать данный статистический критерий для выявления наличия дефектов и их пространственного положения.

Согласно общей теории, развитой в статье [199], максимальное отклонение эмпирического моста при основной гипотезе имеет распределение, близкое к распределению Колмогорова, а при альтернативной растет пропорционально величине неоднородности (в нашем случае Ц– изменению массы этажа). Поэтому при снижении величины отклонения от нормы со 100% до 50% и 25% следует ожидать, что применяемая статистика критерия уменьшится приблизительно в 2 и 4 раза соответственно. Так как для распределения Колмогорова достигнутый уровень значимости имеет асимптотику $\alpha \sim 2\exp(-2x^2)$, то в случае, когда для 100%-го отклонения он достаточно мал, для 50% и 25% он будет соответственно составлять примерно $\alpha^{1/4}$ и $\alpha^{1/16}$. В частности, если $\alpha < 10^{-6}$, то для отклонения в 50% гипотеза об однородности отвергается на уровне значимости 0,032, а для отклонения в 25% - лишь на уровне 0,42. Таким

образом, следует считать разработанный подход приемлемым при 50% отклонения массы этажа от номинала и малоприемлемым при 25%.

В результате проведенного анализа проведены исследования по возможности обнаружения дефекта, состоящие из следующих шагов.

1. Вычисляются кумулятивные суммы значений по осям координат.

2. Кумулятивные суммы ортогонализуются с кумулятивными суммами в подвале (на минус втором этаже), то есть с управляющим воздействием.

3. От ортогонализованных кумулятивных сумм переходим к значениям процесса, а затем к приращениям процесса.

4. Квадраты приращений процесса суммируются по всем координатам и этажам.

5. По этим суммированным данным строится эмпирический мост, и определяется интервал стационарности (на этом интервале максимум модуля эмпирического моста не должен превышать 5).

6. На интервале стационарности строятся эмпирические мосты по квадратам прираще-ний процесса на каждом этаже и по каждой координате.

7. Процесс считается неоднородным, если достигнутый уровень значимости принимает значения меньше 0,01, то есть максимум модуля эмпирического моста больше 1,63.

8. Момент появления неоднородности определяется как момент наибольшего отклонения эмпирического моста от оси абсцисс.

Пункты 1-2 обосновываются необходимостью изучать не управляющее воздействие, а реакцию конструкции на него. Пункты 3-4 необходимы в силу того, что процесс представляет собой быстрые

колебания, и отличия нагруженного и ненагруженного режимов могут быть найдены только при анализе динамики амплитуд этих колебаний. Пункт 5 необходим для определения интервала стационарности процесса, так как вне этого интервала методы обнаружения разладки не работают. Пункты 6-7 описывают методику применения эмпирического моста к анализу однородности [199]. Пункт 8 описывает процедуру отыскания состоятельной оценки момента разладки [115]. Отметим, что для практического применения предлагаемого алгоритма требуется отладка на данных, полученных не в результате моделирования, а в результате прямой записи наблюдений. Проблема здесь состоит в том, что трудно получить наблюдения с наличием дефекта: как описано выше, на модельных данных чувствительность метода позволяет обнаруживать, что масса этажа увеличилась на 50%, но для реальных объектов такое увеличение массы просто недопустимо или крайне трудоемко. Есть предположение (следующее из всей логики статистического анализа), что по более длинной записи наблюдений удастся обнаружить значительно меньшие отклонения массы, однако (ввиду трудоемкости загрузки здания) опыты следует начать либо с масштабных моделей зданий, либо с наблюдений за естественным изменением массы (выпадение снега на крышу, возможно, дает неоднородность требуемого масштаба).

5.11 Результаты главы 6

В главе 6 продемонстрированы применения статистических критериев, основанных на методе эмпирического моста, к выбору вероятностной модели.

- Анализируются две модели медицинских данных: зависимости

концентрации маркеров от массы тела. Отдается предпочтение более простой модели как более логичной и вполне удовлетворяющей статистическому критерию.

- Проанализированы три модели цены квадратного метра квартиры в зависимости от характеристик, приведенных в справочнике по недвижимости. Показано, что первая модель (в отличие от второй и третьей) отвергается статистическим критерием на низком уровне значимости.
- Обнаружено скачкообразное изменение характеристик модели при анализе зависимости цены автомобиля от года выпуска. Показано, что исправленная модель хорошо согласуется с данными.
- Разработан алгоритм анализа неоднородности строительной конструкции по записи ее колебаний.

ЗАКЛЮЧЕНИЕ

В заключении приведены основные результаты работы, составляющие научную новизну, теоретическую и практическую значимость. Обоснована достоверность результатов и сформулированы положения, выносимые на защиту.

Научная новизна

1. Впервые введен широкий класс алгоритмов апостериорного обнаружения разладки в модели выборки с единых позиций: алгоритмы построены на основании функционалов от эмпирического моста. Рассмотрены функционалы, основанные на взвешенных суммах; L_p-нормы и их модификации; L_∞-норма и размах эмпирического моста.

2. Впервые проведено сравнение алгоритмов апостериорного обнаружения разладки в модели выборки с точки зрения их относительной асимптотической эффективности по Питмену. Выбран наилучший критерий, основанный на L_∞-норме эмпирического моста.

3. Впервые разработан алгоритм применения статистического критерия, основанного на норме эмпирического моста, к анализу однородности текста на естественном языке. Выявлена зависимость реально достигаемого уровня значимости критерия от объема текста, приводящая к неадекватности модели выборки для текста большого объема.

4. Впервые предложены модели фрактального броуновского моста и склейки фрактальных броуновских движений для временного ряда, построенного по тексту на естественном языке.

5. Впервые предложены центрированный знаковый метод, модифицированный знаковый метод и бинарный знаковый метод оценивания параметра Херста. Построен статистический критерий проверки гипотезы фрактальности.

6. Впервые статистический критерий проверки гипотезы фрактальности применен для проверки фрактальности текстов на естественном языке.

7. Впервые построен статистический критерий проверки разладки фрактального гауссовского шума, основанный на разности оценок параметра Херста. Критерий применен к анализу однородности текста на естественном языке. Разработан алгоритм выявления склейки текстов.

8. Впервые разработан класс критериев обнаружения разладки регрессии с циклическим трендом, основанных на значениях эмпирического моста. На численном примере проведено сравнение мощностей разработанных критериев.

Теоретическая значимость

1. Метод построения критериев наличия разладки на основании функционалов от эмпирического моста систематизирует процедуры построения таких критериев и может быть использован для исследования разладки в более сложных, в том числе регрессионных, моделях.

2. Процедура сравнения критериев наличия разладки и выявления лучшего в широком классе может быть применена к другим вероятностным моделям.

3. Разработанные модификации знаковых методов оценивания параметра Херста становятся классом методов, инвариантных от-

носительно строго монотонных преобразований пространства значений.

4. Построены статистические модели временного ряда, образованного по тексту на естественном языке, в однородном и неоднородном случаях. Эти модели протестированы на адекватность методами математической статистики. Проведенный анализ позволяет выбирать правильную модель в зависимости от объема текста, а также различать однородные и неоднородные тексты.

5. Разработаны статистические критерии обнаружения разладки регрессии с циклическим трендом.

Практическая значимость

1. Разработаны критерии апостериорного обнаружения разладки в модели выборки, позволяющие решать задачи об однородности случайных последовательностей. Обоснован выбор критерия, имеющего наибольшую относительную асимптотическую эффективность по Питмену, и потому наиболее полезного при практическом различении близких гипотез, то есть в ситуации, когда математические ожидания до и после разладки различаются незначительно.

2. Разработаны статистические критерии, позволяющие по статистике пермен знаков оценивать параметр Херста. В частности, принимать или отвергать модель фрактального броуновского движения. Также разработаны критерии, позволяющие диагностировать разладку в модели фрактального броуновского движения.

3. Методика применения разработанных статистических критериев к анализу однородности текста на естественном языке позво-

ляет анализировать тексты на однородность.

4. Предложенные статистические критерии позволяют тестировать наличие разладки регрессионных моделей с циклическим трендом. В частности, они применялись в качестве решающих правил для выявления наличия или отсутствия изменений прочностных характеристик высотных и уникальных зданий на основании записей их колебаний, выполненных НПО «Содис» (г. Москва).

5. Предложенные критерии позволяют обнаруживать разладку в регрессионных моделях. Так, они применялись для выбора модели зависимости концентрации маркеров в крови от массы тела; зависимости цены квартиры от ее характеристик; зависимости цены автомобиля от года выпуска.

Достоверность результатов диссертации подтверждается их совпадением в частных случаях с результатами расчетов, выполненных другими авторами и с помощью других методов. Теоретические результаты опубликованы в ведущих журналах, докладывались на крупных международных конференциях и представлены в их публикациях. Они известны в научном сообществе и цитируются в работах других авторов.

На защиту выносятся

1. Построение класса статистических критериев (решающих правил) и соответствующих им алгоритмов апостериорного обнаружения разладки в модели выборки на основании функционалов от эмпирического моста.

2. Сравнение алгоритмов апостериорного обнаружения разладки в

модели выборки с точки зрения их относительной асимптотической эффективности по Питмену.

3. Алгоритм применения статистического критерия, основанного на норме эмпирического моста, к анализу однородности текста на естественном языке. Выявление зависимости реально достигаемого уровня значимости критерия от объема текста, приводящая к неадекватности модели выборки для текста большого объема.

4. Модели фрактального броуновского моста и склейки фрактальных броуновских движений для временного ряда, построенного по тексту на естественном языке.

5. Центрированный знаковый метод, модифицированный знаковый метод и бинарный знаковый метод оценивания параметра Херста. Построение решающего правила и алгоритма проверки гипотезы фрактальности.

6. Алгоритм проверки фрактальной модели для текстов на естественном языке с помощью специально разработанного статистического критерия.

7. Разработка алгоритма проверки разладки фрактального гауссовского шума, основанного на разности оценок параметра Херста, его применение к анализу однородности текста на естественном языке. Разработка алгоритма выявления склейки текстов.

8. Функциональная предельная теорема для числа разных элементов выборки и алгоритмы ее применения к анализу текстов.

9. Алгоритмы обнаружения разладки регрессии на порядковые статистики и регрессии с циклическим трендом, основанные на функционалах от эмпирического моста.

Автор выражает глубокую признательность своим учителям Ю. Е. Хайкину, А. И. Саханенко, С. Г. Фоссу, В. А. Селезневу.

Список литературы

[1] Андерсон Т. Введение в многомерный статистический анализ. — М., 1963.

[2] Андерсон Т. Статистический анализ временных рядов. Ц– М.: Мир, 1976.

[3] Анго А. Математика для электро- и радиоинженеров. — М., 1964.

[4] Аркашов Н. С., Борисов И. С. Гауссовская аппроксимация процессов частных сумм скользящих средних // Сиб. мат. журнал, Т. 45, N 6, 2004. — С. 1221–1255.

[5] Бендерская Е. Н., Жукова С. В. Обработка текстовой информации с использованием хаотической нейронной сети // В сб.: «Квантитативная лингвистика: исследования и модели (КЛИМ - 2005)», материалы Всероссийской научной конференции. Новосибирск, НГПУ. 2005. — С. 271–282.

[6] Биллингсли П. Сходимость вероятностных мер. — М.: «Наука», 1977.

[7] Бойко Ю. П. Селективность авторского стиля на уровне предложения // В сб.: «Квантитативная лингвистика: исследования и модели (КЛИМ - 2005)», материалы Всероссийской научной конференции. Новосибирск, НГПУ. 2005. — С. 303–315.

[8] Боровков А. А. Теория вероятностей. — М.: «Эдиториал УРСС», 1999.

[9] Боровков А. А. Математическая статистика. — Новосибирск: «Наука», 1997.

[10] Бродский Б. Е., Дарховский Б. С. Асимптотический анализ некоторых оценок в апостериорной задаче о разладке // Теория вероятн. и ее примен., 35:3, 1990. — С. 551–557.

[11] Бродский Б. Е., Дарховский Б. С. Сравнительный анализ некоторых непараметрических методов скорейшего обнаружения момента «разладки» случайной последовательности // Теория вероятн. и ее примен., 35:4, 1990. — С. 655–668.

[12] Бродский Б. Е., Дарховский Б. С. Алгоритм апостериорного обнаружения многократных разладок случайной последовательности // Автомат. и телемех., N . 1, 1993. — С. 62–67.

[13] Бродский Б. Е., Дарховский Б. С. Проблемы и методы вероятностной диагностики // Автомат. и телемех., N 8, 1999. — С. 3–50.

[14] Вальд А. Последовательный анализ. — М.: «Физматлит», 1960.

[15] Васильченко С. Г. Алгоритм обнаружения моментов разладки случайной последовательности // Фундамент. и прикл. матем., 8:3, 2002. — С. 655-Ц665.

[16] Ватсон Г. Н. Теория бесселевых функций. ч. 1-2. — М., 1949.

[17] Вашак П. Длина слова и длина предложения в текстах одного автора // Вопросы статистической стилистики — Киев: Наукова думка, 1974. — С. 314–329.

[18] Виноградов В. В. Проблема авторства и теория стилей. — М., 1961.

[19] Водный кадастр СССР. Т. VI. — М., 1968.

[20] Воеводин В. В., Тыртышников Е. Е. Вычислительные процессы с теплицевыми матрицами. — М.: «Наука», 1987.

[21] ГОСТ Р ИСО 5479-2002 — Статистические методы. Проверка отклонения распределения вероятностей от нормального распределения. М., 2002.

[22] «Д». Стремя «Тихого Дона». Загадки романа. — Paris: YMCA Press, 1974.

[23] Давыдов Ю. А. Принцип инвариантности для стационарных процессов // Теория вероятностей и ее применения. Т. 24, N 3, 1970. — С. 487–498.

[24] Дарховский Б. С. Непараметрический метод для апостериорного обнаружения момента «разладки» последовательности независимых случайных величин, Теория вероятн. и ее примен., 21:1,1976. — С. 180–184.

[25] Дарховский Б. С. Задача о неопределенной «разладке» случайной последовательности // Теория вероятн. и ее примен., 56:1, 2011. — С. 30–46.

[26] Дарховский Б. С. Обнаружение разладки случайной последовательности при минимальной априорной информации // Теория вероятн. и ее примен., 58:3, 2013. — С. 585–590.

[27] Дарховский Б. С., Бродский Б. Е. Апостериорное обнаружение момента «разладки» случайной последовательности // Теория вероятн. и ее примен., 25:3, 1980. — С. 635–639.

[28] Дарховский Б. С., Бродский Б. Е. Непараметрический метод скорейшего обнаружения изменения среднего случайной последовательности // Теория вероятн. и ее примен., 32:4, 1987. — С. 703–711.

[29] Дарховский Б. С., Пирятинская А. Новый подход к проблеме сегментации временных рядов произвольной природы // Сто-

хастическое исчисление, мартингалы и их применения, Сборник статей. К 80-летию со дня рождения академика Альберта Николаевича Ширяева, Тр. МИАН, 287, МАИК, М., 2014. — С. 61–74.

[30] Дрейпер Н.Р., Смит Г. Прикладной регрессионный анализ. –Ц М.: Диалектика, 2007.

[31] Ермолаев Г. Загадки «Тихого Дона» // Slavic and European Journal, V. 18, N 3, 1974.

[32] Ермолаев Г. Кто написал «Тихий Дон» // Slavic and European Journal, V. 20, N 3, 1976.

[33] Ермоленко Г. В. Лингвистическая статистика. Краткий очерк и библиографический указатель. — Алма-Ата, 1970.

[34] Закревская Н. С. Вероятностные модели текста // В сб.: Материалы IXL международной научной студенческой конференции «Студент и научно-технический прогресс». Математика. Новосибирск, НГУ, 2001. — С. 136.

[35] Закревская Н. С. Оценивание параметра дробного броуновского движения. // В сб.: «Наука. Техника. Инновации.» Материалы докладов региональной научной конференции студентов, аспирантов, молодых ученых. Часть 1. Новосибирск, НГТУ, 2002. — С. 199–200.

[36] Закревская Н. С. Сравнение марковских и элементарных вероятностных моделей статистик текста // В сб.: Материалы XL международной научной студенческой конференции «Студент и научно-технический прогресс». Математика. Новосибирск, НГУ, 2002. — С. 142–143.

[37] Закревская Н. С. Вероятностные модели текста // В сб.: Информатика и проблемы телекомуникаций. Международная научно-техническая конференция. Материалы конференции. Новосибирск, СибГУТИ, 2002. — С. 120–121.

[38] Закревская Н. С. Проверка гипотезы о фрактальном броуновском движении // В сб.: «Наука. Технологии. Инновации.» Материалы докладов всероссийской научной конференции молодых ученых. Часть 1. НГТУ, Новосибирск, 2003. — С. 221–222.

[39] Закревская Н. С. Оценивание параметра фрактального броуновского моста // В сб.: «Наука. Технологии. Инновации.» Материалы всероссийской научной конференции молодых ученых. Часть 1. Новосибирск, НГТУ, 2004. — С. 210.

[40] Закревская Н.С. Исследование однородности текста с помощью модели скользящего среднего // В сб.: «Квантитативная лингвистика: исследования и модели (КЛИМ - 2005)», материалы Всероссийской научной конференции. Новосибирск, НГПУ, 2005. — С. 26–34.

[41] Золотарев В.М. Современная теория суммирования независимых случайных величин. — М.: Наука, 1986.

[42] Ибрагимов И. А., Линник Ю. В. Независимые и стационарно связанные величины. М.: Наука, 1965.

[43] Ито К., Маккин Г. Диффузионные процессы и их траектории. — М.: Мир, 1968.

[44] Карлин С. Основы теории случайных процессов. — М.: Мир, 1971.

[45] Кац М. Вероятность и смежные вопросы в физике. — М.: Мир, 1965.

[46] Кашьяп Р. Л., Рао А. Р. Построение динамических стохастических моделей по экспериментальным данным. — М.: Наука, 1983.

[47] Кйецаа Г. Борьба за «Тихий Дон». — Pergamob Press, USA, 1977.

[48] Клигене Н., Тельксниc Л. Методы обнаружения моментов изменения свойств случайных процессов // Автоматика и телемеханика. — 1983. — N 10, с. 5–56.

[49] *Колмогоров А. Н.* Кривые в гильбертовом пространстве, инвариантные относительно однопараметрической группы движений // Доклады Академии Наук СССР, Т. 26, N 1, 1940. — С. 6–9.

[50] Колодный Л. Вихри над «Тихим Доном». Фрагменты прошлого: истоки одного навета XX века // Московская правда, 5 и 7 марта 1989.

[51] Корн Г., Корн Т. Справочник по математике для научных работников и инженеров. — М., 1984.

[52] Крамер Г., Лидбеттер М. Стационарные случайные процессы. — М., 1969.

[53] Кроновер Р. М. Фракталы и хаос в динамических системах. Основы теории. — М.: Постмаркет, 2000.

[54] Кукушкина О.В., Поликарпов А.А., Хмелев Д.В. Определение авторства текста с использованием буквенной и грамматической информации // Проблемы передачи информации, Т. 37, вып. 2, 2001. — С. 96–108.

[55] Кукушкина О. В., Макаров А. Г., Поддубный В. В., Поликарпов А. А., Шевелев О. Г. Анализ количественных характиристик авторского стиля романа «Тихий Дон» и его соотношение с другими текстами М. А. Шолохова на основе иерархической кластеризации // Загадки и тайны "Тихого Дона": двенадцать лет поисков и находок. — М. : «АИРО — XXI», 2010. — С. 127–130.

[56] Кэрролл Л. Алиса в стране чудес. Алиса в зазеркалье. — М., «Правда» 1982.

[57] Кэрролл Л. Алиса в Стране Чудес. — Новосибирск, Новосибирское книжное издательство, 1987.

[58] Леви П. Стохастические процессы и броуновское движение. — М.: Наука, 1972.

[59] Левин Б. Р. Теоретические основы статистической радиотехники. — М., 1989.

[60] Лемешко Б. Ю., Постовалов С. Н. Применение непараметрических критериев согласия при проверке сложных гипотез // Автометрия, N 2, 2001. — С. 88–102.

[61] Лотов В. И. Асимптотические разложения в последовательном критерии отношения правдоподобия // Теория вероятностей и ее применения, т. 32, N 1, 1987. — С. 62–72.

[62] Малинкин В.Б., Алгазин Е.И., Левин Д.Н., Попантонопуло В.Н. Инвариантный метод анализа телекоммуникационных систем передачи информации. — Новосибирск: СибГУТИ, 2006.

[63] Мандельброт Б. Теория информации и психолингвистика: теория частот слов // в сб.: Математические методы в социальных науках. — М., 1973. — С. 316–337.

[64] Марков А. А. Об одном применении статистического метода // Известия Академии Наук, Сер. 6, Т. 10, вып. 4, 1916.

[65] Мартынов Г. В. Критерии омега-квадрат. М.: Наука, 1978.

[66] Марусенко М. А. Атрибуция анонимных и псевдонимных литературных произведений методами распознавания образов — Л.: Филол. ф-т СПбГУ, 1990.

[67] Марусенко М. А., Мельникова Е. Е., Родионова Е. С. Атрибуция анонимных и псевдонимных статей, опубликованных в журналах «Время» и «Эпоха» в 1861–1865 гг // В сб.: «Квантитативная лингвистика: исследования и модели (КЛИМ - 2005)», материалы Всероссийской научной конференции. Новосибирск, НГПУ. 2005. — С. 283–293.

[68] Медведев Р. Кто написал «Тихий Дон»?. — Paris: Christian Bourg. Edit., 1975.

[69] Молчан Г. М. О максимуме дробного броуновского движения // Теория вероятностей и ее применения, Т. 44, N 1, 1999. — С. 111–115.

[70] Морозов Н.А. Лингвистические спектры: средство для отличения плагиатов от истинных произведений того или иного известного автора. Стилеметрический этюд // Известия отд. русского языка и словесности Имп.акад.наук, Т.20, Кн.4, 1915.

[71] Надточий Е. Д. Статистическая диагностика авторских различий в синтаксисе: автореф. ... канд. филол. наук — Л., 1983.

[72] Никитин Я. Ю. Асимптотическая эффективность непараметрических критериев. — М.: «Физматлит», 1995.

[73] Никифоров И. В. Последовательное обнаружение изменения свойств временных рядов. — М.: «Наука», 1983.

[74] Питербарг В. И. Асимптотические методы в теории гауссовских случайных процессов и полей. Ч– М.: Изд-во МГУ, 1988.

[75] Питмен Э. Основы теории статистических выводов. — М.: «Мир», 1986.

[76] Поддубный В.В., Поликарпов А.А. Диссипативная стохастическая динамическая модель развития языковых знаков //Компьютерные исследования и моделирование. 2011. Т. 3, N 2. — С. 103–124.

[77] Поддубный В. В., Шевелев О. Г. Сравнение и кластерный анализ текстов по частотным признакам на основе гипергеометрического критерия // В сб.: «Квантитативная лингвистика: исследования и модели (КЛИМ - 2005)», материалы Всероссийской научной конференции. Новосибирск, НГПУ. 2005. — С. 205–217.

[78] Поддубный В.В., Шевелев О.Г. Сравнительный анализ стилей текстов по частотным признакам на основе гипергеометрического критерия // Информационные технологии и математическое моделирование: Материалы III Всероссийской научно-практической конференции (11-12 декабря 2004 г.) Ч. 2. — Томск: Изд-во Том. ун-та, 2004. — С.48–51.

[79] Рыжиков Ю. Вычислительные методы. — СПб., 2007.

[80] Селезнева Л.В. К вопросу моделирования обобщенного броуновского движения // Наука. Техника. Инновации. Матер. регион. конф. молодых ученых, ч.1, 2002. — С. 215–216.

[81] Семянникова Н. В. Верификация некоторых методов атрибуции на переводных текстах // В сб.: «Квантитативная лингвистика: исследования и модели (КЛИМ - 2005)», материалы Всероссийской научной конференции. Новосибирск, НГПУ, 2005. — С. 294–302.

[82] Скороход А. В. Элементы теории вероятностей и случайных процессов. — Киев: Вища школа, 1980.

[83] Соколова Д. О., Спектор А. А. Непараметрическое обнаружение стохастических сигналов, основанное на пересечениях с «нулем»// Вестн. Том. гос. ун-та. Управление, вычислительная техника и информатика.— 2013. — N 1 (22). — С.138–146.

[84] Торговицкий И. Ш. Методы определения момента изменения вероятностных характеристик случайных величин // Зарубежная радиоэлектроника. — 1976. — N 1. — С. 3–52.

[85] Феллер В. Введение в теорию вероятностей и ее приложения. Т. 2. — М.: Мир, 1967.

[86] Феддер. Фракталы. — М., 1983.

[87] Федунец Н.И., Черников Ю.Г. Методы оптимизации. — М., 2009.

[88] Феллер В. Введение в теорию вероятностей и ее приложения. Том 1. — Москва, "Мир 1967.

[89] Фоменко В.П., Фоменко Т.Г. Авторский инвариант русских литературных текстов // Методы количественного анализа текстов нарративных источников. М.: Ин-т истории СССР, 1983. — С. 86–109.

[90] Хмелев Д.В. Распознавание автора текста с использованием цепей А.А. Маркова // Вестн. МГУ. Сер. 9, Филология. 2000. N 2. — С.115–126.

[91] Хмелев Д.В. Как определить писателя? // Компьютерра, N 9, 2000.

[92] Хьетсо Г., Густавссон С., Бекман Б., Гил С. Кто написал «Тихий Дон». (Проблема авторства «Тихого Дона»). — М.: Книга, 1989.

[93] Цветаева М. И. Избранное. Лирика. — Ростов-на-Дону, "Феникс 1995.

[94] Чебунин М. Г. Оценивание параметров вероятностных моделей по числу различных элементов выборки // Сиб. журн. индустр. матем. – 2014. – Т. 17. – є 3. – С. 135-147.

[95] Чебунин М. Г. Оценивание числа ячеек по числу занятых ячеек при случайном размещении // Вестн. НГУ. Сер. матем., мех., информ. – 2014. – Т. 14. – є 3. – С. 107-113.

[96] Шахраманьян А.М. Системы мониторинга и прогноза технического состояния зданий и сооружений. Теория и практика // Русский инженер, N 1 (28), 2011. –Ц С. 54-Ц64.

[97] Шевелев О.Г. Анализ частоты встречаемости различных длин предложений в литературном тексте как возможной характеристики авторского стиля с помощью самоорганизующихся карт Кохонена // Нейроинформатика и ее приложения: Материалы XII Всероссийского семинара, 1–3 октября 2004 г. — Красноярск: ИВМ СО РАН, 2004. — С.177–178.

[98] Ширяев А. Н. Основы стохастической финансовой математики. Т. 1. Факты. Модели. — М., 1998.

[99] Шрейдер Ю. А. О возможности теоретического вывода статистических закономерностей текста (к обоснованию закона Ципфа). Проблемы передачи информации, Т.3, N 1, 1967. — С. 57–63.

[100] Шрейдер Ю. А., Шаров Н. А. Системы и модели. — М., 1982.

[101] Шубик С. А. Размер предложения в немецкой художественной прозе // Сборник статей по методике преподавания иностранных языков и филологии — Вып. 4. Л., 1969. — С. 77–79.

[102] Щербаков М. Песни. Тексты. Фотографии. CD. — М., "Triada 2000.

[103] Яглом А. М. Корреляционная теория стационарных случайных функций с примерами из метеорологии. — Л.: Гидрометеоиздат, 1981.

[104] Adler R. J. An introduction to continuity, extrema, and related topics for general Gaussian processes. Inst. Math. Statist. Lecture Notes — Monograph Series, Vol. 12, Inst. Math. Statist. — Hayward, CA, 1990.

[105] D'Agostino R.,Pearson E. S. Tests for departure from normality. Empirical results for the distributions of b^2 and $\sqrt{b^1}$ // Biometrika, Vol. 60, No. 3, 1973. — P. 613–622.

[106] Azais J. M. Conditions for convergence of number of crossings to the local time. Application to stable processes with independent increments and to Gaussian processes // Probability and mathematical statistics, Vol. 11, 1990. — P.19–36.

[107] Bahadur R. R. On the number of distinct values in a large sample from an infinite discrete distribution // Proceedings of the

National Institute of Sciences of India. – 1960. – V. 26A. – є 2. – P. 67-75.

[108] Barbour A. D. Univariate approximations in the infinite occupancy scheme // Alea. – 2009. – V. 6. – P. 415-433.

[109] Barbour A. D., Gnedin A. V. Small counts in the infinite occupancy scheme // Electronic Journal of Probability. – 2009. – V. 14. – є 13. – P. 365-384.

[110] Bender C. An Ito formula for generalized functionals of a fractional Brownian motion with arbitrary Hurst parameter // Stochastic Processes and Their Applications, V. 104, 2003. — P. 81–106.

[111] Bhattacharyya G. K., Johnson R. A. Nonparametric tests for shift at an unknown time point // Ann. Math. Statist, V. 39, 1968. — P. 1731–1743.

[112] Bischoff W. A functional central limit theorem for regression models // Ann. Stat., Vol. 26, N 4, 1998. — P. 1398–1410.

[113] Blok H. J. On the nature of the stock market: simulations and experiments. — The University of British Columbia, 2000.

[114] Brodsky E., Darkhovsky B.S. Nonparametric Methods in Change Point Problems. — Springer, 1993.

[115] Carlstein E. Nonparametric change-point estimation // Ann. of Statist., V. 16, No. 1, 1988. — P. 188–197.

[116] Carmona P., Coutin L. Fractional Brownian motion and the Markov property // Elect. Comm. in Probab., V. 3, 1998. — P. 95–107.

[117] Chen L., Shapiro S. S. An Alternative Test for Normality Based on Normalized Spacings // J. Statistical Computation and Simulation, Vol.53, No.3–4, 1995. — P. 269–288.

[118] Craigmile P. F. Simulating a class of stationary Gaussian processes using the Davies—Harte algorithm, with applications to long memory processes // J. Time Series Analys., Vol. 24, 2003. — P. 505–511.

[119] Dahlhaus R. Efficient parameter estimation for self-similar processes // The Annals of Statistics, Vol. 17, No. 4, 1989. — P. 1749–1766.

[120] Deitrich C. R., Newsam G. N. Fast and exact simulation of stationary Gaussian processes through circulant embedding of the covariance matrix // SIAM J. Sci. Comput., Vol. 18, 1997. — P. 1088–1107.

[121] Dümbgen L. The asymptotic behavior of some nonparametric change-point estimators // The Annals of Statist., V. 19, No. 3, 1991. — P. 1471–1495.

[122] Durieu O., Wang Y. From infinite urn schemes to decompositions of self-similar Gaussian processes // Preprint. – 2015.

[123] Dutko M. Central limit theorems for infinite urn models // Ann. Probab. – 1989. – V. 17. – P. 1255-1263.

[124] Feuerverger A., Hall P., Wood T. A. Estimation of fractal index and fractal dimension of a Gaussian process by counting the number of level crossings // Journal of time series analysis, Vol. 15, No. 6, 1994. — P. 587–606.

[125] *Gnedin A., Hansen B., Pitman J.* Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws // Probability Surveys. – 2007. – V. 4. – P. 146–171.

[126] Henry M., Zaffarony P. The long-range dependence paradigm for macroeconomics and finance // In: Theory and Applications of Long-Range Dependence (P. Doukhan, G. Oppenheim and M. S. Taqqu, eds.). Birkhauser, Boston, 2002. — P. 417–438.

[127] Herdan E. Calculus of legomena. — N.- Y., 1964.

[128] Hurst H. E., Black R. P., Sinaika Y. M. Long term storage in reservoirs. An experimental study. — London: Constable, 1965.

[129] Hwang H.-K., Janson S. Local Limit Theorems for Finite and Infinite Urn Models // The Annals of Probability. – 2008. – V. 36. – є 3. – P. 992–1022.

[130] Jennane R., Harba R., Jacquet G. Analysis methods for fractional brownian motion: theory and comparative results // Traitement du Signal. Vol. 13, No. 4, 1996. — P. 289-Ц302.

[131] Kallianpur J., Oodaira H. Freidlin—Wentzell type estimates for abstract Wiener spaces // Sankhy$\overline{a}$, Ser. A, Vol. 40, 1978. — P. 116–137.

[132] *Karlin S.* Central Limit Theorems for Certain Infinite Urn Schemes // Jounal of Mathematics and Mechanics. – 1967. – V. 17. – є 4. – P. 373-401.

[133] Key E. S. Rare Numbers // Journal of Theoretical Probability, Vol. 5, No. 2, 1992. — P. 375–389.

[134] Konstantopoulos T., Sakhanenko A. Convergence and convergence rate to fractional Brownian motion for weighted random sums. Preprint. — Iztapalapa, 2000.

[135] Leland W. E., Taqqu M. S., Willinger W., Wilson D. V. On the self-similar nature of Ethernet traffic (extended version) // IEEE/ACM Trans. Networking, Vol. 2, 1994. — P.1–15.

[136] Lu Q.Q. Linear regression under multiple changepoints. –Ц Athens, Georgia, 2004.

[137] Lévi P. Processus stochasticques et movement Brownian. — Paris: Gauthier-Villars, 1948.

[138] Mackay A. L. The tetrahedron in curved space — a problem // Hyperspace, Vol. 4, No. 1, 2000. — P. 19–22.

[139] MacNeill I.B. Limit processes for sequences of partial sums of regression residuals // Annals of probability, Vol. 6, N 4, 1978. –Ц P. 695-Ц698.

[140] Mandelbrot B. B. Long-run linearity, locally Gaussian processes, H-spectra and infinite variances // International Economic Review, V. 10, 1969. — P. 82–113.

[141] Mandelbrot B. B. Fractals: Form, Chance, and Dimencion. — San Francisco: Freeman, 1977.

[142] Mandelbrot B. B. The Fractal Geometry of Nature — San Francisco: Freeman, 1982.

[143] Mandelbrot B. B. Comment on Computer Rendering of Fractal Stochastic Models // Communications of the ACM, V. 25, N 8, 1982. — P. 581–583.

[144] Mandelbrot B., Van Ness J. Fractional Brownian motions, fractional noise and applications // SIAM Review, Vol. 10, N 4, 1968. — P. 422–437.

[145] Mandelbrot B. B., Wallis J. R. Some long-run properties of geophysical records // Water Resources Research, Vol. 5, 1969. — P. 321–340.

[146] Mickoch T., Resnick S., Rootzen H., Stegeman A. Is network traffic approximated by stable Levy motion or fractional Brownian motion? // Ann. Appl. Prob., V. 12, N 1, 2002. — P. 23–68.

[147] Mikhailov V. G. Asymptotic Normality of the Number of Empty Cells for Group Allocation of Particles // Theory Probab. Appl. – 1980. – V. 25. – є 1. – P. 82–90.

[148] Mirakhmedov S. M. Asymptotic normality associated with generalized occupancy problems // Statistics and Probability Letters. – 2007. – V. 77. – є 15. – P. 1549–1558.

[149] Nolan J. P. An algorithm for evaluating stable densities in Zolotarev's (M) parameterization // Math. Comput. Model., Vol. 29, 1999. — P. 229–233.

[150] Norros I. On the use of fractional Brownian motion in the theory of connectionless networks // IEEE J. Select. Areas Commun., V. 13, 1995. — P. 953–962.

[151] Norros I., Valkeila E., Virtamo J. An elementary approach to a Girsanov formula and other analytical results on fractional Brownian motions // Bernoulli, Vol. 5, No. 4, 1999. — P. 571–587.

[152] Samorodnitsky G., Taqqu M. S. Stable Non-Gaussian Processes: Stochastic Models with Infinite Variance. — New York, London, 1994.

[153] Saupe D. Algorithms for random fractals. In: The Science of Fractal Images (Peitgen H.-O., Saupe D., editors). Springer-Verlag, New York, 1988. — P. 71–113.

[154] Peitgen H.-O., Jurgens H., Saupe D. Fractals for the Classroom, Parts 1–2, Introduction to Fractals and Chaos. — New York: Springer-Verlag, 1992.

[155] Peitgen H.-O., Jurgens H., Saupe D. Chaos and Fractals: New Frontiers of Science. — New York: Springer-Verlag, 1992.

[156] Shaban S. A. Change-point problem and two-phase regression: An annotated bibliograthy // Internat. Statist. Rev., V. 48, 1980. — P. 83–93.

[157] Shapiro S. S., Wilk M. B. An Analysis of Variance Test for Normality (Complete Samples) // Biometrika, Vol. 52, No. 3/4, 1965. — P. 591–611.

[158] Sherman L. A. Analitics of literature. A manual for the objective study of English prose and poetry. — Boston, 1893.

[159] Sperry P. Short course in spherical trigonometry. — Richmond, USA, Johnson Publishing Company, 1928.

[160] Voss R. F. Random Fractals: Characterization and Measurement, Scaling Phenomena in Disodered Systems. — New York: Plenum Press, 1985.

[161] Wieand H. S. A condition under which the Pitman and Bahadur approaches to efficiency coincide // The Annals of Statist., V.4, No. 5, 1976. — P. 1003–1011.

[162] Wiener N. Differential space // Journal Math. and Phys. Vfssachusettt's Inst. of Technology, V. 2, 1923. — P. 131–174.

Публикации по теме диссертации

Индексируемые в базах цитирования

[163] Закревская Н. С., Ковалевский А. П. Однопараметрические вероятностные модели статистик текста // Сибирский журнал индустриальной математики. 2001. — Т. IV, N 2 (8). — С. 142–153. RSCI (ядро РИНЦ).

[164] Гусарова Г. В., Ковалевский А. П., Макаренко А. Г. Критерии наличия разладки // Сибирский журнал индустриальной математики. 2005. — Т. VIII, No. 4 (24). — С. 18–33. RSCI (ядро РИНЦ).

[165] Ковалевский А. П., Топчий В. А., Фосс С. Г. О стабильности системы обслуживания с континуально ветвящимися жидкостными пределами // Проблемы передачи информации. 2005. — Т 41, вып. 3. — С. 76–104. RSCI (ядро РИНЦ).

Kovalevskii A. P., Topchii V. A., Foss S. G. On the Stability of a Queueing System with Uncountably Branching Fluid Limits // Problems of Information Transmission. 2005. — V. 41, Issue 3. — P. 254–279. SCOPUS.

DOI 10.1007/s11122-005-0030-6

[166] Ковалевский А. П. Модифицированный знаковый метод тестирования фрактальности гауссовского шума // Проблемы передачи информации, 2008. — Т. 44, вып. 1. — С. 45–58. RSCI (ядро РИНЦ).

Kovalevskii A. P. Modified Sign Method for Testing the Fractality of Gaussian Noise // Problems of Information Transmission, 2008. — Vol. 44, No. 1. — P. 40–52. SCOPUS.

DOI 10.1134/S0032946008010043

[167] Ковалевский А.П., Костин В.С., Хиценко В.Е. Моделирование и идентификация последовательности зависимых случайных величин с симметричным устойчивым распределением // Сибирский журнал индустриальной математики, Том 13, 2010. — N 4 (44). — С. 25–37. RSCI (ядро РИНЦ).

[168] Аркашов Н.С., Ковалевский А.П. Вероятностная модель цен на квартиры // Сибирский журнал индустриальной математики, Том 15, 2012. — N 2 (50). — С. 11–20. RSCI (ядро РИНЦ).

[169] Ковалевский А. П., Шаталин Е.В. Асимптотика сумм остатков однопараметрической линейной регрессии, построенной по порядковым статистикам // Теория вероятностей и ее применения, Т. 59, N 3. — 2014. — С. 452–467. RSCI (ядро РИНЦ).

DOI 10.4213/tvp4579

Kovalevskii A. P., Shatalin E. V. Asymptotics of Sums of Residuals of One-Parameter Linear Regression on Order Statistics // Theory of Probability and Its Applications, Vol. 59, No. 3. — 2015. — P. 375–387. WoS, SCOPUS.

DOI 10.1137/S0040585X97T987193

[170] Kovalevskii A. P., Shatalin E. V. A limit process for a sequence of partial sums of residuals of a simple regression on order statistics with Markov-modulated noise // Probability and Mathematical Statistics, Vol. 36.1. — 2016. — С. 113–120. SCOPUS.

[171] Chebunin M., Kovalevskii A. Functional central limit theorems for certain statistics in an infinite urn scheme // Statistics and Probability Letters, V. 119. — 2016. — С. 344-Ц348. WoS, SCOPUS.

DOI 10.1016/j.spl.2016.08.019

[172] Philonenko P., Postovalov S., Kovalevskii A. The limit test statistic distribution of the maximum value test for right-censored data // Journal of Statistical Computation and Simulation. 2016. — Vol. 86, iss. 17. — P. 3482–3494. WoS, SCOPUS.

DOI 10.1080/00949655.2016.1164703

[173] Ковалевский А. П. Тестирование нормальности очень малых выборок // Сибирские электронные математические известия, Т. 14. — 2017. — С. 1207–1214. WoS, SCOPUS.

DOI 10.17377/semi.2017.14.102

Другие статьи

[174] Филиппова Т.А., Ковалевский А.П., Русина Н.О. Основные вопросы маркетинга и менеджмента в энергетике // Научный вестник НГТУ, 1995. — N 1. — С. 161–169.

[175] Кувшинова М. А., Ковалевский А. П., Асланова И. В. Моделирование показателей энергопроизводства в системе поддержки принятия управленческих решений // Сборник научных трудов НГТУ, 2000. — No. 4 (21). — С. 133–138.

[176] Kovalevskii A. Dependence of increments in time series via large deviations // Proceedings of the 7th Korea-Russia International Symposium on Science and Technology. Ulsan, Korea, 2003. — P. 262–267.

[177] Закревская Н.С., Ковалевский А.П., Селезнева Л.В. Процесс Сопа // Научный вестник НГТУ. 2004. — N 3. — С. 13–19.

[178] Закревская Н.С., Ковалевский А.П. Алгоритм идентификации фрактального броуновского движения по разности оценок // Сборник научных трудов НГТУ, 2004. — N 2 (36). — С. 29–36.

[179] Ковалевский А. П. Применение принципа инвариантности к анализу однородности текста // В сб.: «Квантитативная лингвистика: исследования и модели (КЛИМ - 2005)», материалы Всероссийской научной конференции. Новосибирск, НГПУ. 2005. — С. 195–204.

[180] Обухова О.О., Трунов А.Н., Ковалевский А.П. и др. Динамика продукции интерферона у больных герпетической инфекцией на фоне иммунокоррекции // Вестник новых медицинских технологий, 2008. — Т. XV, N 2. С. 141–143.

[181] Алгазин Е.И., Ковалевский А.П., Малинкин В.Б. Оценка помехоустойчивости инвариантной системы связи при когерентном приеме // Электросвязь, 2009. – N 8. — С. 48–50.

[182] Алгазин Е.И., Ковалевский А.П., Левин Д.Н. Оценка помехоустойчивости системы обработки информации, инвариантной к мультипликативной помехе // Радиотехника, 2009. — N 6. — С. 28–31.

[183] Алгазин Е.И., Ковалевский А.П., Малинкин В.Б. Передача сигналов инвариантным методом при наличии аддитивной стационарной гауссовской помехи с корреляционной функцией общего вида // Вестник СибГАУ, вып. 1 (22), 2009. — С. 32–35.

[184] Алгазин Е.И., Касаткина Е.Г., Ковалевский А.П., Малинкин В.Б. Помехоустойчивость инвариантной системы передачи информации, основанной на когерентном приеме и при наличии слабых корреляционных связей // Вестник СибГАУ, вып. 2 (23), 2009. — С. 55–58.

[185] Алгазин Е.И., Ковалевский А.П., Касаткина Е.Г., Малинкин В.Б. Инвариантная когерентная система при комплексном воз-

действии помех // Вестник Тамбовского государственного технического университета, Т. 15, No. 2, 2009. — С. 295–302.

[186] Алгазин Е.И., Ковалевский А.П., Касаткина Е.Г., Малинкин В.Б. Инвариантная система при наличии аддитивной стационарной гауссовской помехи с корреляционной функцией общего вида и собственных шумов генераторного оборудования // Омский научный вестник, серия «Приборы, машины и технологии», є 2 (80), 2009. — С. 223–227.

[187] Алгазин Е.И., Ковалевский А.П., Малинкин В.Б. Передача сигналов инвариантным методом с последующей нелинейной обработкой // Вестник СибГАУ, вып. 3 (24), 2009. — С. 20–23.

[188] Алгазин Е.И., Ковалевский А.П., Малинкин В.Б. Инвариантная система при нелинейной обработке сигналов // Омский научный вестник, серия «Приборы, машины и технологии», N 3 (83), 2009. — С. 272–275.

[189] Алгазин Е.И., Ковалевский А.П., Малинкин В.Б. Передача сигналов инвариантным методом с последующей нелинейной обработкой при наличии слабой корреляции // Вестник СибГАУ, вып. 4 (25), 2009. — С. 96–98.

[190] Алгазин Е.И., Ковалевский А.П., Малинкин В.Б. Инвариантная система при нелинейной обработке сигналов и наличии слабой корреляции // Омский научный вестник, серия «Приборы, машины и технологии», N 1 (87), 2010. — С. 202–205.

[191] Алгазин Е. И., Ковалевский А. П., Малинкин В. Б. Способы повышения помехоустойчивости системы обработки информации, инвариантной к мультипликативной помехе // Радиотехника, 2010. — N 1. — С. 44–47.

[192] Алгазин Е.И., Ковалевский А.П., Малинкин В.Б. Вопросы реализации оптимальной инвариантной системы передачи информации // Материалы X международной конференции «Актуальные проблемы электронного приборостроения», Том 4, 2010. — С. 123–125.

[193] Алгазин Е. И., Ковалевский А. П. Помехоустойчивость инвариантной системы при нелинейной обработке сигналов // В сб.: Современные проблемы радиоэлектроники. Красноярск, СФУ, 2011. — С. 505–509.

[194] A Posterior Change-Point Analysis in Application to the Dynamics of Enteric Infections and Water Turbidity in Ural Region of Russia. Kovalevsky A., Gubarev V., Loktev V. et al. // In: 22nd Annual Conference of The International Environmetrics Society, Book of Abstracts, Hyderabad, India, January 1–6, 2012. — P. 74.

[195] Kovalevskii A. A regression model for prices of second-hand cars // Applied methods of statistical analysis. Applications in survival analysis, reliability and quality control. Novosibirsk, 2013. — P. 124–128.

[196] Ковалевский А. П. Сравнение статистических критериев разладки модели с циклическим трендом // Обозрение прикладной и промышленной математики, Т. 20, вып. 4, 2013. — С. 552–553.

[197] Шаталин Е. В., Ковалевский А. П. Асимптотика эмпирического моста в линейных регрессионных моделях, построенных по порядковым статистикам // Обозрение прикладной и промышленной математики, Т. 20, вып. 4, 2013. — С. 573–574.

[198] Шаталин Е.В., Ковалевский А.П. Асимптотика эмпирического моста в линейных регрессионных моделях, построенных по порядковым статистикам // Материалы XIV всероссийского симпозиума по прикладной и промышленной математике (осенняя сессия), Великий Новгород, — 2013. — С. 573–574.

[199] Ковалевский А.П. Статистические критерии обнаружения разладки регрессии с циклическим трендом // Научный вестник НГТУ. — 2013. — N 3 (52). — С. 55-62.

[200] Kovalevskiy A., Shatalin E. Limit processes for sequences of partial sums of residuals of regressions against order statistics with Markov-modulated noise // Conference program and abstract book of 11th International conference on ordered statistical data, Bedlewo(Poland). — 2014. — P. 37–38.

[201] Ковалевский А. П., Шахраманьян А. М. Анализ дефектов строительных конструкций методом эмпирического моста // Научный вестник НГТУ. — 2014. — N 3 (56). — С. 171–180.

[202] Ковалевский А. П., Шаталин Е.В. Выбор регрессионной модели зависимости массы тела от роста с помощью эмпирического моста // Вестник Томского государственного университета. Математика и механика, No.5(37). — 2015. — С. 35–47. (РИНЦ).

[203] Ковалевский А. П. Оценивание параметра закона Ципфа-Мандельброта по последовательности количеств разных элементов выборки // Обозрение прикладной и промышленной математики, 2017. — Т. 24, вып. 4. — С. 348–349.

[204] Kovalevskii A. P. Asymptotics of an empirical bridge of a regression on concomitants // 13 International conference on

ordered statistical data (OSD 2018): conference program and abstract book, Spain, Cadiz, 22–25 May 2018. — 2018. — P. 29–30.

Монография

[205] Малинкин В.Б., Алгазин Е.И., Ковалевский А.П. Инвариантные системы связи. — Красноярск, 2010. — 202 с.

Авторские свидетельства

[206] Алгазин Е. И., Ковалевский А. П., Малинкин В. Б. Инвариантная система передачи информации по каналам с переменными параметрами. Патент на полезную модель N 85280. Зарегистрировано в Государственном реестре полезных моделей Российской Федерации 27 июля 2009 г.

[207] Свидетельство на программу для ЭВМ 2012660948. Российская Федерация. Программа расчета вероятности ошибок инвариантной к мультипликативной помехе системы, основанной на использовании линейного детектора / Алгазин Е. И., Ковалевский А. П., Малинкин А. В.; правообладатель Новосиб. гос. техн. ун-т. — 2012618953; заявл. 19.10.12; зарегистрировано 30.10.12. — 1 с. — Тип ЭВМ: IBM PC — совместимый с ПК; язык: FORTRAN; ОС: Microsoft Windows 9X/NT/2000/2003/XP; объем: 0,4 Мб.

[208] Свидетельство на программу для ЭВМ 2012660949. Российская Федерация. Программа расчета вероятности ошибок инвариантной к мультипликативной помехе системы, основанной на использовании поднесущей / Алгазин Е. И., Ковалевский А. П., Малинкин А. В.; правообладатель Новосиб. гос. техн. ун-т. — 2012618954; заявл. 19.10.12; зарегистрировано 30.10.12. — 1 с. — Тип ЭВМ: IBM PC — совместимый с ПК; язык: FORTRAN; ОС: Microsoft Windows 9X/NT/2000/2003/XP; объем: 0,4 Мб.

[209] Свидетельство на программу для ЭВМ 2012660950. Российская Федерация. Программа расчета вероятности ошибок инвариантной к мультипликативной помехе системы, основанной на использовании синхронного детектора / Алгазин Е. И., Ковалевский А. П., Малинкин А. В.; правообладатель Новосиб. гос. техн. ун-т. — 2012618955; заявл. 19.10.12; зарегистрировано 30.10.12. — 1 с. — Тип ЭВМ: IBM PC — совместимый с ПК; язык: FORTRAN; ОС: Microsoft Windows 9X/NT/2000/2003/XP; объем: 0,4 Мб.

[210] Патент N 2014121189/08. Российская Федерация. МПК H03D3/00. Способ фазовой обработки сигналов / Алгазин Е. И., Ковалевский А. П.; заявитель и патентообладатель Новосиб. гос. техн. ун-т; заявл. 26.05.2014; опубл. 10.06.2016.