Московский государственный университет
имени М. В. Ломоносова

На правах рукописи

ГОРШЕНИН Андрей Константинович

АСИМПТОТИЧЕСКИЕ СВОЙСТВА СТАТИСТИЧЕСКИХ ПРОЦЕДУР АНАЛИЗА СМЕСЕЙ ВЕРОЯТНОСТНЫХ РАСПРЕДЕЛЕНИЙ

Специальность 01.01.05 — теория вероятностей
и математическая статистика

АВТОРЕФЕРАТ

диссертации на соискание ученой степени
кандидата физико–математических наук

Москва – 2011

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

Научный руководитель:	доктор физико-математических наук, профессор В. Ю. Королев
Официальные оппоненты:	доктор физико-математических наук, профессор И. Н. Володин
	доктор физико-математических наук В. П. Будаев
Ведущая организация:	Московский государственный институт радиотехники, электроники и автоматики (технический университет)

Защита диссертации состоится 23 сентября 2011 г. в 11 часов на заседании диссертационного совета Д 501.001.44 в Московском государственном университете имени М.В. Ломоносова по адресу: 119991, ГСП-1, Москва, Ленинские горы, МГУ, 2-й учебный корпус, факультет ВМК, аудитория 685. Желающие присутствовать на заседании диссертационного совета должны сообщить об этом за два дня по тел. 939-30-10 (для оформления заявки на пропуск).

С диссертацией можно ознакомиться в библиотеке факультета ВМК МГУ. С текстом автореферата можно ознакомиться на официальном сайте ВМК МГУ http://cs.msu.ru в разделе «Наука» – «Работа диссертационных советов» – «Д 501.001.44».

Автореферат разослан ____ августа 2011 г.

Ученый секретарь
диссертационного совета
профессор Н. П. Трифонов

Общая характеристика работы

Актуальность:

Во многих ситуациях удобными математическими моделями стохастических хаотических процессов являются подчиненные винеровские процессы, по сути представляющие собой процессы броуновского движения со случайным временем (или со случайными параметрами сноса и диффузии). Математическим обоснованием такого подхода являются, в частности, предельные теоремы для обобщенных процессов Кокса, которые являются в некотором смысле наилучшими моделями нестационарных хаотических случайных блужданий и демонстрируют высокую адекватность при их использовании для описания динамики биржевых цен или характеристик турбулентной плазмы на временны́х микромасштабах. С помощью соответствующих предельных теорем такие модели распространяются на временны́е макромасштабы и трансформируются в упоминавшиеся выше подчиненные винеровские процессы[1]. В рамках таких моделей распределения приращений рассматриваемых процессов в общем случае имеют вид сдвиг-масштабных смесей нормальных законов.

При изучении тонкой стохастической структуры хаотических процессов наибольший интерес представляет скорость изменения процесса (то есть его волатильность). При этом, в отличие от многих стандартных определений термина «волатильность», в данной работе будет использоваться понятие многомерной волатильности, которое основано на возможности аппроксимации произвольной сдвиг-масштабной смеси нормальных законов конечной смесью вида

$$\sum_{i=1}^{k} p_i \Phi\Big(\frac{x-a_i}{\sigma_i}\Big),$$

где $\Phi(x)$ функция распределения стандартного нормального закона, $k \geqslant 1$ – известное натуральное число, $p_i \geqslant 0$, $\sum\limits_{i=1}^{k} p_i = 1$, $a_i \in \mathbb{R}$, $\sigma_i > 0$, $i = 1, \ldots, k$. В рамках такой модели распределений приращений хаотических стохастических процессов волатильность трактуется как дисперсия приращения, которая равна

$$D = \sum_{i=1}^{k} p_i \sigma_i^2 + \sum_{i=1}^{k} p_i (a_i - \overline{a})^2, \quad \overline{a} = \sum_{i=1}^{k} p_i a_i.$$

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

Здесь первое слагаемое, не зависящее от параметров сдвига компонент, описывает диффузионную компоненту волатильности, тогда как второе слагаемое, не зависящее от параметров диффузии компонент, описывает динамическую компоненту волатильности.

С целью анализа стохастической структуры рассматриваемой системы, в рамках которой развивается изучаемый процесс, необходимо осуществить декомпозицию волатильности на динамическую и диффузионную составляющую. В рамках указанной выше модели типа конечной смеси распределений вероятности эта задача сводится к задаче статистического разделения конечных смесей, то есть задаче отыскания статистических оценок параметров смеси. Данная задача является весьма важной при изучении скрытых тенденций на финансовых рынках, при исследовании корреляционной структуры хаотических процессов в физике турбулентной плазмы, при анализе информационных потоков в вычислительных или телекоммуникационных системах.

Для решения задачи статистического разделения смесей используются различные методы, наиболее популярным из которых является EM-алгоритм[2], который представляет собой итеративный метод для нахождения оценок максимального правдоподобия.

Несмотря на свою популярность и относительную эффективность, EM-алгоритм не лишен ряда существенных недостатков. Например, существуют проблемы неустойчивости по отношению к исходным данным (оценки могут радикально измениться при замене всего лишь одного наблюдения в выборке из $200-300$ наблюдений) и неустойчивости по отношению к выбору начального приближения (от этого может зависеть скорость сходимости, причем весьма существенно). К тому же алгоритм работает с заранее заданным числом компонент, которое может не соответствовать реальному распределению выборки.

В силу неустойчивости EM-алгоритма по отношению к исходным данным возникает необходимость использования робастных оценок на шагах EM-алгоритма, то есть оценок, обладающих нечувствительностью к малым отклонениям от предположений. В качестве робастных оценок можно рассмотреть так называемые M-оценки. M-оценка – всякая оценка T_n, определяемая как решение экстремальной задачи на минимум вида

$$\sum_{i=1}^{n} \rho(x_i; T_n) \to \min,$$

[2] *A. Dempster, N. Laird and D. Rubin.* Maximum likelihood estimation from incompleted data // Journal of the Royal Statistical Society, 1977. Series B. Vol. 39(1). P. 1–38.

где $\rho(\cdot)$ – произвольная функция. М-оценки допускают обобщение на многопараметрический случай, что позволяет одновременно выписывать оценки данного типа для сдвига и масштаба. Известно[3], что медиана является робастной М-оценкой параметра сдвига. Более того, медиана является единственной М-оценкой, инвариантной относительно масштаба. Поэтому в данной работе значительное внимание уделяется построению и применению медианных модификаций алгоритмов ЕМ-типа.

Одним из важнейших недостатков классического ЕМ-алгоритма является то, что он в ряде ситуаций выбирает первый попавшийся локальный максимум[4]. То есть, являясь методом локальной оптимизации, он приводит не к глобальному максимуму функции правдоподобия, а к тому локальному максимуму, который является ближайшим к начальному приближению. Довольно эффективный способ преодоления данного недостатка заключается в случайном «встряхивании» наблюдений (выборки) на каждой итерации. Этот способ лежит в основе *SEM-алгоритма*[5] (от Stochastic EM-algorithm, стохастический (или случайный) ЕМ-алгоритм). Отличие заключается в добавлении дополнительного *S-шага*, на котором и реализуется указанное встряхивание «выборки».

Изучение свойств SEM-алгоритма проводилось для случая неполных данных, а также с введение дополнительных ограничений[6,7]. В данной работе основное внимание уделяется изучению применения SEM-алгоритма к задаче разделения конечных смесей вероятностных распределений и, прежде всего, к разделению смесей нормальных законов (в частности, с применением его новой версии – медианного SEM-алгоритма), а также доказательству важных свойств сходимости данного алгоритма для произвольного конечного числа компонент без дополнительных предположений о параметрах метода. Вопросы, относящиеся к данной тематике, ранее либо не исследовались, либо изучались лишь для некоторых частных случаев.

Для классического SEM-алгоритма известны результаты о свойствах сходимости для случая смеси только двух законов, однако

[3] *П. Хьюбер.* Робастность в статистике. М.: Мир, 1984. – 304 с.

[4] *В. Ю. Королев.* Вероятностно-статистический анализ хаотических процессов с помощью смешанных гауссовских моделей. Декомпозиция волатильности финансовых индексов и турбулентной плазмы. – М.: ИПИ РАН, 2007. – 363 с.

[5] *M. Broniatowski, G. Celeux and J. Diebolt.* Reconnaissance de mélanges de densités par un algorithme d'apprentissage probabiliste // Data Analysis and Informatics, 1984. Vol. 3. P. 359–373.

[6] *E. H. Ip.* A Stochastic EM Estimator in the Presence of Missing Data. – Theory and Practice. PhD Dissertation, Stanford University, 1994.

[7] *S. F. Nielsen.* Stochastic EM algorithm: Estimation and asymptotic results // Bernoulli, 2000. № 6. P. 457–489.

приведенная техника доказательства не допускает обобщения даже на случай смеси трех законов[8]. Более того, предлагается рассмотреть дополнительные ограничения, которые фактически предназначены для того, чтобы исключить случай пустых кластеров, а также учесть возможность считать пустым не только кластер, не содержащий элементов выборки, но и содержащий некоторое их число. Очевидным недостатком данного подхода является тот факт, что приходится принудительно задавать число компонент в подгоняемой смеси, которое на практике обычно неизвестно. Способы преодоления указанного недостатка также рассматриваются в диссертации.

Алгоритмы EM-типа могут применяться как важная составная часть некоторой более сложной процедуры, называемой методом скользящего разделения смесей (CPC-методом). Данный метод позволяет учесть изменения в эволюции процесса с течением времени. Такой подход позволяет решить задачу декомпозиции волатильности в динамике, отследить появление и исчезновение факторов, формирующих структуру процесса в каждый момент времени.

Важным параметром в модели типа смесей вероятностных распределений является число компонент. Алгоритмы EM-типа обычно подразумевают явное задание этого числа. При этом включение в модель дополнительных параметров увеличивает ее согласие с данными. Однако в данной ситуации возникают две существенные сложности. Во-первых, увеличение числа параметров приводит к существенному повышению вычислительной сложности алгоритма. Во-вторых, в ряде ситуаций использование максимального числа компонент может не приводить к увеличению согласия. К примеру, для масштабных смесей известен эффект насыщения, когда согласие не увеличивается уже со значений числа компонент, равного $4 - 5$. Для сдвиг-масштабных смесей известен эффект перетекания волатильности, когда при небольшом числе компонент (около $2 - 3$) бóльшее влияние имеет диффузионная компонента, а при увеличении числа компонент – динамическая. Таким образом, задание слишком большого числа компонент может критически влиять на соответствие модели исходным данным или на интерпретацию получаемых результатов. Поэтому задача исследования подходов к определению точного числа компонент является исключительно важной и во многом определяющей для успешного применения подобных моделей и методов на практике.

Многие существующие подходы к определению числа компо-

[8] *G. Celeux, J. Diebolt.* Asymptotic properties of a stochastic EM algorithm for estimating mixing proportions // Communications in statistics. Stochastic models 1993. Vol. 9. P. 599–613.

нент смеси носят название информационных, так как основываются на понятии расстояния Кульбака–Лейблера[9], также называемого энтропией по Кульбаку. В качестве примеров можно привести критерий Акаике[10], байесовский информационный критерий[11], критерий Ло[12]. Первые два критерия позволяют учесть увеличение согласия с данными при увеличении числа параметров, однако они подразумевают использование некоторой штрафной функции за включение в модель новых параметров. Критерий Ло не требует штрафных функций, однако его статистика обладает весьма сложным распределением при выполнении нулевой гипотезы, а именно взвешенным χ^2-распределением. При этом определение параметров данного распределения представляет собой достаточно серьезную вычислительную задачу даже на небольших объемах выборки и малом числе компонент в смеси (например, уже при максимальном числе компонент, равном трем).

Общим недостатком подобных критериев является то, что для корректности их применения требуется выполнение достаточно жестких условий регулярности, которые для реальных ситуаций могут не быть справедливыми. Так, например, для смесей нормальных законов нарушается предположение о конечности функции правдоподобия, поэтому формальное применение данных критериев может приводить к ошибочным результатам.

Чтобы минимизировать возможные ошибки, возникающие из-за необходимости задавать в явном виде точное число компонент алгоритмам EM-типа, в диссертации предложено использовать статистический подход к определению числа компонент по выборке. Исходя из особенностей применения предлагаемых алгоритмов, были выделены две практически значимые модели смесей вероятностных распределений, в которых необходимо правильно оценивать число компонент (названные *моделью добавления компоненты* и *моделью расщепления компоненты*). При этом ключевым моментом является переход от проверки гипотез о значении *натуральнозначного дискретного* параметра (равного числу компонент смеси) к проверке гипотез о значении непрерывного параметра (соответствующего весу компоненты, значимость которой проверяется). При таком

[9] *S. Kullback and R. A. Leibler.* On Information and Sufficiency // Annals of Mathematical Statistics, 1951.Vol. 22. P. 79–86.

[10] *H. Akaike.* Information theory and an extension of the maximum likelihood principle.// In: B. N. Petrov and F. Csake (eds.) Second International Symposium on Information Theory. – Budapest, 1973. P. 267–281.

[11] *G. Schwartz.* Estimating the dimension of a model // The Annals of Statistics, 1978. Vol. 6. P. 461–464.

[12] *Y. Lo, N. R. Mendell and D. B. Rubin.* Testing the number of components in a normal mixture // Biometrika, 2001. Vol. 88. №. 3. P. 767–778.

переходе естественно возникает задача проверки простой гипотезы против сложной альтернативы. Для построения критерия и исследования его свойств при решении данной задачи используется асимптотический подход.

В рамках такого подхода, также называемого подходом Питмэна[13], размер и мощность критерия одновременно отделены от нуля, при этом важную роль играют асимптотический дефект[14] и потеря мощности. При этом предполагается, что распределение статистики и мощность критерия зависят от некоторого неизвестного параметра t, $0 < t \leqslant C$, $C > 0$. Однако величина, определяющая потерю мощности, позволяет сравнить мощность некоторого критерия, не зависящего от неизвестного параметра t, с мощностью наиболее мощного критерия, зависящего от t. Таким образом, можно гарантировать, что, с одной стороны, полученный критерий будет асимптотически наиболее мощным, а с другой стороны, возможно его корректное применение на практике. Величина же дефекта критерия говорит о том, сколько дополнительных наблюдений необходимо для того, чтобы мощность данного критерия совпала с мощностью наиболее мощного критерия. Важную роль в развитии методологии доказательств в данной области сыграли работы Л. ЛеКама[15,16], которые позволили получать выражения для потери мощности без построения асимптотических разложений (см. работы Д. М. Чибисова[17,18]). Наконец, в книге В. Е. Бенинга[19] были получены выражения для асимптотического дефекта и потери мощности, использование которых позволило в данной работе в явном виде получить потерю мощности и асимптотический дефект предложенных асимптотически наиболее мощных критериев.

Цель работы:

Целью данной работы является исследование асимптотических свойств предложенных в диссертации статистических процедур анализа смесей вероятностных распределений, предназначенных

[13] *E. J. G. Pitman.* Lecture notes on nonparametric statistical inference. Lectures given for the University of North Carolina, Institute of Statistics, 1948.

[14] *J. L. Hodges, Jr., and E. L. Lehmann.* Deficiency // Ann. Math. Statist., 1970. Vol. 41. P. 783–801.

[15] *L. LeCam.* An extension of Wald's theory of statistical decision functions // Ann. Math. Statist., 1955. Vol. 26. P. 69–81.

[16] *L. LeCam.* Asymptotic Methods in Statistical Decision Theory. – New York: Springer, 1986.

[17] *D. M. Chibisov.* Asymptotic expansions and deficiencies of tests // In: Proc. Intern. Congr. Math.,Warszawa, 1983. Vol. 2. P. 1063–1079.

[18] *D. M. Chibisov.* Calculation of the deficiency of asymptotically efficient tests // Theory Probab. Appl., 1985. Vol. 30. P. 289–310.

[19] *V. E. Bening.* Asymptotic Theory Of Testing Statistical Hypothesis: Efficient Statistics, Optimality, Power Loss and Deficiency. – Untrecht: VSP, 2000. – 277 p.

для решения задачи статистического определения параметров смесей, в том числе, для определения числа компонент.

Методика исследования:
Для решения задач в первой главе используется методы математического анализа, теории вероятностей, а также аппарат марковских цепей. Вторая глава существенно использует методы математического анализа, свойства метрики Леви, методы статистической проверки гипотез, а также асимптотический подход Питмэна. Результаты данной главы проверялись с помощью компьютерного моделирования и программной реализации построенных критериев на встроенном языке программирования пакета MATLAB. Третья глава базируется на анализе реальных данных с финансовых рынков и экспериментальных измерений параметров турбулентной плазмы. Тестирование этих данных проводилось с помощью изученных в первых двух главах методов, которые были реализованы программно на различных языках программирования (встроенный язык программирования пакета MATLAB, C++).

Научная новизна:
Все основные результаты диссертации являются новыми и состоят в следующем:

1. Получено обоснование возможности использования медианных модификаций алгоритмов EM-типа для смесей нормальных законов.

2. Установлены свойства получаемой на итерационных шагах SEM-алгоритма последовательности оценок параметров идентифицируемых сдвиг-масштабных смесей вероятностных распределений с произвольным конечным числом компонент. Доказано, что последовательность SEM-оценок параметров смеси представляет собой конечную однородную апериодическую эргодическую марковскую цепь. Данный результат означает корректность использования стохастических алгоритмов EM-типа для получения оценок компонент смеси: доказан факт сходимости распределения итерационной последовательности оценок к стационарному распределению, а также установлена независимость от начального приближения. В частности, эти результаты справедливы для конечных сдвиг-масштабных смесей нормальных законов.

3. Доказаны теоремы устойчивости конечных масштабных смесей нормальных законов к возмущениям параметров в терминах расстояния Леви. Получены двусторонние оценки для

расстояний Леви между смесями через расстояние Леви между смешивающими распределениями в рамках моделей добавления и расщепления компоненты. Данный результат может быть использован для обоснования эквивалентности задач проверки гипотез о значении дискретного и непрерывного параметра для статистического определения числа компонент произвольных конечных смесей вероятностных распределений, а также для доказательства корректности использования различных моделей типа конечных смесей нормальных законов, в частности, сеточных методов разделения смеси.

4. Построены асимптотически наиболее мощные критерии проверки гипотез о числе компонент конечной смеси вероятностных распределений и исследованы их асимптотические свойства, в частности, установлена асимптотическая нормальность критериев, выписаны выражения для потери мощности и асимптотического дефекта. Найдены условия их применимости к анализу практически значимых моделей вида конечных сдвиг-масштабных смесей нормальных и гамма-распределений, а также для случая смесей равномерных распределений. Продемонстрирована высокая вычислительная эффективность полученных критериев по сравнению с известными.

5. Рассмотренные в диссертации методы и статистические процедуры эффективно применены к исследованию стохастической структуры конкретных сложных хаотических систем, в частности, плазменной турбулентности.

Практическая значимость: Результаты диссертации имеют теоретический характер. Однако они направлены на повышение эффективности практического применения статистических процедур анализа смешанных вероятностных моделей. Все описанные методы имеют строгие математические обоснования и в тоже время успешно применены к анализу статистических или экспериментальных данных в различных областях, таких как финансовые рынки или физика турбулентной плазмы.

Апробация работы:

Результаты работы неоднократно докладывались и обсуждались на научном семинаре кафедры Математической статистики факультета ВМК МГУ «Теория риска и смежные вопросы» (2008 – 2011 гг.), Международной научной конференции студентов, аспирантов и молодых ученых «Ломоносов» (2008, 2009 гг.),

научной конференции «Тихоновские чтения» (2010 г.), международной научной конференции «Интеллектуальная обработка информации» (2010 г.), XII Всероссийском Симпозиуме по прикладной и промышленной математике (2011 г.), международной научной конференции «Моделирование нелинейных процессов и систем» (2011 г.).

Методы, описанные в диссертации, реализованы программно на различных языках программирования, получены свидетельства о государственной регистрации программ для ЭВМ №№ 2009610873, 2010611909, 2010611910, 2010611911, 2011610584, 2011610587, 20116119047, 20116119048. Результаты диссертации были использованы при проведении анализа экспериментальных исследований стохастических плазменных процессов в стелларатоpе Л-2М и линейной установке ТАУ-1 в Институте общей физики им. А. М. Прохорова Российской Академии Наук.

Публикации:

Материалы диссертации опубликованы в 14 печатных работах ([1] – [14]), из них 5 статей опубликованы в журналах, включенных в перечень ВАК ([2], [3], [8], [9], [13]).

Структура и объем диссертации:

Диссертация состоит из введения, трех глав, разбитых на 10 параграфов, и списка литературы, содержащего 87 наименований. Общий объем работы составляет 175 страниц.

Содержание работы

Первая глава посвящена исследованиям свойств различных итерационных методов оценивания параметров смесей вероятностных распределений.

В §1.1 дано описание медианной модификации EM-алгоритма, а также обосновывается целесообразность использования робастных оценок медианного типа на E-этапе EM-алгоритма в задаче разделения конечных смесей нормальных законов. Показано, что медианные оценки естественным образом возникают на E-этапе EM-алгоритма в задаче разделения конечных смесей двойных экспоненциальных распределений (распределений Лапласа) с теми же самыми значениями параметров сдвига и масштаба компонент, что и у исходной смеси нормальных законов. В свою очередь, двойное экспоненциальное распределение можно представить в виде масштабной смеси нормальных законов при стандартном показательном смешивающем распределении. Таким образом, медианная модификация

EM-алгоритма по сути сводится к замене исходной задачи разделения конечных смесей нормальных законов задачей разделения конечных смесей распределений Лапласа с теми же самыми значениями параметров сдвига и масштаба компонент. При указанной замене исходные данные представляются в виде «зашумленной» выборки, причем «зашумление» производится с помощью умножения параметров масштаба компонент на случайную величину со стандартным показательным распределением, а подлежащие оцениванию параметры положения (сдвига) компонент остаются неизменными. Показано, что оценки, получаемые с помощью медианной версии EM-алгоритма в задаче разделения конечных смесей нормальных законов, приближают оцениваемые параметры постольку, поскольку соответствующая последовательность оценок, получаемая EM-алгоритмом, сходится к оценкам максимального правдоподобия аналогичных параметров в модели вида конечных смесей распределения Лапласа.

В §1.2 дано общее описание SEM-алгоритма, а также приводятся подробные формулы для важного частного случая конечных сдвиг-масштабных смесей нормальных распределений (в частности, рассматривается и медианная версия SEM-алгоритма для смесей нормальных распределений). Описаны свойства последовательности SEM-оценок, которые строятся алгоритмом при решении задачи разделения конечных смесей вероятностных распределений с произвольным числом компонент. Основной результат формулируется в виде следующей теоремы.

Теорема 1.1. *Последовательность оценок* $\left\{\theta^{(m)}\right\}$*, получаемая SEM-алгоритмом в задаче разделения идентифицируемых смесей с произвольным конечным числом компонент, представляет собой конечную однородную апериодическую эргодическую марковскую цепь.*

Данная теорема играет ключевую роль в обосновании корректности использования стохастических EM-алгоритмов для оценивания параметров смесей. Доказательство заключается в последовательной проверке свойств марковской цепи, которой является последовательность SEM-оценок.

Как уже было отмечено, в работах, посвященных исследованию свойств SEM-алгоритма, предполагается выполнение ряда дополнительных условий. Так, в первых работах, посвященных данной тематике, устанавливались свойства лишь для двухкомпонентной смеси (при этом отмечалась невозможность обобщения приведенных доказательств на произвольное число компонент). Затем была доказана сходимость SEM-алгоритма для произвольного числа компонент, установлены асимптотические свойства последователь-

ности SEM-оценок (асимптотическая нормальность) при выполнении достаточно сложных для проверки на практике условий. Более того, некоторые условия для реальных данных вообще могут не выполняться (например, предположение строгой положительности весов компонент смеси может нарушаться в силу того, что этапы SEM-алгоритма непосредственно не запрещают весам обращаться в нуль). Теорема 1.1 устанавливает свойства оценок SEM-алгоритма для произвольного числа компонент без введения дополнительных предположений о параметрах метода.

Вторая глава посвящена построению наиболее мощных критериев проверки гипотез о числе компонент смеси. Для формализации задачи предложены две модели: добавления компоненты и расщепления компоненты. Рассматриваются сдвиг-масштабные смеси произвольных абсолютно непрерывных распределений.

С целью формирования гипотез в задаче статистической проверки гипотез о числе компонент смеси и количественной оценки того, насколько может измениться модель при добавлении или изъятии компоненты, в §2.1 рассматривается задача оценки устойчивости конечных масштабных смесей нормальных законов относительно смешивающего распределения в рамках упомянутых выше двух специальных моделей добавления и расщепления компоненты. Основные результаты данного раздела сформулированы в теоремах $2.1 - 2.4$.

Предположим, что каждое из независимых наблюдений имеет распределение, представимое в виде конечной масштабной смеси нормальных законов вида

$$G(x) = \sum_{i=1}^{k} p_i \Phi(x\sigma_i),\ \sum_{i=1}^{k} p_i = 1,\ p_i \geqslant 0,\ \sigma_i > 0,\ i = \overline{1,k}. \qquad (1)$$

Очевидно, что функция распределения $G(x)$ из соотношения (1) может быть представлена в виде

$$G(x) = \mathbb{E}\Phi(Ux),$$

где U – дискретная случайная величина, принимающая значения σ_i с вероятностями p_i, $i = 1, \ldots, k$. Обозначим через $\rho(F, G)$ равномерное расстояние между функциями распределения $F(x)$ и $G(x)$, а через $L(F, G)$ – соответствующее расстояние Леви.

В модели добавления компоненты предполагается, что каждое из независимых наблюдений имеет распределение, представимое в виде

$$G_p(x) = (1-p)\sum_{i=1}^{k} p_i \Phi(x\sigma_i) + p\Phi(x\sigma),$$

где все величины σ_i, p_i, $i = 1, \ldots, k$, считаем известными, а $\sigma > 0$ и $0 \leqslant p \leqslant 1$ считаем параметрами модели. Без ограничения общности для определенности считаем, что $0 < \sigma \leqslant \sigma_1 \leqslant \sigma_2 \leqslant \ldots \leqslant \sigma_k$.

В модели расщепления компоненты предполагается, что каждое из независимых наблюдений имеет распределение, представимое в виде

$$G_p(x) = \sum_{i=1}^{k-1} p_i \Phi(x\sigma_i) + (p_k - p)\Phi(x\sigma_k) + p\Phi(x\sigma), \qquad (2)$$

где все величины σ_i, p_i, $i = 1, \ldots, k$, считаем известными, а $\sigma > 0$ и $0 \leqslant p \leqslant p_k$ считаем параметрами модели. Без ограничения общности для определенности будем считать, что выполнены соотношения $0 < \sigma_1 \leqslant \sigma_2 \leqslant \ldots \leqslant \sigma_{k-1} \leqslant \sigma \leqslant \sigma_k$.

Отметим, что условие отделенности параметров масштаба от нуля в обеих моделях также является достаточно общим и означает, что рассматриваются невырожденные нормальные законы с конечными дисперсиями.

Для моделей добавления и расщепления компоненты в диссертации доказываются четыре теоремы об устойчивости, связывающие двойными неравенствами расстояния Леви между смесями и смешивающими распределениями (теоремы 2.1 – 2.4). В качестве примера приведем одну из теорем для модели расщепления компоненты. Здесь и далее $\varphi(\cdot)$ обозначает плотность стандартного нормального закона.

Теорема 2.3. *В рамках модели расщепления компоненты* (2) *справедливы неравенства*

$$C_2^{[2]}(\sigma_1, \sigma_k) L(G, G_p) \leqslant L(U, U_p) \leqslant C_1^{[2]}(\sigma_k) L^{1/2}(G, G_p),$$

где коэффициенты $C_j^{[2]}$, $j = 1, 2$, *не зависят от величин* p *и* σ *и имеют вид*

$$C_1^{[2]}(\sigma_k) = \varphi^{-1/2}(\sigma_k) \left(1 + \frac{\sigma_k}{\sqrt{2\pi}}\right)^{1/2},$$

$$C_2^{[2]}(\sigma_1, \sigma_k) = \frac{\sigma_1 \sqrt{2\pi e}}{\max\{1, \sigma_k\}}.$$

Доказанные теоремы позволяют переформулировать задачу проверки гипотез о значении дискретного (натуральнозначного) параметра, равного числу компонент смеси, в терминах задачи проверки гипотез о значении непрерывного параметра, принимающего

значения из отрезка $[0,1]$. Переход от дискретного случаю к непрерывному играет важную роль при построении асимптотически оптимальных критериев проверки гипотез о числе компонент.

В §2.2 строится асимптотически наиболее мощный критерий в рамках модели добавления компоненты и исследуются его свойства.

Пусть k – некоторое известное натуральное число. Требуется проверить гипотезу

$$H_0 : \ K = k$$

против альтернативы

$$H_1 : K = k + 1,$$

где через K обозначено «истинное» число компонент в смеси. Для удобства асимптотического анализа сведем задачу проверки гипотез о значении дискретного параметра K к задаче проверки гипотез о значении непрерывного параметра: рассматривается простая гипотеза вида

$$H_0 : \ \theta = 0$$

против последовательности сложных альтернатив вида

$$H_1 : \ \theta = \frac{t}{\sqrt{n}} > 0,$$

где t – неизвестный параметр.

Модель добавления компоненты в общем случае имеет вид ($\psi_i(x)$ – плотности, $p_i \geqslant 0$, $i = 1, \ldots, k$, $\theta \in [0,1]$, $\sum\limits_{i=1}^{k} p_i = 1$)

$$p(x,\theta) = (1-\theta)\sum_{i=1}^{k} p_i \psi_i(x) + \theta\psi_{k+1}(x) \equiv (1-\theta)f(x) + \theta g(x). \quad (3)$$

Первый из основных результатов данной главы сформулирован в следующей теореме. Здесь и далее u_α обозначает $(1-\alpha)$-квантиль стандартного нормального закона.

Теорема 2.5. *Пусть для $s = 2,3,4$ моментные характеристики $\Psi_s = \mathbb{E}_0 \left(g(X_1)/f(X_1)\right)^s$ для функций $f(x)$ и $g(x)$ из соотношения (3) конечны, а соответствующая смесь идентифицируема. Тогда для модели добавления компоненты критерий проверки гипотезы о том, что смесь является k-компонентной, против альтернативы, что смесь является $(k+1)$-компонентной, основанный на статистике*

$$T_1 = n^{-1/2} \sum_{i=1}^{n} \left(\frac{g(X_i)}{f(X_i)} - 1 \right),$$

обладает следующими свойствами:

1. *При справедливости нулевой гипотезы статистика* T_1 *имеет нормальное распределение с параметрами* 0 *и* $\Psi_2 - 1$ *при* $n \to \infty$*:*

$$\mathfrak{L}(T_1 \mid H_0) \to N(0, \Psi_2 - 1).$$

2. *При справедливости альтернативы статистика* T_1 *имеет нормальное распределение с параметрами* $t\,(\Psi_2 - 1)$ *и* $\Psi_2 - 1$ *при* $n \to \infty$*:*

$$\mathfrak{L}(T_1 \mid H_{n,1}) \to N(t\,(\Psi_2 - 1)\,, \Psi_2 - 1).$$

3. *Данный критерий является асимптотически наиболее мощным критерием для заданного уровня* $\alpha \in (0,1)$ *с предельной мощностью вида*

$$\beta^*(t) = \Phi(t\sqrt{\Psi_2 - 1} - u_\alpha).$$

4. *Потеря мощности этого критерия равна*

$$r(t) = \lim_{n\to\infty} n(\beta^*(t) - \beta_n(t)) = \frac{t^3\,\varphi\left(u_\alpha - t\sqrt{\Psi_2 - 1}\right)}{8\sqrt{\Psi_2 - 1}} \times$$
$$\times\Big(\Psi_4 + 2\Psi_3 - \Psi_2^2 - \Psi_2 - \frac{(\Psi_3 - 1)^2}{\Psi_2 - 1} - 1\Big).$$

5. *Асимптотический дефект этого критерия равен*

$$d = \frac{2r(t)}{t\sqrt{I}\varphi(t\sqrt{I} - u_\alpha)} =$$
$$= \frac{t^2}{4\,(\Psi_2 - 1)}\Big(\Psi_4 + 2\Psi_3 - \Psi_2^2 - \Psi_2 - \frac{(\Psi_3 - 1)^2}{\Psi_2 - 1} - 1\Big).$$

Здесь $\beta_n(t)$ – мощность критерия, основанного на статистике T_1.

Отметим, что условия теоремы обеспечивают выполнение условий регулярности, что устанавливает следующая лемма.

Лемма 2.1. *Пусть фишеровская информация* I *для плотности* $p(x,\theta)$ *для модели добавления компоненты конечна. Тогда выполнены условия регулярности.*

Здесь же получены достаточные условия конечности моментных характеристик Ψ_s, $s = 2, 3, 4$, которые для случая конечных смесей нормальных и гамма-распределений имеют вид

$$\sigma_{k+1}^2 < \frac{4}{3} \cdot \max_{1 \leqslant j \leqslant k} \sigma_j^2$$

для смесей нормальных распределений и

$$\beta_{k+1} \geqslant \max\left\{\frac{1}{4}\min_{1\leqslant i\leqslant k}(3\beta_i+1), \frac{1}{2}\min_{1\leqslant i\leqslant k}(\beta_i+1)\right\},\ \alpha_{k+1} > \frac{3}{4}\min_{1\leqslant j\leqslant k}\alpha_j.$$

для смесей гамма-распределений.

Для корректного рассмотрения примеров доказывается теорема об условиях идентифицируемости смесей равномерных распределений.

Теорема 2.6. *Пусть* $A(M) = \bigcup\limits_{i\in M}[a_i, b_i]$, *где* M – *некоторое подмножество номеров. Обозначим семейство конечных смесей равномерных распределений через*

$$H = \left\{F(x) = \sum_{i=1}^{k} p_i F_i(x),\ \sum_{i=1}^{k} p_i = 1,\ F_i \in \mathfrak{F}\right\},$$

где $\mathfrak{F} = \{F(x, a_i, b_i),\ x \in \mathbb{R},\ -\infty < a_i < b_i < \infty,\ i \in \mathbb{N}\}$ – *некоторое множество функций распределения равномерных законов (возможно, конечное). Семейство* H *идентифицируемо тогда и только тогда, когда*

$$A(M_1) \setminus A(M_2) \neq \varnothing,$$

для всех возможных различных M_1 *и* M_2, $M_i \subseteq \mathbb{N}$.

В §2.3 рассматривается асимптотически наиболее мощный критерий для модели расщепления компоненты, которая в общем случае формализуется следующим образом ($\psi_i(x)$, $\psi(x)$ – плотности, $p_i \geqslant 0$, $i = 1, \ldots, k$, $0 \leqslant \theta \leqslant p_k$, $\sum\limits_{i=1}^{k} p_i = 1$):

$$p(x, \theta) = \sum_{i=1}^{k} p_i\psi_i(x) + \theta\cdot(\psi(x) - \psi_k(x)) = f(x) + \theta\cdot g(x). \qquad (4)$$

Второй основной результат данной главы сформулирован в следующей теореме.

Теорема 2.7. *Пусть выполнены достаточные условия конечности моментных характеристик* $\Psi_s = \mathbb{E}_0\left(g(X_1)/f(X_1)\right)^s$, $s = 2, 3, 4$, *для функций* $f(x)$ *и* $g(x)$ *из соотношения* (4), *а соответствующая смесь идентифицируема. Тогда для модели расщепления компоненты критерий проверки гипотезы о том, что смесь является* k*-компонентной, против альтернативы, что смесь является* $(k+1)$*-компонентной, основанный на статистике*

$$T_2 = n^{-1/2}\sum_{i=1}^{n}\frac{g(X_i)}{f(X_i)},$$

обладает следующими свойствами:

1. *При справедливости нулевой гипотезы эта статистика имеет нормальное распределение с параметрами* 0 *и* Ψ_2 *при* $n \to \infty$*:*
$$\mathfrak{L}(T_2 \mid H_0) \to N(0, \Psi_2).$$

2. *При справедливости альтернативы эта статистика имеет нормальное распределение с параметрами* $t\Psi_2$ *и* Ψ_2 *при* $n \to \infty$*:*
$$\mathfrak{L}(T_2 \mid H_{n,1}) \to N(t\Psi_2, \Psi_2).$$

3. *Данный критерий является асимптотически наиболее мощным критерием для заданного уровня* $\alpha \in (0,1)$ *с предельной мощностью вида*
$$\beta^*(t) = \Phi(t\sqrt{\Psi_2} - u_\alpha).$$

4. *Потеря мощности для этого критерия составляет*
$$r(t) = \frac{t^3}{8\sqrt{\Psi_2}} \cdot \varphi\left(u_\alpha - t\sqrt{\Psi_2}\right)\left(\Psi_4 - \Psi_2^2 - \frac{\Psi_3^2}{\Psi_2}\right).$$

5. *Асимптотический дефект для этого критерия равен*
$$d = \frac{t^2}{4\Psi_2}\left(\Psi_4 - \Psi_2^2 - \frac{\Psi_3^2}{\Psi_2}\right).$$

В данной модели выполнение условий регулярности устанавливается следующей леммой.

Лемма 2.2 *Пусть при* $k = 1$ *в равенстве* (4) *конечен интеграл*
$$\int_{-\infty}^{\infty} \psi^2(x)\psi_1^{-1}(x)\,dx,$$
а при $k \geqslant 2$ *конечен интеграл*
$$\int_{-\infty}^{\infty} g^2(x)\left(\sum_{i=1}^{k-1} p_i\psi_i(x)\right)^{-1} dx.$$
Тогда выполнены условия регулярности.

В этом параграфе также получены достаточные условия конечности моментных характеристик Ψ_s, $s = 2, 3, 4$, которые для случая конечных смесей нормальных и гамма-распределений имеют вид

$$\sigma^2 < \frac{4}{3}\sigma_k^2,\ k \geqslant 1, \quad \sigma^2 < 2 \max_{1 \leqslant j \leqslant k-1} \sigma_j^2,\ \sigma_k^2 < 2 \max_{1 \leqslant j \leqslant k-1} \sigma_j^2,\ k \geqslant 2.$$

для смесей нормальных распределений и

$$\beta \geqslant \max\left\{\frac{1}{4}(3\beta_1 + 1), \frac{1}{2}(\beta_1 + 1)\right\},\ \alpha > \frac{3}{4}\alpha_1, \quad k = 1,$$

$$\beta \geqslant \max\left\{\frac{1}{4}(3\beta_k + 1), \frac{1}{2}(\beta_k + 1), \frac{1}{2} \min_{1 \leqslant i \leqslant k-1} (\beta_i + 1)\right\},$$

$$\alpha > \max\left\{\frac{1}{2} \min_{1 \leqslant j \leqslant k-1} \alpha_j, \frac{3}{4}\alpha_k\right\},$$

$$\beta_k \geqslant \frac{1}{2} \min_{1 \leqslant i \leq k-1} (\beta_i + 1),\ \alpha_k > \frac{1}{2} \min_{1 \leqslant j \leqslant k-1} \alpha_j, \quad k \geqslant 2.$$

для смесей гамма-распределений.

В §2.4 рассматривается эффективность применения полученных асимптотически наиболее мощных критериев на практике. Проверяется правильность различения малых весов (вплоть до значений 0.01) на различных объемах выборки. Показано, что число успехов приближается к 100%, при этом число ошибок заведомо не превосходит уровень значимости критерия для каждого из случаев. Отмечены преимущества использования данных критериев по сравнению с критерием Ло.

Третья глава посвящена применению введенных в Главах 1 и 2 алгоритмов и техник повышения их эффективности. Отмечены новые для ряда практических областей результаты, которые были получены только с использованием полученных в диссертации методов.

В §3.1 описывается общая схема анализа хаотических процессов с применением метода скользящего разделения смесей (СРС-метод).

В §3.2 рассматривается анализ реальных данных с финансовых рынков с использованием СРС-метода. Найдены и проинтерпретированы портреты волатильности для различных финансовых индексов. Наибольшее внимание уделяется применению стохастических модификаций алгоритмов ЕМ-типа.

В §3.3 рассматривается анализ хаотических процессов в турбулентной плазме с использованием СРС-метода для различных алгоритмов ЕМ-типа. С помощью подобного анализа впервые была

определена структура хаотических процессов, протекающих в турбулентной плазме – было найдено их число $(3-5)$, определены параметры.

В §3.4 рассматривается альтернативный CPC-методу подход в анализе хаотических процессов в турбулентной плазме, базирующийся на рассмотрении «производных» величин от выборки (гистограммы, спектры), который можно рассматривать как одну из разновидностей бутстреп-процедур. Данный подход позволяет отследить, прежде всего, особенности функционирования системы на протяжении некоторого периода времени, за которое была построена анализируемая выборка. При этом объем выборки заранее предполагается весьма значительным (порядка нескольких сотен тысяч наблюдений), а истинная структура системы неизвестной. Проведенный анализ позволил получить взаимосвязь между результатами для гистограмм и для спектров, что заранее не предсказывалось теорией. Однако высокая степень согласия полученных результатов с экспериментальными данными позволяет предполагать, что данная связь является неслучайной, а потому представляет значительный интерес для исследований.

Работа выполнена под руководством доктора физико-математических наук, профессора Виктора Юрьевича Королева, которому автор выражает искреннюю благодарность.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИИ

1. *Г. М. Батанов, А. К. Горшенин, В. Ю. Королев, Д. В. Малахов, Н. Н. Скворцова.* Анализ статистических характеристик турбулентных пульсаций с помощью алгоритмов EM-типа // Материалы научной конференции «Тихоновские чтения». Москва, 2010. С. 62–63.

2. *Г. М. Батанов, А. К. Горшенин, В. Ю. Королев, Д. В. Малахов, Н. Н. Скворцова.* Эволюция вероятностных характеристик низкочастотной турбулентности плазмы в микроволновом поле // Математическое моделирование, 2011. Т. 23. № 5. С. 35–55.

3. *В. Е. Бенинг, А. К. Горшенин, В. Ю. Королев.* Асимптотически оптимальный критерий проверки гипотез о числе компонент смеси вероятностных распределений // Информатика и ее применения, 2011. Т. 5. Вып. 3. С. 4–15.

4. *А. К. Горшенин.* Медианные модификации стохастического EM-алгоритма для разделения смесей вероятностных распре-

делений и их применение к декомпозиции волатильности финансовых временных рядов // Сборник тезисов лучших дипломных работ 2008 года. М.: Издательский отдел факультета ВМиК МГУ им. М. В. Ломоносова, 2008. С. 62–63.

5. *А. К. Горшенин.* Применение медианной модификации SEM-алгоритма к задаче разделения смесей вероятностных распределений // Материалы XV Международной научной конференции студентов, аспирантов и молодых ученых «Ломоносов-2008»: секция «Вычислительная математика и кибернетика». М.: Издательский отдел факультета ВМиК МГУ им. М. В. Ломоносова, 2008. С. 30.

6. *А. К. Горшенин.* Проверка гипотез о числе компонент смеси вероятностных распределений // Обозрение прикладной и промышленной математики, 2011. Т. 18. Вып. 2.

7. *А. К. Горшенин.* Сравнение модификаций EM-алгоритма для декомпозиции волатильности финансовых временных рядов // Материалы XVI Международной научной конференции студентов, аспирантов и молодых ученых «Ломоносов-2009»: секция «Вычислительная математика и кибернетика». М.: Издательский отдел факультета ВМиК МГУ им. М. В. Ломоносова, 2009. С. 22.

8. *А. К. Горшенин.* Проверка статистических гипотез в модели расщепления компоненты // Вестник Московского Университета, 2011. Серия 15, Вычислительная математика и кибернетика. Т. 4.

9. *А. К. Горшенин, В. Ю. Королев, Д. В. Малахов, Н. Н. Скворцова.* Анализ тонкой стохастической структуры хаотических процессов с помощью ядерных оценок // Математическое моделирование, 2011. Т. 23. № 4. С. 83–89.

10. *А. К. Горшенин, В. Ю. Королев, Д. В. Малахов, Н. Н. Скворцова.* Бутстреп-методология структурного исследования хаотических процессов // Материалы научной конференции «Тихоновские чтения». Москва, 2010. С. 63–64.

11. *А. К. Горшенин, В. Ю. Королев, Д. В. Малахов, Н. Н. Скворцова.* Бутстреп-методология исследования структуры хаотических процессов // Материалы Второй международной научной конференции «Моделирование нелинейных процессов и систем». Москва, 2011. С. 219.

12. *А. К. Горшенин, В. Ю. Королев, А. М. Турсунбаев.* Медианные модификации EM-алгоритма для разделения смесей вероятностных распределений и их применение к декомпозиции волатильности финансовых индексов // Статистические методы оценивания и проверки гипотез, 2008. С. 169–195.

13. *А. К. Горшенин, В. Ю. Королев, А. М. Турсунбаев.* Медианные модификации EM- и SEM-алгоритмов для разделения смесей вероятностных распределений и их применение к декомпозиции волатильности финансовых временных рядов // Информатика и ее применения, 2008. Т. 2. Вып. 4. С. 12–47.

14. *А. К. Горшенин, В. Ю. Королев, С. Я. Шоргин.* CPC-методы как методы интеллектуального анализа данных при исследовании реальных хаотических процессов // Интеллектуальная обработка информации: 8-я международная конференция. Республика Кипр, г. Пафос, 17-24 октября 2010 г.: Сборник докладов. – М.: МАКС Пресс, 2010. С. 224–227. ISBN: 978-5-317-03409-2.

В работах [1, 2] Горшениным А. К. получены результаты анализа экспериментальных данных, построены портреты волатильности.

В работе [3] Горшенину А. К. принадлежат формулировка и доказательство теоремы о свойствах асимптотически оптимального критерия, доказательство достаточных условий конечности моментных характеристик, формулировка и доказательство теоремы об условиях идентифицируемости смесей равномерных распределений.

В работах [9, 10, 11] Горшенину А. К. принадлежат разработка предлагаемой в статьях бутстреп-процедуры, ее программная реализация, а также проведение анализа экспериментальных измерений параметров турбулентной плазмы.

В работе [12] Горшенину А. К. принадлежат программная реализация медианных методов и их применение к реальным финансовым данным.

В работе [13] Горшенину А. К. принадлежат программная реализация стохастических и медианных методов, построение портретов волатильности финансовых и тестовых данных.

В работе [14] Горшенину А. К. принадлежит описание особенностей применения CPC-методов к реальным хаотическим процессам.