На правах рукописи

СМИРНОВ Павел Олегович

РОБАСТНЫЕ МЕТОДЫ И АЛГОРИТМЫ ОЦЕНИВАНИЯ КОРРЕЛЯЦИОННЫХ ХАРАКТЕРИСТИК ДАННЫХ НА ОСНОВЕ НОВЫХ ВЫСОКОЭФФЕКТИВНЫХ И БЫСТРЫХ РОБАСТНЫХ ОЦЕНОК МАСШТАБА

Специальность 05.13.18 – Математическое моделирование,
численные методы и комплексы программ

АВТОРЕФЕРАТ

диссертации на соискание учёной степени
кандидата физико-математических наук

Санкт-Петербург – 2013

Работа выполнена в Федеральном государственном бюджетном образовательном учреждении высшего профессионального образования «Санкт-Петербургский государственный политехнический университет».

Научный руководитель:	доктор физико-математических наук, профессор, ШЕВЛЯКОВ Георгий Леонидович
Официальные оппоненты:	НИКИТИН Яков Юрьевич, доктор физико-математических наук, профессор, заведующий кафедрой теории вероятностей и математической статистики ФГБОУ ВПО «Санкт-Петербургский государственный университет»
	ПРОУРЗИН Владимир Афанасьевич, кандидат физико-математических наук, старший научный сотрудник лаборатории методов анализа надёжности ФГБУН «Институт проблем машиноведения» Российской академии наук
Ведущая организация:	ФГБУН «Институт проблем управления им. В. А. Трапезникова» Российской академии наук

Защита состоится 26 марта 2014 г. в 18 часов на заседании диссертационного совета Д 212.229.13 при ФГБОУ ВПО «Санкт-Петербургский государственный политехнический университет», расположенном по адресу: 195251, Санкт-Петербург, Политехническая ул., д. 29, I уч. корп., ауд. 41.

С диссертацией можно ознакомиться в фундаментальной библиотеке ФГБОУ ВПО «Санкт-Петербургский государственный политехнический университет» по адресу 195251, Санкт-Петербург, Политехническая ул., д. 29. Автореферат диссертации доступен на официальном сайте СПбГПУ (http://www.spbstu.ru/).

Автореферат разослан «_____» февраля 2014 г.

Ученый секретарь
диссертационного совета Д 212.229.13,
доктор технических наук, профессор  Григорьев Борис Семёнович

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования. При исследовании закона распределения одномерных случайных величин по результатам наблюдений основное внимание уделяется описанию таких двух главных характеристик как его мера положения (некоторое типичное значение) и мера разброса значений вокруг этой центральной точки. Так, широко использующийся нормальный закон распределения случайных величин полностью характеризуется первыми двумя моментами — математическим ожиданием (положением) и дисперсией (масштабом), и для их оценки в классической статистике чаще всего используются выборочные аналоги: среднее арифметическое и среднеквадратичное отклонение.

При наличии нескольких случайных величин или одной многомерной величины встаёт вопрос о взаимосвязи составляющих этой системы. Мерой их линейной зависимости является коэффициент корреляции или, в случае размерности больше двух, корреляционная матрица, которая наряду с математическим ожиданием и дисперсией полностью описывает нормально распределённые случайные величины.

Тем не менее, любые модели лишь приближённо описывают реальные явления, и на практике мы сталкиваемся с различными отклонениями от них. В силу этого, использование классических параметрических моделей распределений для оценивания их параметров не всегда оправдано, может привести к некорректным результатам, и, соответственно, поставить под сомнение обоснованность применения связанных с ними процедур. Возникшая на смену классическим моделям непараметрическая статистика, отказавшись от всяких предположений о конкретном виде закона распределения случайных величин, позволила находить приемлемое решение части задач по сравнению законов распределения и некоторых их производных характеристик. В то же время, полное игнорирование параметрических моделей приводит к большой потере информации о форме распределения.

Компромиссное решение предложила робастная статистика, возникшая в середине XX века. Сам термин «робастный» («грубый, сильный, крепкий») ввёл Дж. Бокс в 1953 году, но систематическое развитие она получила с работы Дж. Тьюки, исследующей модели загрязнения распределений. Полноценный теоретический подход к робастности в статистике был предложен Дж. П. Хьюбером в 1964 году, и получил широкую известность в 1981 году с выходом книги, посвящённой минимаксным методам поиска оценок, оптимальных в окрестности предполагаемого распределения. Альтернативный подход через функции влияния был предложен Ф. Хампелем в диссертации в 1968 году и рассмотрен

более подробно в книге 1986 года.

Основная идея робастности — это построение статистических процедур, устойчивых к возможным отклонениям от принятых вероятностных моделей распределений данных. Подходы Хьюбера и Хампеля отличаются различным выбором используемых мер устойчивости рассматриваемых робастных оценок, но, несмотря на эти различия, как правило, «хорошие» робастные оценки в смысле Хьюбера практически близки «хорошим» робастным оценками в смысле Хампеля, а иногда они и совпадают.

В нашей стране теория устойчивых статистических методов также активно развивается, одной из первых вех была вышедшая в 1931 году статья А. Н. Колмогорова «Метод медианы в теории ошибок», подробно рассматривающая преимущества медианы перед средним арифметическим в том случае, если «гипотеза нормального распределения не удовлетворяет фактам». Изучение и дальнейшая разработка вероятностно-статистических методов, их внедрение в научную, инженерную и медицинскую практику было одной из задач, поставленных перед Межфакультетской («колмогоровской») лабораторией статистических методов при кафедре теории вероятностей МГУ. Похожие на хьюберовские оценки параметров многомерных распределений, при которых занижается вклад выдающихся значений на периферии, рассматривал Л. Д. Мешалкин, предложивший в 1970 году экспоненциальное взвешивание наблюдений. Этот подход и связанные с ним результаты развил А. М. Шурыгин, исследуя применимость методов классической статистики и теории вероятностей к решению реальных задач геофизики.

Значительный вклад в теорию робастного (учитывающего фактор неопределённости) управления внёс Я. З. Цыпкин, с 1956 года и до своей кончины в 1997 году заведующий лабораторией №7 Института автоматики и телемеханики (в настоящее время — лаборатория адаптивных и робастных систем им. Я. З. Цыпкина Института проблем управления РАН). За цикл работ «Робастность в задачах оценивания, оптимизации и устойчивости» Я. З. Цыпкин и Б. Т. Поляк были награждены премией А. А. Андронова.

Научная школа непараметрической и робастной статистики была создана в Томске Ф. П. Тарасенко, первоочередное внимание в которой уделялось непараметрическим методам. Характерной особенностью томской группы статистиков является последовательное использование функционального представления статистических процедур, при которой статистики порождаются путем подстановки различных оценок распределений в характеристический функционал рассматриваемой задачи. Много усилий на обобщение и развитие именно робастных статистических процедур направил В. П. Шуленин, в 1993 году

опубликовавший монографию по робастной статистике, и совсем недавно, в 2012 году, выпустивший учебное пособие в трёх томах, посвящённых отдельно достижениям в параметрической, непараметрической и робастной статистике.

Ю. С. Харин в связи с организацией кафедры теории вероятностей и математической статистики был приглашён в Минск, где впоследствии занял пост заведующего новой кафедрой математического моделирования и анализа данных Белорусского государственного университета и директора НИИ прикладных проблем математики и информатики БГУ. Тематика научных интересов основанной им кафедры связана с разработкой математических моделей, методов, алгоритмов и программных средств робастного распознавания и анализа стохастических данных для компьютерных систем защиты информации и информационных технологий.

В связи с развитием теории ошибок измерений, изучения случайных ошибок и грубых промахов, возникших в ходе эксперимента, наиболее полно исследованным оказалось робастное оценивание параметра положения распределений случайных величин. В чуть менее разработанной области робастного оценивания параметра масштаба, а тем более, коэффициента корреляции двух зависимых случайных величин остаётся ещё потенциал для исследования с точки зрения увеличения эффективности алгоритмов оценивания (уменьшения разброса значений вычисленных по выборкам оценок).

Внедрение и практическое использование предлагаемых новых робастных методов оценивания параметра масштаба и корреляционных характеристик данных предполагает разработку программно-алгоритмического комплекса, их реализующего.

Цель работы. Целью настоящей диссертационной работы является разработка комплекса новых методов, алгоритмов и программ робастного оценивания корреляционных характеристик данных, обладающих высокой устойчивостью к загрязнениям данных и другим отклонениям от предполагаемой параметрической модели при сохранении высокой асимптотической эффективности.

Задачи исследования.

1. Изучить различные робастные методы оценивания коэффициента корреляции и корреляционных матриц, включая оценки, основанные на оценках масштаба.

2. Исследовать поведение асимптотического смещения и дисперсии оценок коэффициента корреляции, определённых через оценки масштаба, на семействе распределений

в независимых компонентах (которое включает в себя двумерное нормальное распределение).

3. Построить оценки максимального правдоподобия для коэффициента корреляции семейства распределений в независимых компонентах.

4. Предложить быструю высокоэффективную оценку параметра масштаба для использования при оценивании коэффициента корреляции и связанных с ним величин.

5. Исследовать применение предложенных оценок параметра масштаба и коэффициента корреляции в других статистических методах (многомерном статистическом анализе, теории временных рядов).

Научная новизна. В диссертационной работе получены и обоснованы следующие новые результаты, **выносимые на защиту**:

1. Разработаны робастные методы и алгоритмы оценивания корреляционных характеристик данных на основе новых высокоэффективных и быстрых робастных оценок масштаба.

2. Предложено параметрическое семейство новых робастных M-оценок масштаба с абсолютной асимптотической эффективностью на нормальном распределении от 80 до 95%, максимально возможной пороговой точкой 50% и асимптотически линейным ростом времени работы алгоритма $O(n)$ при увеличении размера выборки n.

3. Исследовано применение оценок масштаба для оценивания коэффициента корреляции и корреляционных матриц многомерных распределений из класса распределений с независимыми компонентами, и доказана прямо пропорциональная зависимость асимптотического смещения и дисперсии оценки коэффициента корреляции от асимптотической дисперсии используемой оценки масштаба.

4. Получены оценки максимального правдоподобия и M-оценки для коэффициента корреляции семейства распределений в независимых компонентах, уравнение правдоподобия выражено через оценочную функцию параметра масштаба, и доказана прямо пропорциональная зависимость асимптотической дисперсии оценки коэффициента корреляции от асимптотической дисперсии используемой оценки масштаба.

5. Предложено теоретическое и практическое обоснование необходимого числа повторений эксперимента (≈ 50000) в исследованиях оценок методом Монте-Карло.

6. Разработаны алгоритмы и комплекс программ и библиотек функций, реализующих предлагаемые оценки параметров масштаба, корреляции и корреляционных матриц случайных распределений, а также предоставляющих экспериментальную среду для проведения испытаний Монте-Карло.

Теоретическая и практическая значимость. Доказанная зависимость между асимптотическими дисперсиями оценок коэффициента корреляции и параметра масштаба, на которых они основаны, позволяет повышать статистическую эффективность корреляционных алгоритмов оценивания за счет использования более эффективных оценок масштаба.

Полученные робастные, высокоэффективные оценки параметра масштаба, коэффициента корреляции и корреляционных матриц помогают с большей точностью и устойчивостью к помехам и ошибкам измерений проводить статистический анализ данных.

Методология и методы исследования. Для решения поставленных задач использовался аппарат теории алгоритмов, линейной алгебры, вычислительной математики, математического анализа, теории вероятностей, параметрической и робастной математической статистики.

Степень достоверности результатов. Приведённые в диссертации теоретические результаты подтверждаются как аналитическими исследованиями, так и прямым имитационным моделированием Монте-Карло для различных, в том числе и больших, размеров выборок.

Внедрение результатов исследования. Подготовлена к публикации в свободном доступе библиотека функций для широко используемой бесплатной программной среды статистических вычислений и обработки данных *R Project*, содержащая предложенные в данной работе процедуры оценивания масштаба распределений, коэффициентов корреляции и корреляционных матриц многомерных случайных величин, автоковариационных функций и коэффициентов авторегрессии случайных временных рядов [13].

Апробация работы. Основные положения и результаты диссертации докладывались и обсуждались на международных конференциях: «International Conference on Robust

Statistics» (Чехия, Прага, 2010 год; Испания, Вальядолид, 2011 год), «International Conference on Computer Data Analysis and Modeling» (Беларусь, Минск, 2010 и 2013 год), «IEEE International Conference on Acoustics, Speech and Signal Processing» (Канада, Ванкувер, 2013 год). По материалам диссертации опубликовано двенадцать печатных работ и одна работа в электронном виде, из них две — в ведущих российских изданиях, включённых в перечень ВАК, и две работы опубликованы в международных профильных реферируемых журналах.

Структура и объем работы. Диссертация состоит из введения, трёх глав и заключения, содержит 157 страниц основного текста, включая 18 рисунков и 15 таблиц. Приложение содержит распечатки программных реализаций основных алгоритмов. В списке литературы 128 наименований.

СОДЕРЖАНИЕ РАБОТЫ

Во введении обоснована актуальность диссертационной работы, сформулирована цель и аргументирована научная новизна исследований, показана практическая значимость полученных результатов, представлены выносимые на защиту научные положения.

В первой главе рассмотрена задача оценивания параметра масштаба случайных распределений, введены основные определения, проведён обзор классических и робастных оценок параметра масштаба, методов их построения. Также предложены новые высокоэффективные робастные оценки MQ_n и FQ_n, изучены их характеристики и проведено имитационное моделирование методом Монте-Карло для потверждения теоретических выводов на практике. Эти оценки анонсированы на международной конференции по робастной статистике [5] и опубликованы в журнале [1] из списка рекомендованных ВАК.

В качестве робастной оценки масштаба случайных распределений обычно используется медиана абсолютных отклонений MAD_n, которая достигает максимального значения пороговой точки и имеет достаточно простой, понятный и быстрый алгоритм вычисления. Основным её недостатком является низкая эффективность (т.е. высокая дисперсия) на нормальном распределении, всего 37% по сравнению с классическим среднеквадратичным отклонением SD_n. Рауссеу и Крукс предложили альтернативную оценку с большей эффективностью и такой же высокой пороговой точкой, квартиль абсолютных разностей

Q_n, которая из-за своих характеристик приобрела большую известность в современной робастной статистике.

Одним из препятствий к использованию этой оценки является высокая (по сравнению с другими оценками) асимптотическая сложность алгоритма её вычисления. Нахождение порядковой статистики среди примерно $n(n-1)/2$ пар элементов выборки в общем случае требует $O(n^2)$ времени и столько же памяти. Несмотря на то, что для случая попарных разностей авторы предложили более эффективный алгоритм, требующий только $O(n \log n)$ времени и $O(n)$ памяти, на больших выборках разница становится существенной.

В данной работе предлагаются новые оценки, основанные на функции влияния Q_n и наследующие от неё локальные робастные свойства и асимптотическую эффективность.

Определение 1.1. *Оценкой* MQ_n, построенной для выборки $(x_1, \ldots, x_n)$ из распределения с плотностью $f(x)$, будем называть параметрическое семейство M-оценок масштаба, являющихся решением уравнения

$$\sum_{i=1}^{n} \chi_\alpha(x_i / MQ_n) = 0, \tag{1.1}$$

где оценочная функция χ_α задаётся формулой

$$\chi_\alpha(x) = c_\alpha - 2f(x) - \frac{1}{3}\alpha^2 f''(x), \qquad c_\alpha : \int\limits_{-\infty}^{\infty} \chi_\alpha(x) f(x)\, dx = 0. \tag{1.2}$$

Для нормального распределения с плотностью $\varphi(x) = (2\pi)^{-1/2} \exp(-x^2/2)$

$$\chi_\alpha(x) = c_\alpha - \frac{1}{3}(6 + \alpha^2(x^2 - 1))\varphi(x), \qquad c_\alpha = \frac{12 - \alpha^2}{12\sqrt{\pi}}, \tag{1.3}$$

в важном частном случае при $\alpha = 0$ выражение принимает вид

$$\chi_0(x) = \frac{1}{\sqrt{\pi}} - 2\varphi(x). \tag{1.4}$$

Вычисление оценки MQ_n как решения неявного уравнения (1.1) затруднительно, но возможно применение итеративных схем. В частности, можно ограничиться первой итерацией метода Ньютона, получив, так называемую, одношаговую M-оценку.

Определение 1.2. *Оценкой* FQ_n (Fast Q_n), построенной для выборки из нормального распределения, будем называть параметрическое семейство одношаговых M-оценок масштаба, задаваемых формулой

$$FQ_n = MAD_n \cdot \left(1 - \frac{(6 - \alpha^2)U_0 + \alpha^2 U_2 - 3c_\alpha n}{3(2 - \alpha^2)U_2 + \alpha^2 U_4}\right) \tag{1.5}$$

где

$$U_k = \sum_{i=1}^{n} u_i^k \varphi(u_i), \quad u_i = x_i / \, MAD_n \, .$$

В важном частном случае при $\alpha = 0$ оценка принимает вид

$$FQ_n = MAD_n \cdot \left(1 - \frac{U_0 - n/(2\sqrt{\pi})}{U_2}\right) \tag{1.6}$$

Теорема 1.1. *Пороговая точка оценки MQ_n для значений $\alpha \in [\,0, \sqrt{2}\,]$ задаётся формулой*

$$\varepsilon^* = \frac{12(\sqrt{2}-1) - \alpha^2(2\sqrt{2}-1)}{(6-\alpha^2)2\sqrt{2}}, \tag{1.7}$$

достигая максимума $\varepsilon^* = 1 - 1/\sqrt{2} \approx 0.2929$ (т.е. чуть меньше 30%) при $\alpha = 0$ и уменьшаясь при увеличении параметра α.

Пороговая точка оценки FQ_n не зависит от α и составляет 50%.

Теорема 1.2. *Оценки MQ_n и FQ_n при $\alpha \in [\,0, \sqrt{2}\,]$ на нормальном модельном распределении являются B-робастными. Функции влияния оценок ограничены и имеют вид*

$$\mathrm{IF}(x; MQ, \Phi) = \mathrm{IF}(x; FQ, \Phi) = \frac{2(12-\alpha^2) - 8\sqrt{\pi}\,(6 + \alpha^2(x^2-1))\,\varphi(x)}{3(4-\alpha^2)}. \tag{1.8}$$

Теорема 1.3. *Асимптотическая дисперсия оценок MQ_n и FQ_n на нормальном распределении задаётся формулой*

$$V(MQ, \Phi) = V(FQ, \Phi) = \left(\frac{4-\alpha^2}{8}\right)^{-2} \left(\frac{54 - 12\alpha^2 + \alpha^4}{27\sqrt{3}} - \left(\frac{12-\alpha^2}{12}\right)^2\right), \tag{1.9}$$

Асимптотическая эффективность при $\alpha = 0$ составляет 80.8%, возрастая при увеличении α и достигая своего максимума в 95.9% при $\alpha = 1.4028$.

Теорема 1.4. *Вариант оценки MQ_n, построенный исходя из формулы (1.2) для распределения Коши с плотностью $f(x) = (1/\pi)/(1+x^2)$, является оценкой максимального правдоподобия для данного распределения с максимально возможной асимптотической эффективностью (100%) и пороговой точкой (50%).*

Имитационное моделирование методом Монте-Карло подтверждает хорошие характеристики предложенной оценки. В случае отсутствия загрязнения на нормальном распределении FQ_n и по смещению, и по дисперсии (за исключением очень больших выборок) ведёт себя лучше, чем Q_n. Это особенно заметно на малых выборках ($n = 20$), где смещение сравнимо с лучшим результатом, на порядок превосходя оценку Q_n, а дисперсия

становится наименьшей, показывая наилучшую эффективность среди рассматриваемых робастных оценок.

Проведённые измерения времени вычисления оценок показывают преимущество линейных алгоритмов, в том числе и FQ_n, над более медленным алгоритмом вычисления оценки Q_n, основанной на попарных разностях наблюдений. При размере выборки $n = 1000$ время работы алгоритма Q_n превышает время вычисления оценки FQ_n более чем в 9 раз.

На нормальном распределении с 10%-ным загрязнением в модели больших ошибок Тьюки предлагаемая оценка FQ_n занимает второе место по смещению после наиболее B-робастной M-оценки MAD_n, и имеет лучший результат по дисперсии как для малых, так и для больших выборок. При увеличении доли загрязнения оценка Q_n лишь незначительно обходит её по дисперсии. При подмене нормального распределения на распределение Коши, имеющее тяжёлые хвосты, оценка FQ_n также занимает второе место по смещению после MAD_n, немного уступая Q_n по дисперсии.

В целом, на рассмотренных моделях оценка FQ_n имеет хорошие характеристики как при наличии, так и при отсутствии загрязнения. Это позволяет рекомендовать её как

- более быструю альтернативу используемой в последние годы робастной оценке Q_n;
- более эффективное уточнение давно известной робастной оценки MAD_n.

Предложенную оценку можно использовать как непосредственно для оценивания масштаба симметричных распределений, так и в качестве базового алгоритма в других статистических процедурах (в задачах регрессии, корреляционного анализа [8], и т.п.).

Во второй главе рассмотрена задача оценивания коэффициента корреляции и корреляционных матриц распределений случайных величин, проведён обзор основных классических и робастных оценок. Отдельно изучен класс распределений в независимых компонентах и для него на базе оценок масштаба построены оценки коэффициента корреляции, изучены их характеристики и проведено имитационное моделирование методом Монте-Карло для потверждения теоретических выводов на практике. Также в работе в качестве вспомогательного шага оценивания исследована проблема исправления (приведения к положительно определённому виду) оценок корреляционных матриц, полученных поэлементно при помощи попарных корреляций. Результаты второй главы анонсированны на конференциях [4, 6] и опубликованы в международных рецензируемых журналах [8, 9].

Рассмотрим семейство двумерных распределений вероятностей, определяемое факторизуемой плотностью

$$f(x,y) = \frac{1}{a} g\Big(\frac{u}{a}\Big) \cdot \frac{1}{b} g\Big(\frac{v}{b}\Big), \tag{2.1}$$

где u, v — главные компоненты, которые задаются поворотом системы координат, ортогональным преобразованием

$$u = (x+y)/\sqrt{2}, \quad v = (x-y)/\sqrt{2},$$

параметры a, b играют роль параметров масштаба для некоторой симметричной плотности вероятности $g(t)$.

Простые преобразования показывают, что коэффициент корреляции задаётся формулой

$$\rho = \frac{a^2 - b^2}{a^2 + b^2} = \frac{\mathsf{D}(U) - \mathsf{D}(V)}{\mathsf{D}(U) + \mathsf{D}(V)}. \tag{2.2}$$

Заметим, что если в качестве базовой плотности взять плотность нормального закона распределения $g(t) = \varphi(t) = (2\pi)^{-1/2}\exp(-t^2/2)$, то формула (2.1) как частный случай даст плотность двумерного нормального распределения

$$f(x,y) = \frac{1}{2\pi\sigma^2\sqrt{1-\rho^2}} \exp\left\{-\frac{x^2 - 2\rho xy + y^2}{2\sigma^2(1-\rho^2)}\right\}.$$

Выражение коэффициента корреляции (2.2) приводит к естественной оценке для данного класса

$$\widehat{\rho}_n = \frac{\widehat{a}_n^2 - \widehat{b}_n^2}{\widehat{a}_n^2 + \widehat{b}_n^2} = \frac{S_n^2(\mathbf{u}) - S_n^2(\mathbf{v})}{S_n^2(\mathbf{u}) + S_n^2(\mathbf{v})}, \tag{2.3}$$

где $\widehat{a}_n = S_n(\mathbf{u})$, $\widehat{b}_n = S_n(\mathbf{v})$ — некоторые оценки параметров масштаба a и b по трансформированным выборкам

$$\mathbf{u} = (u_1, \ldots, u_n), \quad u_i = (x_i + y_i)/\sqrt{2},$$
$$\mathbf{v} = (v_1, \ldots, v_n), \quad v_i = (x_i - y_i)/\sqrt{2}.$$

Теорема 2.1. *Пусть $g(t)$ — непрерывная, симметричная плотность распределения вероятностей с конечной дисперсией $\mathsf{D}_g = 1$, а $S_n^2(\mathbf{t})$ — состоятельная, асимптотически несмещённая оценка дисперсии $g(t)$ по выборке $\mathbf{t} = (t_1, \ldots, t_n)$ с математическим ожиданием $\vartheta_n = \mathsf{M}_{g^n}[S_n^2(\mathbf{t})]$ и конечной дисперсией $\delta_n^2 = \mathsf{D}_{g^n}[S_n^2(\mathbf{t})]$, имеющей порядок $1/n$. Тогда оценка $\widehat{\rho}_n$, вычисляемая по формуле (2.3), является состоятельной, асимптотически несмещённой оценкой коэффициента корреляции для закона распределения (2.1)*

со смещением и дисперсией, определяемыми формулами

$$\mathsf{M}(\widehat{\rho}_n) - \rho = -\frac{\rho(1-\rho^2)}{2}\left(\frac{\delta_n^2}{\vartheta_n^2}\right) + o\left(\frac{1}{n}\right), \tag{2.4}$$

$$\mathsf{D}(\widehat{\rho}_n) = \frac{(1-\rho^2)^2}{2}\left(\frac{\delta_n^2}{\vartheta_n^2}\right) + o\left(\frac{1}{n}\right), \tag{2.5}$$

Теорема 2.2. *Пусть $g(t)$ — непрерывная, симметричная плотность распределения вероятностей с конечной дисперсией $\mathsf{D}_g = 1$, а $S_n(\mathbf{t})$ — состоятельная, асимптотически несмещённая, асимптотически нормальная оценка масштаба $g(t)$ по выборке $\mathbf{t} = (t_1, \ldots, t_n)$, которую можно представить в виде функционала S от эмпирического закона распределения $S_n(\mathbf{t}) = S(G_n)$ с существующей функцией влияния $\mathrm{IF}(t, S, G)$ и асимптотической дисперсией $V(S, G)$. Тогда оценка $\widehat{\rho}_n$, вычисляемая по формуле (2.3), является состоятельной, асимптотически несмещённой оценкой коэффициента корреляции для закона распределения (2.1) со смещением и дисперсией, пропорциональными $V(S, G)$, и определяемыми формулами*

$$\mathsf{M}(\widehat{\rho}_n) - \rho = -\frac{2\rho(1-\rho^2)}{n}\,V(S,G) + o\left(\frac{1}{n}\right), \tag{2.6}$$

$$\mathsf{D}(\widehat{\rho}_n) = \frac{2(1-\rho^2)^2}{n}\,V(S,G) + o\left(\frac{1}{n}\right), \tag{2.7}$$

На данном классе помимо естественной оценки коэффициента корреляции (2.3) можно рассмотреть и уравнение правдоподобия для ρ. Оно может быть записано в привычном для M-оценок виде

$$\sum_{i=1}^{n} \psi(u_i, v_i; \rho) = 0, \tag{2.8}$$

где оценочная функция $\psi = \partial \ln f / \partial \rho$ для коэффициента корреляции связана с оценочной функцией χ параметра масштаба базового распределения:

$$\psi(u, v; \rho) = \frac{1}{2}\left[\frac{1}{1+\rho}\,\chi\left(\frac{u}{\sqrt{1+\rho}}\right) - \frac{1}{1-\rho}\,\chi\left(\frac{v}{\sqrt{1-\rho}}\right)\right]. \tag{2.9}$$

Подобная форма записи позволяет легко перейти к M-оценкам коэффициента корреляции (оценкам *типа* максимального правдоподобия) путём выбора произвольной подходящей функции χ.

Вычисление M-оценки возможно при помощи итерационного алгоритма

$$\widehat{\rho}_{k+1} = \widehat{\rho}_k + \frac{(1-\widehat{\rho}_k^2)^2}{n(1+\widehat{\rho}_k^2)B_\chi} \cdot \sum_{i=1}^{n}\left[\frac{1}{1+\widehat{\rho}_k}\,\chi\left(\frac{u_i}{\sqrt{1+\widehat{\rho}_k}}\right) - \frac{1}{1-\widehat{\rho}_k}\,\chi\left(\frac{v_i}{\sqrt{1-\widehat{\rho}_k}}\right)\right], \tag{2.10}$$

где

$$B_\chi = \int_{-\infty}^{\infty} t\chi'(t)g(t)\,dt.$$

Чтобы получить робастные оценки корреляции в качестве базовых разумно брать робастные высокоэффективные оценки масштаба, такие как предложенная FQ_n.

Теорема 2.3. *Асимптотическая дисперсия M-оценки коэффициента корреляции ρ семейства двумерных распределений в независимых компонентах, задаваемой формулами (2.8) и (2.9), пропорциональна асимптотической дисперсии $V(\chi, G)$ M-оценки параметра масштаба, лежащей в её основе, и определяется формулой*

$$V(\psi, F) = \frac{2(1-\rho^2)^2}{(1+\rho^2)} V(\chi, G). \tag{2.11}$$

Корреляционная матрица системы p случайных величин (в том числе и выборочная) является матрицей из коэффициентов попарных корреляций, поэтому очевидный подход к её робастному оцениванию заключается в замене выборочного коэффициента корреляции Пирсона на его робастные аналоги. Недостатком такого подхода является невозможность обеспечить необходимую положительную определённость матрицы, составленной из произвольных оценок. Тем не менее, состоятельные оценки коэффициента корреляции в пределе дают матрицы, удовлетворяющие всем необходимым условиям, т.е. оценка $\widehat{R}$ лежит «близко» к искомой матрице R и может быть скорректирована должным образом.

Этим недостатком не обладают алгоритмы оценивания корреляционных или ковариационных матриц в целом, такие как *эллипсоид минимального объёма* (MVE) или *минимальный ковариационный определитель* (MCD). Корректное вычисление этих оценок, основанных на переборе всех возможных вариантов, требует больших вычислительных затрат, поэтому реальное их использование оказалось возможным только благодаря приближённым алгоритмам, дающим адекватные результаты. К сожалению, даже использующийся в настоящее время робастный алгоритм оценивания *FastMCD* подразумевает большое количество вычислительно непростых итераций, поэтому попарные оценки корреляционных матриц всё ещё представляют интерес.

Среди всех возможных псевдокорреляционных матриц размерности 3×3, т.е. симметричных матриц с единичной диагональю и элементами, ограниченными по модулю единицей, доля положительно полуопределённых (ППО) корреляционных матриц составляет 61.7%. Как было показано экспериментально, эта доля быстро уменьшается с ростом размерности, при $p = 5$ доля не-ППО матриц уже возрастает до 97.8%. Разумеется, не все они могут быть получены в результате оценивания. Так, оценка коэффициента корреляции (2.3), основанная на предлагаемой оценке масштаба FQ_n, при $p = 5$ для умеренных выборок ($n = 100$) порождает 8% матриц, требующих коррекции. Эта доля увеличивается

с ростом размерности и с уменьшением количества элементов в выборке (т.е., с ростом неопределённости). Алгоритмы коррекции включают в себя как методы прямой правки элементов или собственных чисел матрицы, так и решение задачи поиска ближайшей корреляционной матрицы.

В имитационном моделировании методом Монте-Карло были рассмотрены разные оценки коэффициента корреляции двумерного нормального закона распределения, включая предложенные в данной работе: оценку через независимые компоненты r_{FQ} и M-оценку $r_{M\cdot FQ}$, вычисленные по формулам (2.3) и (2.10) соответственно, с использованием введённой оценки масштаба FQ_n. Эксперимент подтверждает хорошие характеристики предложенных оценок.

В случае отсутствия загрязнения на двумерном нормальном распределении оценки показывают умеренное смещение и не самую лучшую, но ожидаемо высокую эффективность, обусловленную низкой асимптотической дисперсией оценки FQ_n. При этом эффективность $r_{M\cdot FQ}$ по отношению к выборочному коэффициенту корреляции превышает 100% для существенных корреляций ($\rho \geqslant 0.5$).

При сферическом засорении в модели больших ошибок Тьюки оценка r_{FQ} показывает наибольшую эффективность среди всех рассматриваемых оценок, с точки зрения смещения уступая оценке, основанной на MAD_n, и двумерному варианту алгоритма *FastMCD*. Тем не менее, последняя оценка имеет слишком низкую эффективность, т.е. её суммарная среднеквадратичная ошибка оказывается велика. Кроме того, как подтвердил эксперимент, оценка r_{MCD} легко подвержена внутреннему загрязнению, теряя все свои преимущества.

В целом, на рассмотренных моделях оценка r_{FQ} имеет хорошие характеристики как при наличии, так и при отсутствии загрязнения. Это позволяет рекомендовать её как более быструю альтернативу используемой в последние годы робастной оценке по алгоритму *FastMCD* на больших выборках и как более эффективную — на выборках небольшого размера.

В третьей главе рассмотрены возможные приложения полученных робастных высокоэффективных оценок параметра масштаба и коэффициента корреляции распределений в теории временных рядов и дескриптивной статистике.

Предложены робастные методы оценивания автоковариационной функции и спектральной плотности мощности стационарных временных рядов, коэффициентов процесса авторегрессии. Предварительные результаты робастного оценивания спектра по методу

Юла-Уолкера, основанному на робастной оценке масштаба FQ_n, показывают устойчивость предложенных оценок к редкой импульсной помехе высокой амплитуды, но для окончательных выводов требуется серьезный сравнительный анализ существующих робастных методов оценивания спектров.

Имитационное моделирование Монте-Карло показывает, что применение новых оценок в разведочном анализе при построении робастных одномерных и двумерных боксплотов приводит к статистически более эффективным результатам. Отбраковка данных по критерию, основанному на предложенной оценке масштаба FQ_n, превосходит результаты, полученные по боксплоту Тьюки, и значительно превосходит классический тест Граббса при различных видах и долях засорения.

Результаты третьей главы анонсированны на конференциях [2, 7, 10–12] и опубликованы в журнале [3], входящем в список ведущих рецензируемых журналов, рекомендованных ВАК.

В заключении сформулированы основные результаты работы, даны рекомендации и перспективы дальнейшей разработки темы. **В приложении А** приведены основные распечатки программных реализаций предложенных методов и алгоритмов.

ЗАКЛЮЧЕНИЕ

В ходе данного исследования были выполнены все поставленные задачи. Предложенные методы и алгоритмы оценивания параметра масштаба и коэффициента корреляции случайных распределений по сравнению с робастными оценками, широко применяющимися на практике в настоящее время, обладают конкурентными преимуществами: высокой статистической эффективностью и скоростью работы при сохранении робастных свойств.

Рассмотренные методы дают хорошие результаты непосредственно как оценки соответствующих характеристик распределений, так и в других задачах математической статистики: при оценивании корреляционных и ковариационных матриц многомерных распределений, автоковариационных функций и спектров плотности мощности стационарных временных рядов, при первичной отбраковке данных в дескриптивной статистике.

Созданный программный комплекс подготовлен к открытой публикации в интернете [13] в виде библиотеки процедур для среды статистических вычислений *R Project*, широко использующейся для статистического моделирования и робастного анализа данных.

СПИСОК РАБОТ АВТОРА ПО ТЕМЕ ДИССЕРТАЦИИ

1. *Смирнов, П. О.* Приближение оценки Q_n параметра масштаба с помощью быстрых M-оценок [Текст] / П. О. Смирнов, Г. Л. Шевляков // *Вестник Сибирского государственного аэрокосмического университета имени академика М. Ф. Решетнева.* — 2010. — Т. 31, № 5. — С. 83–85.
2. *Смирнов, П. О.* Визуализация данных двумерными FQ_n-боксплотами [Текст] / К. Андреа, Г. М. Лаврентьева, П. О. Смирнов, Г. Л. Шевляков // Высокие технологии, фундаментальные исследования, экономика. — Т. 1. — Санкт-Петербург, Россия : Изд-во Политехн. ун-та, 2011. — С. 59–66.
3. *Смирнов, П. О.* Двумерный боксплот на основе высокоэффективных робастных оценок масштаба и корреляции [Текст] / К. Андреа, П. Смирнов, Г. Шевляков // *Вестник Томского государственного университета. Управление. Вычислительная техника и информатика.* — 2013. — Т. 22, № 1. — С. 25–31.
4. *Smirnov, P. O.* Highly efficient robust estimators of a correlation coefficient for bivariate independent component distributions [Text] / G. L. Shevlyakov, P. O. Smirnov // Book of Abstracts: International Conference on Robust Statistics (ICORS 2010). — Prague, Czech Republic : Charles University, 2010. — P. 93–94.
5. *Smirnov, P. O.* On approximation of the Q_n-estimate of scale by fast M-estimates [Text] / P. O. Smirnov, G. L. Shevlyakov // Book of Abstracts: International Conference on Robust Statistics (ICORS 2010). — Prague, Czech Republic : Charles University, 2010. — P. 94–95.
6. *Smirnov, P. O.* Robust estimation of a correlation coefficient: An attempt of survey [Text] / G. L. Shevlyakov, P. O. Smirnov // Proceedings of the 9th International Conference on Computer Data Analysis and Modeling. — Vol. 1. — Minsk, Belarus : Publishing center of BSU, 2010. — P. 108–115.
7. *Smirnov, P. O.* Fast low-complexity bivariate boxplots based on highly efficient and robust estimates of dispersion and correlation [Text] / G. Shevlyakov, K. Andrea, G. Lavrentyeva, P. Smirnov // Book of Abstracts: International Conference on Robust Statistics (ICORS 2011). — Valladolid, Spain : University of Valladolid, 2011. — P. 72.
8. *Smirnov, P. O.* Robust estimation of the correlation coefficient: An attempt of survey [Text] / G. L. Shevlyakov, P. O. Smirnov // *Austrian Journal of Statistics.* — 2011. — Vol. 40, no. 1&2. — P. 147–156.

9. *Smirnov, P. O.* Asymptotically minimax bias estimation of the correlation coefficient for bivariate independent component distributions [Text] / G. L. Shevlyakov, P. O. Smirnov, V. I. Shin, K. Kim // *Journal of Multivariate Analysis*. — 2012. — Vol. 111. — P. 59–65.
10. *Smirnov, P. O.* Detection of outliers with boxplots [Text] / K. Andrea, G. L. Shevlyakov, P. O. Smirnov // Proceedings of the 11th International Conference on Computer Data Analysis and Modeling. — Minsk, Belarus : Publishing center of BSU, 2013. — P. 141–144.
11. *Smirnov, P. O.* Robust versions of the Tukey boxplot with their application to detection of outliers [Text] / Georgy L. Shevlyakov, Kliton Andrea, Lakshminarayan Choudur [et al.] // IEEE International Conference on Acoustics, Speech, and Signal Processing. — Vancouver, Canada : IEEE, 2013. — P. 6506–6510.
12. *Smirnov, P. O.* Some remarks on robust estimation of power spectra [Text] / G. L. Shevlyakov, N. S. Lyubomishchenko, P. O. Smirnov // Proceedings of the 11th International Conference on Computer Data Analysis and Modeling. — Minsk, Belarus : Publishing center of BSU, 2013. — P. 97–104.
13. *Smirnov, P. O.* `robcor`: Robust correlations. R package version 0.1-5 [Electronic resource]. — Vienna, Austria : The Comprehensive R Archive Network, 2013. — URL: http://CRAN.R-project.org/package=robcor (online; accessed: 06.12.2013).