Московский государственный университет
имени М. В. Ломоносова

На правах рукописи

Назаров Алексей Леонидович

Приближенные методы разделения смесей вероятностных распределений

Специальность 01.01.05 — теория вероятностей
и математическая статистика

Автореферат
диссертации на соискание ученой степени
кандидата физико-математических наук

Москва — 2013

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 Л. Г. Афанасьева
	кандидат физико-математических наук В. А. Лапшин
Ведущая организация:	Институт проблем информатики РАН

Защита диссертации состоится 26 апреля 2013 г. в 11 часов на заседании диссертационного совета Д 501.001.44 при Московском государственном университете имени М. В. Ломоносова по адресу: 119991, ГСП-1, Москва, Ленинские горы, МГУ, 2-й учебный корпус, факультет ВМК, аудитория 685. Желающие присутствовать на заседании диссертационного совета должны сообщить об этом за 2 дня до указанной даты по тел. (495) 939-30-10 (для оформления заявки на пропуск).

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

Автореферат разослан марта 2013 года.

Ученый секретарь
диссертационного совета,
к.т.н., в.н.с В.А. Костенко

Общая характеристика работы

Актуальность темы. Смеси вероятностных распределений как математические модели используются во многих задачах, например, при непараметрическом оценивании плотности и в кластерном анализе. Они демонстрируют высокую адекватность при описании неоднородных данных. Смеси вероятностных распределений хорошо зарекомендовали себя при описании хаотических процессов, моделирующих, к примеру, поведение цен финансовых инструментов, турбулентной плазмы.

Первые работы по исследованию и применению смесей вероятностных распределений появились еще в конце XIX века. К числу пионерских работ этого направления можно отнести работы С. Ньюкомба 1886 г.[1] и К. Пирсона 1894 г.[2] В них рассматривается смесь нормальных распределений, использующаяся для моделирования скошенных и островершинных распределений.

В общем случае задача разделения смеси вероятностных распределений заключается в поиске смеси из некоторого допустимого класса, которая в некотором смысле ближе всего к распределению наблюдаемой величины. Так как класс допустимых смесей чаще всего определяется классом допустимых смешивающих распределений, гарантирующих идентифицируемость модели, задача разделения смеси обычно сводится к задаче статистического оценивания смешивающего распределения по реализациям смеси[3,4,5,6].

Традиционными статистическими инструментами при решении задачи разделения смеси являются метод моментов и метод максимального правдоподобия. Как правило, для оценки смешивающего распределения в рассматриваемых задачах используется метод максимального правдоподобия. При этом ищется точка глобального максимума функции правдоподобия, соответствующая семейству допустимых смесей, как функция параметров. Численно данная задача может быть решена с помощью стандартных методов оптимизации[7]. Для решения задачи статистического разделения смесей вероятностных распределений также используются метод моментов, метод минимума

[1] Newcomb S. A generalized theory of the combination of observations so as to obtain the best result // American Journal of Mathematics.– 1886.– Vol. 8, no. 4.– Pp. 343–366.

[2] Pearson K. Contributions to the Mathematical Theory of Evolution // Philosophical Transactions of the Royal Society of London.– 1894.– Vol. 185.– Pp. 71–110.

[3] Everitt B., Hand D. J. Finite Mixture Distributions. Monographs on Applied Probability and Statistics.– Chapman and Hall, 1981.

[4] Titterington D. M., Smith A. F. M., Makov U. E. Statistical analysis of finite mixture distributions.– John Wiley & Sons, 1987.

[5] McLachlan G. J., Basford K. E. Mixture models. inference and applications to clustering.– 1988.

[6] McLachlan G., Peel D. Finite Mixture Models. Wiley series in probability and statistics: Applied probability and statistics.– Wiley, 2004.

[7] Васильев Ф. П. Методы оптимизации.– Москва: Факториал Пресс, 2002.

хи-квадрат, метод наименьших квадратов и пр. (см. работы О.К. Исаенко и В.Ю. Урбаха[8], J. Grim[9] и др.[5,6,7,8]).

Для конечных смесей нормальных распределений задача поиска оценок максимального правдоподобия может быть решена с помощью EM-алгоритма. EM-алгоритмом принято называть схему построения процедур итерационного типа для численного решения задачи поиска экстремума целевой функции в разнообразных задачах оптимизации. В частности, в прикладной статистике эта схема вполне работоспособна при поиске оценок максимального правдоподобия и родственных им в ситуациях, когда функция правдоподобия имеет сложную структуру, из-за которой другие методы оказываются неэффективными или вообще не применимы.

Появление первых электронно-вычислительных машин сделало возможной реализацию довольно сложных итерационных процедур, к числу которых принадлежит EM-алгоритм, и стимулировало дальнейшее развитие идей, лежащих в основе EM-алгоритма. Эти идеи нашли свое отражение в работах M. Healy, M. Westmacott, H.O. Hartley, S.F. Buck, М.И. Шлезингера, N. Day, J. Wolfe, B.J.N. Blight и др.

Поиск оценок весов и параметров компонент смеси с помощью EM-алгоритма работает эффективнее, чем стандартные методы оптимизации[10]. Однако в общем случае функция правдоподобия конечной смеси нормальных распределений нерегулярна, имеет много локальных максимумов (возможно, к тому же, бесконечных). Поэтому при численном решении данной задачи EM-алгоритм так же, как и стандартные методы оптимизации, становится крайне неустойчивым. К сожалению, последнее обстоятельство является серьезным препятствием для корректной интерпретации результатов применения данных алгоритмов к разделению конечных смесей нормальных законов.

В частности, было экспериментально установлено, что EM-алгоритм неустойчив по начальным данным. В некоторых случаях замена лишь одного наблюдения в выборке может кардинально изменить итоговые оценки, полученные с помощью EM-алгоритма[11]. Поэтому необходимо иметь альтернативные методы разделения смесей, ориентированные не на максимизацию «полной» функции правдоподобия, а на оптимизацию других разумных критериев качества получаемых оценок.

Одно из серьезных ограничений использования метода максимального правдоподобия для решения рассматриваемой задачи заключается в том, что класс допустимых смешивающих распределений чаще всего параметризо-

---

[8]Исаенко О. К., Урбах В. Ю. Разделение смесей распределений вероятностей на их составляющие // Итоги науки и техники.– Москва: ВИНИТИ, 1976.– Т. 13 из Теория вероятностей, математическая статистика и теоретическая кибернетика.– С. 37–58.

[9]Grim J. On numerical evaluation of maximum-likelihood estimates for finite mixtures of distributions // Kybernetika.– 1982.– Vol. 18, no. 3.– Pp. 173 – 190.

[10]Королев В. Ю. Вероятностно-статистические методы декомпозиции волатильности хаотических процессов.– Москва: Издательство Московского университета, 2011.

[11]Королев В. Ю. EM-алгоритм, его модификации и их применение к задаче разделения смесей вероятностных распределений. Теоретический обзор.– Москва: ИПИ РАН, 2007.

ван и задается точками из некоторого подмножества евклидова пространства. Это связанно с тем, что поиск точек глобального максимального правдоподобия обычно проводится численно. При этом задача поиска оценки смешивающего распределения, например, среди всех распределений, сосредоточенных на некотором компакте, не может быть решена с помощью данных методов без дополнительных ограничений.

Поиск новых эффективных методов является принципиально важным с точки зрения возможности адекватной практической интерпретации результатов работы алгоритмов разделения смесей. Именно такие альтернативные методы и предлагаются в данной работе.

Цель работы. Целью настоящей диссертации является описание и исследование класса приближенных методов разделения смесей вероятностных распределений – сеточных методов разделения смесей. Изучаются вопросы теоретического обоснования применимости данных методов и исследуются асимптотические свойства оценок, полученных с их помощью.

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

1. Получены верхние оценки устойчивости для масштабных, сдвиговых и сдвиг-масштабных смесей нормальных законов.

2. Исследованы вопросы существования нижних оценок устойчивости для подклассов смесей нормальных законов. Оценки для сдвиговых смесей нормальных распределений получены в явном виде.

3. Разработан, исследован и реализован класс сеточных методов разделения смесей. Проведено тестирование данных алгоритмов на различных наборах данных.

4. Исследованы асимптотические свойства оценок, получаемых с помощью сеточного метода максимального правдоподобия.

5. Доказана функциональная предельная теорема, описывающая сходимость оценок, полученных с помощью сеточного метода максимального правдоподобия для разделения масштабных смесей, при согласованном увеличении размера выборки и числа узлов сетки.

Объектом исследования является модель сдвиг-масштабных смесей нормальных распределений, а также класс сеточных методов разделения смесей, применяемый для статистического оценивания параметров модели по ее реализациям.

Методы исследования. Для решении задач стохастической устойчивости рассматриваемой модели в первой главе используются свойства вероятностных метрик(метрик Леви, Леви–Прохорова и др.), свойства прямых и обратных преобразований Фурье, теорема Планшереля, теорема Прохорова,

а также общие методы теории вероятностей. Асимптотические свойства оценок, полученных с помощью сеточных методов, рассматриваются с использованием результатов теории М-оценок и эмпирических процессов. Доказательство состоятельности данных оценок опираются на теоремы, описывающие сходимость вероятностных мер в пространстве Скорохода $D[0,1]$. Для исследования предложенных в работе алгоритмов применяется аппарат математической статистики и численные методы оптимизации. Сравнительное тестирование сеточных алгоритмов проводилось на выборках логарифмических приращений финансовых индексов Amex, Nasdaq, Nikkei, CAC 40. Данные алгоритмы были реализованы на языке C и с помощью пакета MATLAB.

Обоснованность научных положений. Теоретические положения и выводы диссертации сформулированы в виде утверждений, лемм, теорем и строго доказаны. Достоверность полученных результатов подтверждена экспериментальной проверкой алгоритмов и программ на реальных данных.

Теоретическая и практическая значимость Результаты диссертации являются вкладом в теорию статистического оценивания. Предложенный класс алгоритмов может применяться для численного решения задачи разделения смесей вероятностных распределений. Верхние и нижние оценки устойчивости смесей вероятностных распределений, полученные в работе и использующиеся для теоретического обоснования применимости данных методов, развивают теорию устойчивости стохастических моделей.

Апробация работы и публикации Основные результаты диссертации опубликованы в 10 печатных работах [1–10], 3 из которых – в журналах, включенных в перечень ВАК.

Результаты диссертации докладывались на ежегодной Международной конференции студентов, аспирантов и молодых ученых «Ломоносов» (Москва, Россия, апрель 2009 г. и апрель 2012 г.), на научной конференции «Ломоносовские чтения» 2012 года (Москва, Россия, 16-25 апреля 2012 г.), на XXX Международном семинаре по проблемам устойчивости стохастических моделей (Светлогорск, Россия, 24-30 сентября 2012 г.), на VI Международном рабочем семинаре «Прикладные задачи теории вероятностей и математической статистики, связанные с моделированием информационных систем» (Бер-Шева, Израиль, ноябрь 2012 г.), на научно-исследовательском семинаре «Теория риска и смежные вопросы» на факультете ВМК МГУ, на семинаре «Исследование асимптотического поведения и устойчивости стохастических моделей» на механико-математическом факультете МГУ, на семинаре «Моделирование финансовых рынков» в Высшей школе экономики.

Структура и объем диссертации Диссертация состоит из введения, трех глав, заключения и списка литературы. Полный объем диссертации **111** страниц текста с **12** рисунками. Список литературы содержит **70** наименований. Нумерация теорем и лемм в автореферате совпадает с нумерацией теорем и лемм в диссертации.

Содержание работы

Во **введении** дается обзор существующих методов разделения смесей вероятностных распределений, обосновывается актуальность темы диссертационного исследования, формулируются его цели, кратко излагается содержание работы.

В **первой главе** получены верхние оценки близости сдвиг-масштабных смесей нормальных законов через близость смешивающих распределений в метрике Леви–Прохорова, аналогичные оценки в метрике Леви рассмотрены для масштабных и сдвиговых смесей. Помимо этого, изучены нижние оценки устойчивости для различных подклассов смесей нормальных законов. Показано, что для сдвиговых смесей полученные оценки не могут быть улучшены без дополнительных ограничений. В заключение выписаны достаточные условия существования нижних оценок для произвольных классов смесей. Результаты первой главы используются во второй и третьей главах для обоснования применимости сеточных методов и исследования их свойств.

Основным объектом, который рассматривается во всех главах диссертации, является класс сдвиг-масштабных смесей нормальных законов. Пусть X – случайная величина, имеющая стандартное нормальное распределение, (U, V) – случайный элемент со значениями в $\mathbb{R}\times\mathbb{R}^+$, X и (U, V) заданы на одном вероятностном пространстве и стохастически независимы. Здесь и далее через Φ и ϕ соответственно обозначается функция распределения и плотность стандартного нормального закона.

Рассмотрим случайную величину

$$Y = VX + U, \quad \mathsf{P}(Y < x) = E\Phi\left(\frac{x-U}{V}\right), \quad x \in \mathbb{R}.$$

Распределение случайной величины Y – сдвиг-масштабная смесь нормальных законов. Если U вырождена в 0, распределение случайной величины Y также называется масштабной смесью, а в случае, когда V вырождена в 1, – сдвиговой смесью. В таком контексте чисто сдвиговая смесь $H(x) = E\Phi(x-U)$ является не чем иным, как функцией распределения суммы двух независимых случайных величин X и U, то есть сверткой их функций распределения. В то же время, чисто масштабная смесь $H(x) = E\Phi(\frac{x}{V})$ является не чем иным, как функцией распределения произведения двух независимых случайных величин X и V. Если это специально не оговорено, считается, что все случайные величины и элементы, рассматриваемые в данной главе, определены на одном вероятностном пространстве $(\Omega, \mathcal{A}, \mathrm{P})$.

В дальнейшем будем обозначать через ρ равномерное расстояние между функциями распределения, через L метрику Леви, через Π – метрику Леви-Прохорова.

При выводе верхних оценок для класса сдвиг-масштабных смесей используется следующая лемма, доказанная в первой главе.

Лемма 3. *Распределения P, Q двух случайных элементов X, Y, принимающих конечное число значений в $\mathbb{R}^2$, расстояние Леви-Прохорова между которыми $\Pi(P,Q) \le \delta$, можно представить в виде*

$$X: \quad P(\{x_i\}) = p_i,\ p_i > 0, \quad x_i \in \mathbb{R}^2,\ i = 1, \dots, s, \quad \sum_{i=1}^{s} p_i = 1,$$

$$Y: \quad Q(\{y_i\}) = q_i,\ q_i > 0, \quad y_i \in \mathbb{R}^2,\ i = 1, \dots, k, \quad \sum_{i=1}^{k} q_i = 1,$$

$$p_i = q_i, \quad d(x_i, y_i) \le \delta, \quad i = 1, \dots, n, \quad n \le \min\{s, k\},$$

$$\sum_{i=n+1}^{s} p_i \le \delta, \qquad \sum_{i=n+1}^{k} q_i \le \delta.$$

С помощью приведенной леммы в диссертации получены верхние оценки устойчивости для сдвиг-масштабных смесей нормальных законов.

Теорема 1. *Пусть случайные величины Y_1, Y_2 имеют вид:*

$$Y_1 = V_1 X + U_1 \ \textit{и} \ \ Y_2 = V_2 X + U_2, \quad X \sim \mathcal{N}(0,1), V_1 > 0, V_2 > 0,$$

причем случайная величина X и случайный вектор (U_i, V_i) стохастически независимы для $i = 1, 2$. Тогда справедливо неравенство

$$L(Y_1, Y_2) \le 3\Pi\Big(\big(U_1, C\ln(V_1)\big), \big(U_2, C\ln(V_2)\big)\Big), \tag{1}$$

где

$$C = \frac{1}{\ln(1 + \sqrt{2\pi e})}.$$

Отдельно получена аналогичная оценка расстояния между масштабными и между сдвиговыми смесями через расстояния между смешивающими распределениями в метрике Леви.

Теорема 2. *Пусть случайная величина X имеет стандартное нормальное распределение; U_1, U_2, V_1, V_2 – не зависящие от X случайные величины, $V_i > 0$, $i = 1, 2$. Тогда*

$$L(X + U_1, X + U_2) \le L(U_1, U_2), \quad L(V_1 X, V_2 X) \le 2L(C\ln(V_1), C\ln(V_2)),$$

где $C = \frac{1}{\ln(1+\sqrt{2\pi e})}$.

Структура полученного класса верхних оценок на первый взгляд может показаться необычной из-за того, что в неравенствах используются не сами случайные величины, отвечающие смешивающему распределению параметров масштаба, а их логарифмы. Однако переписать данные оценки с

использованием расстояния Леви между исходными величинами затруднительно. Это связанно с тем, что для двух положительных случайных величин V_1, V_2 и любого положительного числа c справедливо равенство

$$L(\ln(cV_1), \ln(cV_2)) = L(\ln(V_1), \ln(V_2)).$$

В то же время

$$\lim_{c \to +\infty} L(cV_1, cV_2) = \rho(V_1, V_2),$$

а ρ в общем случае не метризует слабую сходимость.

Далее в первой главе получены нижние оценки устойчивости для класса сдвиговых смесей нормальных законов, то есть для сверток некоторых смешивающих распределений со стандартным нормальным законом.

Теорема 3. *Пусть случайная величина X имеет стандартное нормальное распределение; случайные величины U_1, U_2 не зависят от X и имеют конечные первые моменты, F_1, F_2 – соответствующие им функции распределения. Если $L(F_1, F_2) \geq \delta$, то*

$$\|F_1 * \Phi - F_2 * \Phi\|_2 \geq \sqrt{\frac{\delta^3}{2}} \exp\left\{-\frac{2}{\pi^2 \delta^6}\right\}.$$

Нижняя оценка близости сдвиговых смесей, полученная в теореме 3, на первый взгляд может показаться достаточно грубой. В работе построен пример двух смесей, показывающий, что без дополнительных условий данная оценка не может быть принципиально улучшена.

Утверждение 1. *Пусть случайная величина X имеет стандартное нормальное распределение. Существуют не зависящие от X случайные величины U_1, U_2 такие, что расстояние Леви $L(U_1, U_2) \geq \delta$, $\delta \in (0, \frac{4-\pi}{8+\pi})$, а*

$$L(X + U_1, X + U_2) \leq \left(\frac{21}{(4-\pi)^2}\right)^{\frac{1}{3}} \delta \exp\left(-\frac{\pi^2}{3\delta^2}\right).$$

Верхние оценки устойчивости, описывающие близость масштабных смесей через близость соответствующих им смешивающих распределений, получены с использованием метрики Леви в начале первой главы. К сожалению, аналогичные результаты для нижних оценок устойчивости не могут быть выписаны без дополнительных ограничений на смешивающее распределение.

Лемма 7. *Пусть случайная величина X имеет стандартное нормальное распределение. Справедливы следующие утверждения.*

1. *Для любых $\varepsilon \in (0, 1)$, $\delta > 0$ существуют не зависящие от X положительные случайные величины V_1, V_2 такие, что $L(V_1, V_2) \geq \varepsilon$, $L(XV_1, XV_2) < \delta$.*

2. *Для любых* $\varepsilon \in (0,1)$, $\delta > 0$ *существуют не зависящие от* X *положительные случайные величины* V_1, V_2 *такие, что* $L(\ln V_1, \ln V_2) \geq \varepsilon$, $L(XV_1, XV_2) < \delta$.

Если потребовать выполнения некоторых дополнительных условий, можно доказать существование искомых оценок и для сдвиговых, и для масштабных смесей нормальных законов. В диссертации данное утверждение сформулировано для произвольного класса смесей, порожденных некоторым ядром. Смеси нормальных законов являются частным случаем класса, описанного в теореме.

Рассматривается семейство смесей $\mathfrak{A} = \{P_t, t \in T\}$ и семейство соответствующих смешивающих распределений $\mathfrak{B} = \{Q_t, t \in T\}$, заданных на измеримых пространствах $(S_{\mathfrak{A}}, \mathcal{F}_{\mathfrak{A}})$, $(S_{\mathfrak{B}}, \mathcal{F}_{\mathfrak{B}})$ соответственно. Пусть T – некоторое параметрическое множество, и для любого $t \in T$ смеси P_t соответствует смешивающее распределение Q_t. Через $\bar{\mathfrak{B}}$ обозначается семейство, содержащее все распределения, заданные на $(S_{\mathfrak{B}}, \mathcal{F}_{\mathfrak{B}})$ и являющиеся слабыми пределами последовательностей из $\mathfrak{B}$. Семейству $\bar{\mathfrak{B}}$ соответствуют семейство смесей $\bar{\mathfrak{A}}$ и параметрическое множество $\bar{T}$. Предполагается, что α и β метризуют слабую сходимость в $\bar{\mathfrak{A}}$ и $\bar{\mathfrak{B}}$ соответственно.

Теорема 4. *Пусть* $\bar{\mathfrak{A}}$ *идентифицируемо,* $\mathfrak{B}$ *плотно. Если для* $\bar{\mathfrak{A}}$ *существуют верхние оценки устойчивости вида*

$$\forall \varepsilon > 0 \;\; \forall t_1, t_2 \in \bar{T} \;\; \beta(Q_{t_1}, Q_{t_2}) \leq \varepsilon \;\; \Rightarrow \alpha(P_{t_1}, P_{t_2}) \leq g_u(\varepsilon),$$

где g_u *– некоторая вещественная функция, определенная на вещественной положительной полуоси,*

$$\lim_{x \downarrow 0} g_u(x) = 0,$$

то для этого семейства должны существовать нижние оценки устойчивости вида

$$\forall \varepsilon > 0 \;\; \forall t_1, t_2 \in T \;\; \beta(Q_{t_1}, Q_{t_2}) \geq \varepsilon \;\; \Rightarrow \alpha(P_{t_1}, P_{t_2}) \geq g_l(\varepsilon) > 0,$$

где g_l *– некоторая вещественная функция, определенная на вещественной положительной полуоси.*

Доказательство состоятельности оценок, полученных с помощью сеточных методов разделения смесей вероятностных распределений, опирается на теоремы существования, доказанные в первой главе. Кроме этого, приведенные примеры позволяют лучше понять, насколько сильно могут изменяться оценки смешивающего распределения при работе с реализациями одной и той же смеси.

Вторая глава посвящена описанию сеточных методов разделения смесей вероятностных распределений. Работа сеточных методов рассматри-

вается сначала на примере конечной смеси функций распределения вида

$$F(x) = \sum_{i=1}^{k} p_i \Phi\Big(\frac{x - a_i}{\sigma_i}\Big), \qquad x \in \mathbb{R}, \tag{2}$$

где $k \geq 1$ – целое. В классической задаче разделения смесей параметрами, подлежащими статистическому оцениванию, являются тройки (p_i, a_i, σ_i), $i = 1, \ldots, k$, где $a_i \in \mathbb{R}$, $\sigma_i > 0$, $p_i \geq 0$, $p_1 + \ldots + p_k = 1$.

Предполагается, что заранее известны числа $\underline{a}$, $\overline{a}$ и $\overline{\sigma}$ такие, что $\underline{a} \leq a_i \leq \overline{a}$ и $\sigma_i \leq \overline{\sigma}$ при всех $i = 1, \ldots, k$. Другими словами, считается, что известны диапазоны изменения неизвестных параметров a_i и σ_i.

Идея, лежащая в основе рассматриваемого подхода, заключается в замене *интервалов* $[\underline{a}, \overline{a}]$ и $(0, \overline{\sigma}]$ возможных значений неизвестных параметров масштаба σ_i и сдвига a_i *дискретными* множествами известных точек. Эти точки могут быть выбраны, например, исходя из следующих соображений.

Пусть ε_a и ε_σ – положительные числа, определяющие априорные требования к точности оценивания параметров a_i и σ_i:

$$\max_i |a_i - \widehat{a}_i| \leq \varepsilon_a, \qquad \max_i |\sigma_i - \widehat{\sigma}_i| \leq \varepsilon_\sigma, \tag{3}$$

где $\widehat{a}_i$ и $\widehat{\sigma}_i$ – искомые оценки параметров. Числа ε_a и ε_σ также можно интерпретировать как пороги различимости возможных значений параметров: значения a', a'' и σ', σ'' соответственно считаются неразличимыми, если

$$|a' - a''| \leq \varepsilon_a, \qquad |\sigma' - \sigma''| \leq \varepsilon_\sigma. \tag{4}$$

Положим $k_a = [(\overline{a} - \underline{a})/\varepsilon_a] + 1$, $k_\sigma = [\overline{\sigma}/\varepsilon_\sigma] + 1$, где $[z]$ обозначает целую часть числа z. Для $r = 1, 2, \ldots, k_a + 1$ положим $\widetilde{a}_r = \underline{a} + (r-1)\varepsilon_a$. Аналогично для $l = 1, 2, \ldots, k_\sigma$ положим $\widetilde{\sigma}_l = l\varepsilon_\sigma$. Тогда точки с координатами $(\widetilde{a}_r, \widetilde{\sigma}_l)$ образуют узлы конечной сети, покрывающей прямоугольник $\{(a, \sigma) : \underline{a} \leq a \leq \overline{a}, 0 \leq \sigma \leq \overline{\sigma}\}$, который представляет собой множество возможных значений параметров сдвига и масштаба компонент смеси (2). Параметры масштаба не равны нулю, чтобы избежать возможной некорректности. Число узлов полученной сети равно $K = (k_a + 1)k_\sigma$. Для удобства записи и упрощения обозначений перенумеруем каким-либо образом узлы указанной сети, вводя *единый* индекс i для координат $(\widetilde{a}_i, \widetilde{\sigma}_i)$ узла с номером i после перенумерации, $i = 1, \ldots, K$.

Основная идея предлагаемого подхода состоит в аппроксимации смеси (2) смесью с заведомо бóльшим числом *известных* компонент:

$$F(x) = \sum_{i=1}^{k} p_i \Phi\Big(\frac{x - a_i}{\sigma_i}\Big) \approx \sum_{i=1}^{K} \widetilde{p}_i \Phi\Big(\frac{x - \widetilde{a}_i}{\widetilde{\sigma}_i}\Big) \equiv \widetilde{F}(x), \qquad x \in \mathbb{R}. \tag{5}$$

Такое приближение допустимо, поскольку в силу соотношений (3) и (4) для любой пары (a_r, σ_r) параметров компоненты смеси (2) обязательно найдется

практически неотличимая от нее пара $(\widetilde{a}_i, \widetilde{\sigma}_i)$ параметров компоненты смеси $\widetilde{F}$. Веса остальных компонент смеси $\widetilde{F}$, для параметров которых не найдется "близкой" пары параметров (a_r, σ_r) компоненты смеси (2), можно считать равными нулю. Действительно, в этом случае смешивающие распределения будут, например, близки в метрике Леви-Прохорова, а из этого напрямую следует близость самих смесей (см. теорему 1).

В общем случае для произвольного класса сдвиг-масштабных смесей нормальных законов применяется точно такая же аппроксимация. Таким образом, при использовании сеточных методов оценка смешивающего распределения ищется в классе распределений, сосредоточенных в конечном числе точек, то есть на сетке. При этом точки расположены в области, содержащей носитель смешивающего распределения

$$F(x) = E\Phi\left(\frac{x-U}{V}\right) \approx \sum_{i=1}^{K} \widetilde{p}_i \Phi\Big(\frac{x-\widetilde{a}_i}{\widetilde{\sigma}_i}\Big) \equiv \widetilde{F}(x), \qquad x \in \mathbb{R}. \tag{6}$$

Другими словами, при статистическом оценивании в качестве допустимого рассматривается не класс всех распределений, сосредоточенных на $\mathbb{R} \times \mathbb{R}^+$, а лишь некоторое его подмножество. Обосновать такую замену позволяют верхние оценки устойчивости сдвиг-масштабных смесей нормальных законов. Заметим, что неизвестными параметрами смеси $\widetilde{F}$ являются *только* веса $\widetilde{p}_1, \ldots, \widetilde{p}_K$.

В диссертации предложены методы приближенного разделения смесей, основанные на

- (i) минимизации невязки между теоретическими и эмпирическими моментами
- (ii) максимизации сеточной функции правдоподобия.

Показано, что задачи типа (i) могут быть сведены к задачам линейного программирования. Для численного решения задач типа (ii) предложены "усеченный" EM-алгоритм и алгоритм условного градиента. Приведены результаты сравнительного анализа эффективности предложенных методов.

В сеточном методе максимального правдоподобия поиск оценок весов проводится с помощью итерационного процесса. В функцию правдоподобия, соответствующую распределениям, сосредоточенным на сетке, подставляются реализации случайной величины с оцениваемым распределением. Далее численно ищется точка максимума этой функции, как функции параметров. Распределение, соответствующее полученным величинам, используется в качестве оценки истинного. В работе доказывается, что максимизируемая функция выпукла вверх по неизвестным параметрам.

Теорема 5. *Любая «сеточная» функция правдоподобия как функция весов вогнута.*

Теорема 6. *Сеточный метод максимального правдоподобия, реализованный с помощью алгоритма условного градиента, сходится к точке глобального максимума «сеточной» функции правдоподобия.*

Таким образом, в отличие от классического EM-алгоритма, итерационный процесс в предложенном методе является устойчивым по начальным приближениям.

Наибольшую эффективность при решении задачи разделения смесей вероятностных распределений демонстрирует сеточный метод максимального правдоподобия, реализованный с помощью алгоритма условного градиента. Этот вывод сделан как на основе исследования свойств алгоритмов, так и на основе анализа применения различных сеточных методов разделения смесей к решению задачи статистической декомпозиции волатильности финансовых индексов.

В **третьей главе** исследуются асимптотические свойства оценок смешивающего распределения, полученных с помощью сеточного метода максимального правдоподобия для масштабных смесей нормальных законов. К сожалению, классическая теория, описывающая свойства оценок максимального правдоподобия, не может быть применена для изучения свойств последних. Для их исследования используются аппараты М-оценок[12] и эмпирических процессов[13]. Рассмотрены достаточные условия состоятельности оценок, описывающие алгоритм согласованного изменения структуры параметрического множества и числа элементов выборки.

Для формального описания процесса "измельчения" сетки, на которой сосредоточен класс распределений, приближающих исходное смешивающее распределение, вводятся следующие обозначения.

Рассмотрим некоторое множество $Z_0 \subset \mathbb{R}$, содержащее носитель смешивающего распределения. Обозначим через $Z_1 \subset Z_2 \subset \ldots$ последовательность вложенных конечных подмножеств Z_0, исчерпывающую счетное, всюду плотное подмножество Z_0. Сопоставим каждому множеству Z_i параметрическое множество

$$\Theta_i = \Big\{ \{(\sigma_j, f_j)\}_{j=1}^{k_i} |\ \sigma_j \in Z_i,\ \sigma_j < \sigma_{j+1}, \quad f_j = \sum_1^j p_l,$$

$$p_j \geq 0, \quad j = 1, \ldots, k_i, \quad p_1 + \ldots + p_{k_i} = 1, \quad k_i = \#Z_i \Big\}, \quad i = 1, 2, \ldots$$

Рассмотрим метрические пространства $(\Theta_i, d_i) \equiv (\Theta_i, d)$, $i = 1, 2, \ldots$, где расстояние между двумя элементами θ^1, $\theta^2 \in \Theta_i$ задается через

$$d \equiv d_i(\theta^1, \theta^2) = \max_{1 \leq j \leq \#Z_i} |f_j^{(1)} - f_j^{(2)}|,$$

[12]van der Vaart A. W. Asymptotic Statistics (Cambridge Series in Statistical and Probabilistic Mathematics).– Cambridge University Press, 2000.

[13]Pollard D. Empirical Processes: Theory and Applications. Hayward, CA: Institute of Mathematical Statistics, 1990.– Vol. 2 of NSF-CBMS Regional Conference Series in Probability and Statistics.

$i = 1, 2, \ldots$

Далее рассмотрим статистическую структуру $(\mathfrak{X}, \mathcal{F}, \mathcal{P})$, где

$$\mathfrak{X} = \mathbb{R} \times \mathbb{R} \times \ldots,$$

$$\mathcal{F} = \mathcal{B}(\mathbb{R}) \times \mathcal{B}(\mathbb{R}) \times \ldots,$$

$\mathcal{B}(\mathbb{R})$ – борелевская сигма-алгебра, $\mathcal{P} = \{P_\theta, \theta \in \Theta_i, i = 1, 2, \ldots\} \cup P_0$ – параметризованное семейство вероятностных распределений, отвечающих последовательностям независимых одинаково распределенных случайных величин $(X_1, X_2, X_3, \ldots)$. Здесь P_0 – распределение реализаций оцениваемой смеси. Для каждого фиксированного i класс $\{P_\theta, \theta \in \Theta_i\}$ содержит распределения, приближающие истинное среди смесей, соответствующих смешивающим распределениям, сосредоточенным на множестве Z_i.

Семейству $\mathcal{P}$ соответствует семейство

$$\mathcal{P}^1 = P_0^1 \cup \{P_\theta^1, \theta \in \Theta_i,\ i = 1, 2, \ldots\}$$

одномерных распределений. Одномерные распределения данных последовательностей для $\{P_\theta, \theta \in \Theta_i, i = 1, 2, \ldots\}$ являются конечными смесями нормальных законов. При этом каждому фиксированному $\theta \in \Theta_i$, $i = 1, 2, \ldots$, соответствует плотность одномерного распределения последовательности относительно меры Лебега на $\mathbb{R}$:

$$p_\theta(x) = \sum_{j=1}^{k_i} p_j \frac{1}{\sigma_j} \phi\Big(\frac{x}{\sigma_j}\Big), \qquad x \in \mathbb{R},$$

где ϕ – плотность стандартного нормального закона. Заметим, что от P_0, вообще говоря, лишь требуется, чтобы оно соответствовало распределению некоторой масштабной смеси нормальных законов. Будем считать, что P_0^1 имеет плотность p_0 относительно меры Лебега.

Рассмотрим статистики, с помощью которых проводится оценка истинного распределения. Зафиксируем некоторый номер i сетки Θ_i. Как было описано выше, в функцию правдоподобия, соответствующую распределениям из набора $\{P_\theta, \theta \in \Theta_i\}$, подставляются реализации независимых случайных величин с распределением P_0^1. Затем ищется точка максимума этой функции как функции параметров.

В работе исследуются свойства последовательности оценок $\hat{\theta}_n^{(i)}$:

$$\hat{\theta}_n^{(i)} = \arg\max_{\theta \in \Theta_i} M_n(\theta),$$

$$M_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} m_\theta(X_i), \quad m_\theta(x) = \ln \frac{p_\theta(x)}{p_{\theta_0}(x)}, \quad \theta \in \Theta_i.$$

Такие оценки принадлежат классу M-оценок. При выполнении ряда условий они обладают некоторыми важными свойствами. Для их исследования изучаются свойства целевой функции M и последовательности точек $\theta_0^{(i)}$:

$$\theta_0^{(i)} = \arg\min_{\theta \in \Theta} D_{KL}(P_0^1, P_\theta^1),$$

где D_{KL} – расстояние Кульбака–Лейблера. Для распределений P, Q, имеющих плотность p, q относительно меры Лебега,

$$D_{KL}(P,Q) = \int_{\mathbb{R}} \ln \frac{p(x)}{q(x)} p(x)dx$$

(считается, что $\min_{\theta\in\Theta} D_{KL}(P_0^1, P_\theta^1) < \infty$).

Последовательность параметров $\{\theta_0^{(j)}\}_{j=1}^{\infty}$ соответствует последовательности распределений, сосредоточенных на сетках $\{Z_j\}_{j=1}^{\infty}$, ближайших к истинному в смысле расстояния Кульбака–Лейблера.

Используя введенные обозначения, зафиксируем параметрическое множество $\Theta = \Theta_i$ для одного из значений i, $i \in \mathbb{N}$.

Лемма 8. *Для* $d(\theta_0, \theta) > \delta$, $\theta \in \Theta$,

$$P_0^1 m_{\theta_0} - P_0^1 m_\theta \geq C_H(\Theta)\delta^2,$$

где

$$P_0^1 f \equiv \int_{\mathbb{R}} f(x) P_0^1(dx)$$

для интегрируемой функции f, $C_H(\Theta)$ – *константа, зависящая от множества* Θ.

Последовательность рассматриваемых оценок имеет предел.

Теорема 7. *Пусть для последовательности оценок* $\{\hat\theta_n\}_{n=1}^{\infty}$ *существует последовательность* $\{\alpha_n\}_{n=1}^{\infty}$ *такая, что*

$$M_n(\hat\theta_n) \geq M_n(\theta_0) - \alpha_n, \quad \alpha_n \xrightarrow{P_0} 0,$$

и $\int_{\mathbb{R}} \sup\limits_{\theta\in\Theta} m_\theta(x) P_0^1(dx) < \infty$, *тогда для любого* $\varepsilon > 0$

$$P_0(d(\hat\theta_n, \theta_0) \geq \varepsilon) \to 0.$$

Из теоремы 7 следует, что в сеточном методе максимального правдоподобия при достаточно большом размере выборки смешивающее распределение оценивается среди распределений, сосредоточенных на известной сетке, с помощью ближайшего к истинному в смысле расстояния Кульбака–Лейблера.

В работе получены оценки скорости сходимости в сеточном методе разделения смесей вероятностных распределений.

Теорема 8. *Пусть* Θ – *некоторое параметрическое множество из последовательности* $\{\Theta_i, i \in \mathbb{N}\}$, $p_k \geq \frac{c_\sigma}{\sqrt{k}}$ *для всех* $\theta \in \Theta$, $\hat\theta_n$ – *оценка, полученная при использовании сеточного метода, такая что*

$$M_n(\hat\theta_n) \geq M_n(\theta_0) - \frac{C_H(\Theta)}{2nV^2(\Theta)}. \tag{7}$$

Тогда $\forall \lambda > 0$

$$P_0\left(\sqrt{n}V(\Theta)d(\hat{\theta}_n, \theta_0) > \lambda\right) \leq \frac{C}{\lambda},$$

где

$$V(\Theta) = \frac{C_H(\Theta)}{\frac{k}{c_\sigma}\sum_{j=1}^{k}\frac{\sigma_k}{\sigma_j}},$$

C *– некоторая константа.*

Далее исследуется сходимость последовательности получаемых оценок в функциональном пространстве при согласованном увеличении размеров сетки и числа элементов выборки, по которой строится оценка.

Каждому $\theta^{(i)} \in \Theta_i$ ставится в соответствие элемент $x_{\theta^{(i)}}$ пространства Скорохода $D[0,1]$ следующим образом. Пусть F_0 – функция оцениваемого смешиваемого распределения, тогда

$$x_{\theta^{(i)}}(t) = F_{\theta^{(i)}}\left(F_0^{-1}(t)\right), \quad t \in [0,1), \quad x_{\theta^{(i)}}(1) = 1,$$

$$F_0^{-1}(t) = \inf_{x\in\mathbb{R}}\{x|F_0(x) \geq t\}, \quad t \in [0,1),$$

$$F_{\theta^{(i)}}(y) = \min_{j=1\ldots k_i}\{f_j|\sigma_j \geq y\}, \quad y \in \mathbb{R}.$$

С учетом этого $\hat{\theta}_n^{(i)}$ соответствует семейство конечномерных распределений $\mu_{\hat{\theta}_n^{(i)}}$:

$$\mu_{\hat{\theta}_n^{(i)};t_1,t_2,\ldots,t_l}(A) = P_0\left((x_{\hat{\theta}_n^{(i)}}(t_1),\ldots,x_{\hat{\theta}_n^{(i)}}(t_l)) \in A\right),$$

$$0 \leq t_1 < t_2 < \ldots < t_l \leq 1, \qquad A \in \mathcal{B}(\mathbb{R}^l).$$

Данное семейство является согласованным. Справедливо следующее утверждение.

Лемма 12. *Пусть* Θ_i *– некоторое параметрическое множество из последовательности* $\{\Theta_i, i \in \mathbb{N}\}$, $p_{k_i} \geq \frac{c_\sigma}{\sqrt{k_i}}$ *для всех* $\theta \in \Theta_i$, $\hat{\theta}_n^{(i)}$ *– оценка, полученная при использовании сеточного метода,*

$$M_n(\hat{\theta}_n^{(i)}) \geq M_n(\theta_0^{(i)}) - \frac{C_H(\Theta_i)}{2nV^2(\Theta_i)}.$$

Тогда в $D[0,1]$ *существует случайный элемент* $X_{\hat{\theta}_n^{(i)}}$ *с конечномерными распределениями* $\mu_{\hat{\theta}_n^{(i)}}$.

Таким образом, каждому значению параметра θ соответствует некоторая траектория x_θ, а каждой оценке $\hat{\theta}_n^{(i)}$ – случайный процесс $X_{\hat{\theta}_n^{(i)}}$ со значениями в пространстве Скорохода $D[0,1]$ с заданными выше конечномерными распределениями.

Теорема 9. Рассмотрим последовательность случайных элементов $\{X_{\hat{\theta}_n^{(r(n))}}\}_{n=1}^{\infty}$ *со значениями в пространстве Скорохода* $D[0,1]$, *где* $r(n)$ – *возрастающая последовательность натуральных чисел.*

Пусть выполнены следующие условия:

1. *Последовательность распределений* $P^1_{\theta_0^{(n)}}$ *сходится в смысле расстояния Кульбака–Лейблера:*

$$D_{KL}(P_0^1, P^1_{\theta_0^{(n)}}) \to 0 \quad \textit{при } n \to \infty.$$

2. *Для каждого элемента последовательности параметрических множеств* $\{\Theta_i\}_{i=1}^{\infty}$ *справедливо неравенство* $p_{k_i} \geq \frac{c_\sigma}{\sqrt{k_i}}$, *и*

$$M_n(\hat{\theta}_n^{(i)}) \geq M_n(\theta_0^{(i)}) - \frac{C_H(\Theta_i)}{2nV^2(\Theta_i)},$$

где

$$V(\Theta_i) = \frac{C_H(\Theta_i)}{\frac{k_i}{c_\sigma}\sum_{j=1}^{k_i}\frac{\sigma_k^{(i)}}{\sigma_j^{(i)}}},$$

3. *Последовательность чисел* $\{r(n)\}_{n=1}^{\infty}$ *такова, что*

$$\sqrt{n}V(\Theta_{r(n)}) \to \infty \quad \textit{при } n \to \infty,$$

и для некоторого $\alpha > 0$ *существует константа* K*:*

$$\sqrt{n}V(\Theta_{r(n)})\delta_{r(n)}^{2+\alpha} \geq K, \quad \forall n \in \mathbb{N},$$

где $\delta_{r(n)}$ – *длина минимального отрезка, на котором* $x_{\theta_0^{(r(n))}}$ *постоянна.*

4. *Семейство логарифмов случайных величин, отвечающих* $\mathcal{P}^1 = \{P^1_{\theta_i}, \theta_i \in \Theta_i, i = 1, 2, \ldots\} \cup P_0^1$, – *плотно.*

Тогда

$$X_{\hat{\theta}_n^{(r(n))}} \xrightarrow{d} x_0.$$

Как известно, сходимость по распределению к неслучайной функции влечет за собой сходимость по вероятности. Следовательно, оценки смешивающих распределений, полученные с помощью сеточных методов разделения смесей вероятностных распределений, для указанных классов смесей нормальных законов являются состоятельными при согласованном увеличении размера выборки и числа узлов сетки.

Результаты, выносимые на защиту

1. Верхние оценки устойчивости для масштабных, сдвиговых и сдвиг-масштабных смесей нормальных законов.
2. Результаты исследования вопросов существования нижних оценок устойчивости для подклассов смесей нормальных законов.
3. Класс сеточных методов разделения смесей.
4. Результаты исследования асимптотических свойств оценок, получаемых с помощью сеточного метода максимального правдоподобия.
5. Функциональная предельная теорема, описывающая сходимость оценок, полученных с помощью сеточного метода максимального правдоподобия для разделения масштабных смесей, при согласованном увеличении размера выборки и числа узлов сетки.

Список публикаций автора по теме диссертации

1. *Назаров А. Л.* Разделение смесей вероятностных распределений сеточным методом максимального правдоподобия при помощи алгоритма условного градиента // *Сб. статей молодых ученых факультета ВМиК МГУ.* — 2009. — № 6. — С. 128–135.
2. *Korolev V. Y., Nazarov A. L.* Separating mixtures of probability distributions with the grid method of moments and the grid maximal likelihood method. // *Autom. Remote Control.* — 2010. — Vol. 71, no. 3. — Pp. 455–472.
3. *Королев В. Ю., Назаров А. Л.* Разделение смесей вероятностных распределений при помощи сеточных методов моментов и максимального правдоподобия. // *Автомат. и телемех.* — 2010. — Т. 71, № 3. — С. 98–116.
4. *Назаров А. Л.* Об устойчивости смесей вероятностных законов к возмущениям смешивающих распределений // *Статистические методы оценивания и проверки гипотез.* — 2010. — № 22. — С. 154–172.
5. *Назаров А. Л.* Нижние оценки в задаче устойчивости смесей нормальных распределений к возмущениям смешивающих распределений // *Информатика и ее применения.* — 2012. — Т. 6, № 4. — С. 24–32.
6. *Назаров А. Л.* Асимптотические свойства оценок, полученных с помощью сеточных методов разделения смесей вероятностных распределений // *Статистические методы оценивания и проверки гипотез.* — 2012. — № 24. — С. 22–35.
7. *Назаров А. Л.* О состоятельности оценок параметров масштабных смесей нормальных распределений, получаемых с помощью сеточных методов // *Системы и средства информатики.* — 2012. — Т. 22, № 2. — С. 227–243.

8. *Nazarov A. L.* Asymptotic properties of grid method estimators in the normal mixture separation problems // XXX International Seminar on Stability Problems for Stochastic Models (ISSPSM'2012) and VI International Workshop "Applied Problems in Theory of Probabilities and Mathematical Statistics Related to Modeling of Information Systems" (APTP + MS'2012). Book of abstracts. — Moscow: IPI RAS, 2012. — Pp. 56–57.

9. *Nazarov A. L.* Lower bounds for the stability of normal mixture models with respect to perturbations of mixing distribution // XXX International Seminar on Stability Problems for Stochastic Models (ISSPSM'2012) and VI International Workshop "Applied Problems in Theory of Probabilities and Mathematical Statistics Related to Modeling of Information Systems" (APTP + MS'2012). Book of abstracts. — Moscow: IPI RAS, 2012. — Pp. 57–59.

10. *Nazarov A. L.* On the consistency of grid method estimators in gaussian scale mixture separation problems // VI International Workshop "Applied Problems in Theory of Probabilities and Mathematical Statistics Related to Modeling of Information Systems" (APTP + MS'2012), Autumn Session, 2012. Book of abstracts. — Moscow: IPI RAS, 2012. — Pp. 77–81.