На правах рукописи

Браништи Владислав Владимирович

МЕТОДЫ И АЛГОРИТМЫ НАСТРОЙКИ ПРОЕКЦИОННОЙ ОЦЕНКИ ПЛОТНОСТИ ВЕРОЯТНОСТИ СЛУЧАЙНОГО ВЕКТОРА В УСЛОВИЯХ МАЛЫХ ВЫБОРОК

Специальность 05.13.17 – Теоретические основы информатики

АВТОРЕФЕРАТ
диссертации на соискание учёной степени
кандидата физико-математических наук

Красноярск 2019

Работа выполнена в федеральном государственном бюджетном образовательном учреждении высшего образования «Сибирский государственный университет науки и технологий имени академика М. Ф. Решетнёва», г. Красноярск.

Научный руководитель: доктор физико-математических наук, доцент, **Сафонов Константин Владимирович**

Официальные оппоненты: **Кошкин Геннадий Михайлович**, доктор физико-математических наук, профессор, Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский Томский государственный университет», кафедра теоретической кибернетики, профессор

Дронов Сергей Вадимович, кандидат физико-математических наук, доцент, Федеральное государственное бюджетное образовательное учреждение высшего образования «Алтайский государственный университет», кафедра математического анализа, доцент

Ведущая организация: Федеральное государственное бюджетное образовательное учреждение высшего образования «Сибирский государственный университет телекоммуникаций и информатики».

Защита состоится «25» июня 2019 г. в 14.00 часов на заседании диссертационного совета Д 212.099.22, созданного на базе Сибирского федерального университета по адресу: 660074, г. Красноярск, ул. ак. Киренского, 26, аудитория УЛК 112.

С диссертацией можно ознакомиться в библиотеке и на сайте Сибирского федерального университета по адресу http://www.sfu-kras.ru/.

Автореферат разослан «____» мая 2019 г.

Учёный секретарь диссертационного совета  Покидышева Людмила Ивановна

Общая характеристика работы

Актуальность темы и степень её разработанности. Разработка и исследование моделей и алгоритмов анализа данных, обнаружения закономерностей в данных в условиях неопределённости практически всегда предполагает оценивание функции распределения либо плотности вероятности соответствующих величин. Задача оценивания плотности вероятности случайного вектора возникает при разработке методов распознавания образов, фильтрации, распознавания и синтеза изображений.

Имеющиеся в настоящее время методы оценивания функции плотности вероятности можно разделить на параметрические и непараметрические. Параметрические методы используются в случае, когда известна структура закона распределения с точностью до параметров, и задача сводится к построению статистических оценок этих параметров, удовлетворяющих заданным условиям (состоятельность, несмещённость и др.). К числу наиболее разработанных параметрических методов относятся метод моментов, метод максимального правдоподобия, метод минимума χ^2. Однако часто в практических задачах возникают ситуации, когда структура закона распределения неизвестна, т.е. ситуации *непараметрической неопределённости*. При этом априорная информация о функции плотности вероятности $f(x)$ носит более общий характер, например, $f(x)$ может предполагаться непрерывной на данном отрезке, имеющей n-ю производную, имеющей суммируемый квадрат и т.п. Использование параметрических методов при фактическом несовпадении структуры закона распределения приводит к неудовлетворительным результатам. В этом случае используются методы, получившие название непараметрических.

Исторически первой непараметрической оценкой функции плотности вероятности является гистограмма, исследованная К. Пирсоном в 1895 г. Во второй половине 20-го века интерес к непараметрическим методам значительно возрос, о чём свидетельствует ряд работ, посвящённых следующим оценкам: полиграмма (Ф. П. Тарасенко, Е. В. Черепанов, А. И. Рубан), оценка k ближайших соседей (Д. О. Лофтсгарден, К. П. Куэсенберри и др.), оценка Розенблатта – Парзена (В. А. Васильев, А. В. Добровидов, Г. М. Кошкин, А. В. Медведев и др.), проекционная оценка (Н. Н. Ченцов, Дж. С. Ватсон, А. А. Новосёлов, В. Н. Вапник и др.).

При использовании непараметрических методов представляет интерес исследование сходимости получаемых оценок к истинной функции плотности вероятности по заданной метрике, а также оценка скорости сходимости. В связи с этим возникает задача оптимальной настройки оценок функции плотности вероятности. Так, одной из первых формул для расчёта числа интервалов группирования одинаковой длины при построении гистограммы яв-

ляется формула Стэрджеса. В случае использования полиграммы или оценки k ближайших соседей подлежит настройке численный параметр, определяющий степень сглаженности полученной оценки.

При использовании проекционной оценки плотности вероятности случайного вектора $\mathbf{x} = (x_1, \ldots, x_k)$:

$$\hat{f}(\mathbf{x}) = \sum_{j=0}^{l} a_j \psi_j(\mathbf{x})$$

настройке подлежат как численные параметры l, a_1, …, a_l, так и ортогональная система функций $\{\psi_j\}$. При этом оптимального набора функций ψ_j для всех плотностей не существует, так как очевидно, что для данной функции плотности вероятности $f(\mathbf{x})$ оптимальным будет любая система, в которой $\psi_0(\mathbf{x}) \equiv \frac{1}{\|f\|} f(\mathbf{x})$. Тогда $l = 0$ и $a_0 = \|f\|$.

Аналогично, при использовании оценки Розенблатта – Парзена:

$$\hat{f}(x_1, \ldots, x_k) = \frac{1}{n} \sum_{i=1}^{n} \prod_{j=1}^{k} \frac{1}{h_j} \Phi_j \left(\frac{x_j - x_{ij}}{h_j} \right)$$

настройке подлежат как параметры h_1, …, h_k, так и «ядерные» функции Φ_1, …, Φ_k. Как и в случае проекционной оценки, несложно подобрать оптимальные параметры для данного закона распределения.

Задача настройки непараметрических оценок значительно усложняется при отсутствии информации о законе распределения. В большинстве работ, посвящённых этой проблеме, исследования преимущественно выполняются в предположении, что объём выборки $n \to \infty$. Так, асимптотические свойства проекционной оценки исследуются в работах Н. Н. Ченцова, А. А. Новосёлова, Дж. С. Ватсона и др. Для оценки Розенблатта – Парзена в работе В. А. Епанечникова для этого случая получено решение для формы ядра Φ в классе усечённых функций, дифференцируемых в заданном интервале. Однако в анализе данных задачу оценивания плотности часто приходится решать при малых n, например, при обработке биомедицинских данных и данных, касающихся производства и эксплуатации дорогостоящих технических систем. Исследования показали, что результаты, полученные при $n \to \infty$, могут оказаться неоптимальными для малых n.

В этих условиях представляет интерес исследование проекционной оценки, так как, в отличие от других видов непараметрических оценок, например, оценки Розенблатта – Парзена или оценка k ближайших соседей, проекционная оценка не содержит в себе всей выборки и допускает компактное аналитическое выражение. Это оказывается более удобным при теоретическом анализе, в приложениях, а также повышает быстродействие алгоритмов классификации и восстановления зависимостей.

Целью диссертационной работы является разработка эффективных методов и алгоритмов настройки непараметрических оценок в условиях малых выборок.

Поставленная цель достигается путём решения следующих **задач**:

а) провести сравнительный анализ известных методов настройки непараметрических оценок;

б) осуществить расширение области применимости проекционной оценки;

в) исследовать возможность применения метода моментов для настройки проекционной оценки и выполнить его обобщение;

г) разработать алгоритмы настройки коэффициентов и длины ряда проекционной оценки функции плотности вероятности случайного вектора, ориентированные на решение задач восстановления зависимостей, классификации и оценивания количества информации;

д) сравнить разработанные методы и алгоритмы с известными алгоритмами настройки проекционной оценки на малых выборках.

Соответствие диссертации паспорту специальности. Диссертационная работа соответствует области исследований специальности 05.13.17 – Теоретические основы информатики по п. 5 «Разработка и исследование моделей и алгоритмов анализа данных, обнаружения закономерностей в данных и их извлечениях, разработка и исследование методов и алгоритмов анализа текста, устной речи и изображений» и п. 7 «Разработка методов распознавания образов, фильтрации, распознавания и синтеза изображений, решающих правил. Моделирование формирования эмпирического знания».

Методы исследования. Основные результаты получены на основе методов теории вероятностей, математической статистики, функционального анализа и теории меры, а также матричного анализа. При численных расчётах функционалов качества получаемых оценок использован метод статистических испытаний.

Научная новизна:

1. Впервые использовано весовое расширение пространства L_2 при построении проекционной оценки для любых функций плотности вероятности, в том числе, с несуммируемым квадратом. Тем самым расширена область применения проекционных оценок, в частности, при решении задач обнаружения закономерностей в данных и распознавания образов.

2. Разработан новый метод настройки коэффициентов проекционной оценки функции плотности вероятности случайного вектора, являющийся обобщением метода моментов. Метод позволяет повысить эффективность проекционной оценки в условиях малых выборок.

3. Предложен новый метод оценивания длины ряда проекционной оцен-

ки, в которой коэффициенты настраиваются методом моментов или его обобщением.

4. Разработаны алгоритмы настройки коэффициентов и длины ряда проекционной оценки функции плотности вероятности случайного вектора, которые ориентированы на решение задач восстановления зависимостей, классификации и оценивания количества информации. Предложенные алгоритмы являются более результативными для проекционной оценки в условиях малых выборок, чем алгоритмы, реализующие традиционный подход.

Теоретическая и практическая значимость работы. Работа носит теоретический характер. Результаты могут быть использованы при решении задач восстановления зависимостей, классификации и оценивания количества информации для построения проекционных оценок функции плотности вероятности.

Положения, выносимые на защиту диссертационной работы.

1. Доказательство сходимости проекционной оценки в весовом пространстве $L_{2,w}(\mathbb{R}^k)$ к функции плотности вероятности для любого закона распределения непрерывного случайного вектора при подходящей весовой функции w.

2. Метод настройки коэффициентов проекционной оценки функции плотности вероятности случайного вектора, представляющий собой обобщение метода моментов.

3. Метод оценивания длины ряда проекционной оценки функции плотности вероятности случайного вектора.

4. Алгоритмы настройки коэффициентов и длины ряда проекционной оценки функции плотности вероятности случайного вектора, предназначенные для решения прикладных задач на малых выборках.

Достоверность результатов работы подтверждается математическими доказательствами основных положений, а также численными экспериментами.

Апробация результатов работы. Результаты диссертационной работы докладывались автором на следующих конференциях: Всероссийской конференции «Наука. Технологии. Инновации» (Новосибирск, 2007 г.); Всероссийской конференции «Молодёжь и наука» (Красноярск, 2007, 2014 гг.); Всероссийской конференции «Актуальные проблемы авиации и космонавтики» (Красноярск, 2014, 2015, 2017 гг.); Всероссийской конференции «Наука и АСУ – 2014» (Москва, 2014 г.); Международной конференции «Решетнёвские чтения» (Красноярск, 2014, 2016 гг.).

Результаты работы обсуждались на научно-исследовательских семинарах в Сибирском федеральном университете и Сибирском государственном университете науки и технологий имени академика М. Ф. Решетнёва.

Публикации. По результатам диссертационного исследования опубликовано 12 работ, из которых 4 изданы в журналах, рекомендованных ВАК, 7 в тезисах и трудах конференций и 1 свидетельство о регистрации программы, зарегистрированное в Реестре программ для ЭВМ.

Структура работы. Диссертационная работа изложена на 125 страницах и состоит из введения, трёх глав, заключения и списка литературы. Библиографических список включает 160 наименований.

Содержание диссертации

Во **введении** обоснована актуальность темы диссертационной работы, определены цель и задачи исследования, указаны применяемые в работе методы, представлены основные результаты.

В **главе 1** даётся обзор и сравнительный анализ основных методов оценивания функции плотности вероятности (решается задача а) диссертационного исследования). Приводятся необходимые сведения из теории меры и функционального анализа. Определяется пространство $L_p(\Omega, \Sigma, \mu)$ μ-интегрируемых в p-й степени функций, заданных на множестве Ω, где μ – мера, определённая на системе подмножеств Σ множества Ω, $p \geq 1$. Указываются достаточные условия на μ для того, чтобы пространство $L_2(\Omega, \Sigma, \mu)$ было гильбертовым. Приводится постановка задачи оптимизации оценки функции плотности вероятности.

Рассмотрим некоторый класс $\mathcal{F} = \{\hat{f}_\alpha(\mathbf{x}) \mid \alpha \in \mathrm{A}\}$ оценок функции плотности вероятности $f(\mathbf{x})$, где α – набор параметров, A – некоторое множество. Критерием близости оценки $\hat{f}_\alpha(\mathbf{x})$ к истинной плотности является следующий функционал:

$$Q_p\{\hat{f}\} = M\left\{\left\|\hat{f} - f\right\|_{L_p}^p\right\}. \tag{1.1}$$

Набор параметров α выбирается, исходя из условия

$$Q_p\{\hat{f}_\alpha\} \to \min_\alpha.$$

Если $p = 2$, то функционал (1.1) допускает следующее преобразование:

$$Q_2\{\hat{f}_\alpha\} = M\left\{\left\|\hat{f}_\alpha\right\|^2 - 2\left(\hat{f}_\alpha, f\right)\right\} + \|f\|^2.$$

Слагаемое $\|f\|^2$, независимое от α, при минимизации обычно опускается. Тогда приходим к следующей задаче:

$$W\{\hat{f}_\alpha\} = M\left\{\left\|\hat{f}_\alpha\right\|^2 - 2\left(\hat{f}_\alpha, f\right)\right\} \to \min_\alpha,$$

эквивалентной задаче минимизации функционала (1.1). В ряде работ (А. В. Лапко, А. В. Медведев, А. А. Новосёлов) этот подход используется при настройке непараметрических оценок функции плотности вероятности.

Проекционная оценка функции плотности вероятности случайной величины определяется следующим образом:

$$\hat{f}(x) = \sum_{j=0}^{l} a_j \varphi_j(x). \tag{1.2}$$

Настройка коэффициентов a_j и длины ряда l традиционно осуществляется по формулам:

$$a_j = \frac{1}{n} \sum_{i=0}^{n} \varphi_j(x_i), \quad j = 0, \ldots, l; \tag{1.3}$$

$$\hat{W}_l = \sum_{j=0}^{l} \left(\frac{2}{n} s^2_{\varphi_j} - m^2_{\varphi_j} \right); \tag{1.4}$$

$$\hat{l} = \arg\min_{l} \hat{W}_l. \tag{1.5}$$

Оценка Розенблатта — Парзена определяется формулой:

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} \Phi \left(\frac{x - x_i}{h} \right). \tag{1.6}$$

Известные методы настройки параметра размытости h основаны на построении несмещённой оценки функционала качества:

$$\hat{W}(h) = \frac{1}{hn^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \tau \left(\frac{x_i - x_j}{h} \right) - \frac{2}{hn(n-1)} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} K \left(\frac{x_j - x_i}{h} \right); \tag{1.7}$$

$$\hat{h} = \arg\min_{h} \hat{W}(h). \tag{1.8}$$

В главе 1 указаны и другие методы настройки параметра размытости h.

Глава 2 посвящена исследованию проекционной оценки. В ней решаются задачи б), в) диссертационного исследования. Основные результаты второй главы опубликованы в работах [1, 2, 3, 4, 5, 11, 12].

Построение проекционной оценки является мощным непараметрическим методом восстановления функции плотности вероятности. В отличие от оценки Розенблатта — Парзена и других непараметрических оценок проекционная оценка не содержит в себе всей исследуемой выборки и допускает лаконичное математическое выражение. В § 2.1 решается проблема применимости проекционной оценки для оценивания функций плотности вероятности с несуммируемым квадратом.

Пусть $k \in \mathbb{N}$, μ – мера Лебега в $\mathbb{R}^k$, Ω – μ-измеримое подмножество в $\mathbb{R}^k$, $p \geq 1$, $w : \Omega \to \mathbb{R}$, причём функция $w(\mathbf{x})$ является μ-измеримой, положительной почти всюду на Ω и выполняется условие $\operatorname{ess\,sup}\limits_{\mathbf{x}\in\Omega} w(\mathbf{x}) < +\infty$.

Вводится следующее определение.

Определение 2.1. ***Весовое функциональное пространство*** $L_{p,w}(\Omega)$ *определим как множество μ-измеримых функций $f : \Omega \to \mathbb{R}$, удовлетворяющих неравенству*

$$\left(\int_\Omega |f(\mathbf{x})|^p w(\mathbf{x}) d\mu\right)^{1/p} < +\infty.$$

Функцию $w(\mathbf{x})$ при этом будем называть ***весовой функцией****.*

Верны предложения.

Предложение 2.2. *Всякое пространство $L_{p,w}(\Omega)$ является полным по своей норме $\|\cdot\|_{p,w}$.*

Предложение 2.3. *Всякое пространство $L_{p,w}(\Omega)$ является сепарабельным.*

Таким образом, при $p = 2$ пространство $L_{2,w}(\Omega)$ является гильбертовым. Следовательно, в этом пространстве определена проекционная оценка.

Предложение 2.4. *Пусть $\boldsymbol{\xi}$ – непрерывный случайный вектор, $f(\mathbf{x})$ – его функция плотности вероятности, множество $\Omega \subseteq \mathbb{R}^k$ удовлетворяет условию*

$$\int_\Omega f(\mathbf{x}) d\mu(\mathbf{x}) = 1,$$

$p \geq 1$. Тогда существует такая весовая функция $w(\mathbf{x})$, что $f \in L_{p,w}(\Omega)$.

Доказаны следующие необходимые и достаточные условия на весовые функции w_1 и w_2 для того, чтобы пространство $L_{p,w_1}(\Omega)$ было шире пространства $L_{p,w_2}(\Omega)$.

Следствие 2.3.3. Пространство $L_{p,w_1}(\Omega)$ является расширением пространства $L_{p,w_2}(\Omega)$, т.е. $L_{p,w_2}(\Omega) \subset L_{p,w_1}(\Omega)$, тогда и только тогда, когда одновременно выполняются условия:

1) $\operatorname{ess\,inf}\limits_{\mathbf{x}\in\Omega} \dfrac{w_1(\mathbf{x})}{w_2(\mathbf{x})} = 0;$

2) $\operatorname{ess\,sup}\limits_{\mathbf{x}\in\Omega} \dfrac{w_1(\mathbf{x})}{w_2(\mathbf{x})} < +\infty.$

Также доказано более сильное утверждение о том, что при таком расширении пространство $L_{p,w_1}(\Omega)$ содержит более широкое множество функций плотности вероятности.

Теорема 2.3. *Пусть $w(\mathbf{x})$ – измерима и положительна почти всюду на Ω. Тогда измеримая функция $g(\mathbf{x})$, удовлетворяющая условиям*

1) $\int_\Omega |g(\mathbf{x})| d\mu = +\infty;$

2) $\int_\Omega |g(\mathbf{x})| w(\mathbf{x}) d\mu < +\infty$

существует тогда и только тогда, когда

$$\operatorname{ess\,inf}_{\mathbf{x}\in\Omega} w(\mathbf{x}) = 0. \tag{2.1}$$

Теорема 2.4. *Пусть весовая функция $w(\mathbf{x})$ удовлетворяет равенству* (2.1), *$p > 1$. Тогда найдётся измеримая на Ω функция $g(\mathbf{x})$, принадлежащая множеству $(L_{p,w}(\Omega) \setminus L_p(\Omega)) \cap L_1(\Omega)$.*

При настройке проекционной оценки в многомерном случае представляет интерес вопрос нахождения базиса в пространстве $L_{2,w}(\Omega)$, где $\Omega \subseteq \mathbb{R}^k$. Оказалось, что если выполняется условие

$$w(x_1, \ldots, x_k) = w_1(x_1) \ldots w_k(x_k). \tag{2.2}$$

причём $w_i : \Omega_i \to \mathbb{R}$, $\Omega_1 \times \cdots \times \Omega_k \supseteq \Omega$, то мера μ_w, индуцированная весовой функцией w является прямым произведением мер μ_{w_i}:

$$\mu_w = \mu_{w_1} \otimes \cdots \otimes \mu_{w_k}, \quad \mu_{w_i}(X) = \int_X w_i(x) dx,$$

и, следовательно, базисом в пространстве $L_{2,w}(\Omega)$ является система функций

$$\psi_{j_1,\ldots,j_k}(x_1, \ldots, x_k) = \varphi_{1,j_1}(x_1) \cdot \ldots \cdot \varphi_{k,j_k}(x_k), \quad j_1, \ldots, j_k = 0, 1, \ldots,$$

где $\{\varphi_{1,j}\}_{j=0}^\infty$, …, $\{\varphi_{k,j}\}_{j=0}^\infty$ – базисы в пространствах $L_{2,w_1}(\Omega_1)$, …, $L_{2,w_k}(\Omega_k)$ соответственно. Преимуществом данного подхода является упрощение выкладок и уменьшение вычислительной сложности при построении проекционной оценки. Недостатком является ограничение множества восстанавливаемых функций плотности вероятности. Так, доказана следующая теорема.

Теорема 2.6. *При $k \geq 2$ существуют функции $f \in L_1(\Omega)$, не принадлежащие никакому пространству $L_{p,w}(\Omega)$, в котором весовая функция имеет вид* (2.2).

Примером такой функции плотности вероятности при $k = 2$ является

$$f(x_1, x_2) = \frac{3}{8\sqrt{|x_1 - x_2|}}, \quad x_1, x_2 \in [0; 1], \quad x_1 \neq x_2.$$

В § 2.2 исследуется проекционная оценка плотности вероятности, в которой параметры настраиваются методом моментов, а также предлагается некоторое обобщение последнего.

Применение метода моментов для оценивания коэффициентов a_j проекционной оценки сводится к решению системы линейных уравнений, которая в матричном виде записывается следующим образом:

$$\begin{pmatrix} (1,\varphi_0)_w & \dots & (1,\varphi_l)_w \\ \vdots & \ddots & \vdots \\ (x^l,\varphi_0)_w & \dots & (x^l,\varphi_l)_w \end{pmatrix} \cdot \begin{pmatrix} a_0 \\ \vdots \\ a_l \end{pmatrix} = \begin{pmatrix} \hat{\nu}_0 \\ \vdots \\ \hat{\nu}_l \end{pmatrix}, \tag{2.3}$$

где $(x,y)_w$ – скалярное произведение в весовом пространстве $L_{2,w}$, $\hat{\nu}_j = \frac{1}{n}\sum_{i=1}^{n} x_i^j$ – j-й выборочный начальный момент исследуемой случайной величины, причём $\hat{\nu}_0 \equiv 1$. Если основная матрица системы (2.3) не вырождена, то она имеет единственное решение, которое берётся в качестве искомых оценок a_j.

Основная идея обобщения метода моментов состоит в том, что для оценивания того же количества параметров a_j используется большее количество выборочных начальных моментов $\hat{\nu}_j$. Пусть требуется оценить $(l+1)$ коэффициентов a_0, …, a_l. Выберем произвольное натуральное $l' > l$ и запишем для него соответствующую систему (2.3):

$$\begin{pmatrix} (1,\varphi_0)_w & \dots & (1,\varphi_{l'})_w \\ \vdots & \ddots & \vdots \\ (x^{l'},\varphi_0)_w & \dots & (x^{l'},\varphi_{l'})_w \end{pmatrix} \cdot \begin{pmatrix} a_0 \\ \vdots \\ a_{l'} \end{pmatrix} = \begin{pmatrix} \hat{\nu}_0 \\ \vdots \\ \hat{\nu}_{l'} \end{pmatrix}. \tag{2.4}$$

Если основная матрица системы (2.4) не вырождена, то имеется единственное решение $(a_0, a_1, \dots, a_{l'})^T$, из которого выбирается подматрица $(a_0, a_1, \dots, a_l)^T$, которая берётся в качестве набора искомых коэффициентов для оценки (1.2). Данные значения обозначаются через $a_j^{(l')}$.

Для оценок $a_j^{(l')}$ получены следующие свойства:

– при использовании ортонормированной системы Лежандра выполняется равенство $a_j^{(l')} = a_j^{(l)}$ при любом $l' > l$;

– в общем случае $a_j^{(l')} \neq a_j^{(l)}$, но при определённых условиях на базис при неограниченном увеличении l' и фиксированных n и l оценка $a_j^{(l')}$ сходится к традиционной оценке $a_j = \frac{1}{n}\sum_{i=1}^{n} \varphi_j(x_i)$ почти наверное;

– в общем случае оценки a_j не являются несмещёнными;

– оценка плотности, в которой коэффициенты рассчитываются обобщённым методом моментов при подходящем выборе параметра l' , ближе в среднем квадратичном к истинной плотности при любой длине ряда l.

В § 2.3 рассматривается задача оценивания параметров l и l'. Идея метода взята из работы А. А. Новосёлова. Оптимальные значения l^* и $(l')^*$

определяются из условия

$$Q\left\{\hat{f}_l^{(l')}\right\} = M\left\{\left\|\hat{f}_l^{(l')} - f\right\|^2\right\} \to \min_{l,l'}.$$

Вводится функционал

$$W\left\{\hat{f}_l^{(l')}\right\} = M\left\{\left\|\hat{f}_l^{(l')}\right\|^2 - 2\left(\hat{f}_l^{(l')}, f\right)\right\},$$

минимизация которого эквивалентна минимизации функционала Q:

$$\arg\min_{l,l'} Q\left\{\hat{f}_l^{(l')}\right\} = \arg\min_{l,l'} W\left\{\hat{f}_l^{(l')}\right\}.$$

В ходе исследования удалось построить несмещённую оценку функционала $W\left\{\hat{f}_l^{(l')}\right\}$:

$$\hat{W}_{l,l'} = \operatorname{tr} \mathbf{B}_{l,l'}^{-1}\left(\hat{\boldsymbol{\nu}}\hat{\boldsymbol{\nu}}^T\left(\mathbf{B}_{l,l'}^{-1}\right)^T - 2\hat{\mathbf{G}}\right), \tag{2.5}$$

где

$$\hat{\mathbf{G}} = \|g_{j_1,j_2}\|_{\substack{j_1=0,\ldots,l' \\ j_2=0,\ldots,l}},$$

$$g_{j_1,j_2} = \frac{1}{n(n-1)}\left(\sum_{i=1}^n x_i^{j_1}\sum_{i=1}^n \eta_{j_2}(x_i) - \sum_{i=1}^n x_i^{j_1}\eta_{j_2}(x_i)\right).$$

Тогда оценки $\hat{l}$ и $\hat{l}'$ находятся путём минимизации $\hat{W}_{l,l'}$:

$$\left(\hat{l}, \hat{l}'\right) = \arg\min_{l,l'} \hat{W}_{l,l'}.$$

Было проведено сравнение (см. табл. 2.1) предложенных методов настройки проекционной оценки ($\hat{f}_3(x)$ и $\hat{f}_4(x)$) с традиционными методами ($\hat{f}_1(x)$ и $\hat{f}_2(x)$), которое показало, что независимо от используемого базиса и восстанавливаемого распределения обобщённый метод моментов на малых выборках даёт близкие или лучшие результаты. В табл. 2.1 приведено сравнение качества оценок при $n = 100$.

Таблица 2.1. Сравнение качества оценок

Распределение	$\hat{f}_1(x)$	$\hat{f}_2(x)$	$\hat{f}_3(x)$	$\hat{f}_4(x)$
равномерное	0.466 ± 0.033	$\mathbf{0.433 \pm 0.030}$	$\mathbf{0.433 \pm 0.030}$	$\mathbf{0.433 \pm 0.030}$
кубическое	0.225 ± 0.032	$\mathbf{0.198 \pm 0.028}$	$\mathbf{0.198 \pm 0.028}$	$\mathbf{0.198 \pm 0.028}$
показательное	0.078 ± 0.013	0.071 ± 0.010	0.049 ± 0.008	$\mathbf{0.047 \pm 0.008}$
нормальное	0.051 ± 0.006	$\mathbf{0.047 \pm 0.005}$	0.050 ± 0.005	0.048 ± 0.005

В § 2.4 предложенный подход распространяется на многомерный случай. Была получена несмещённая оценка функционала W, путём минимизации которой находятся оценки $\hat{l}_1, \ldots, \hat{l}_k$. Сравнение с традиционным подходом на тестовых распределениях показало, что предложенный подход даёт лучшие результаты.

В этом же параграфе рассматривается задача подбора весовой функции. Показано, что при определённых условиях пространство $L_{2,w}(\Omega)$ можно расширить следующим образом:

$$w_0(\mathbf{x}) = \begin{cases} w(\mathbf{x}) \left(\sum_{j=1}^{k} |x_j - a_j|^{\alpha_j} \right)^p, & x \in U_\delta(\mathbf{a}) \\ w(\mathbf{x}), & x \in \Omega \setminus U_\delta(\mathbf{a}) \end{cases}, \tag{2.6}$$

где $\delta > 0$, $U_\delta(\mathbf{a})$ – δ-окрестность точки $\mathbf{a}$, а показатели α_j удовлетворяют следующему неравенству:

$$\frac{1}{\alpha_1} + ... + \frac{1}{\alpha_k} \leq 1.$$

В **главе 3** рассматриваются три задачи анализа данных, решаемые с помощью представленных в главе 2 результатов: восстановление функции регрессии, классификация и оценивание количества информации. Предложены алгоритмы настройки коэффициентов и длины ряда проекционной оценки функции плотности вероятности случайного вектора, предназначенные для решения прикладных задач на малых выборках. Алгоритмы реализованы на языке Wolfram Language. Приводится сравнение разработанных алгоритмов с известными алгоритмами настройки на малых выборках.

В § 3.1 рассматривается задача восстановления функции регрессии. Даётся описание трёх алгоритмов её решения. Первый алгоритм $\bar{\varphi}_1$ основан на проекционной оценке функции плотности вероятности, в которой параметры настраиваются при помощи метода моментов; два других $\bar{\varphi}_2$, $\bar{\varphi}_3$ – на оценке Розенблатта – Парзена, с разными способами настройки параметра размытости. Сравнение разработанных алгоритмов на тестовых задачах по критерию средней близости к истинной функции регрессии показало, что алгоритмы, основанные на оценке Розенблатта – Парзена, эффективнее, чем алгоритм, основанный на проекционной оценке, причём способ настройки параметра размытости оказался несущественным (табл. 3.1).

Таблица 3.1. Результаты восстановления функции регрессии

Обозначение	**Оценка плотности**	**Оценка параметров**	$Q\{\hat{I}\}$
$\bar{\varphi}_1$	проекционная (1.2), базис Лежандра	оценка Ченцова	0.1101 ± 0.0054
$\bar{\varphi}_2$	Розенблатта — Парзена (1.6), параболическое ядро	минимизация оценки функционала качества	$\mathbf{0.03896 \pm 0.00074}$
$\bar{\varphi}_3$	Розенблатта — Парзена (1.6), параболическое ядро	метод максимального правдоподобия	$\mathbf{0.03906 \pm 0.00078}$

В § 3.2 приводится постановка задачи классификации. Для решения данной задачи разработано 3 алгоритма, в которых при синтезе решающих правил использованы различные оценки плотности. Качество классификации

оценивалось по средней близости восстановленных областей классификации к истинным областям:

$$Q\{\hat{\eta}\} = M\{\rho(\hat{S}, S)\} = M\{\mu(S_1 \cap \hat{S}_1) + \mu(S_2 \cap \hat{S}_2)\}, \tag{3.1}$$

где S_1, S_2 – истинные области классификации, $\hat{S}_1$, $\hat{S}_2$ – соответствующие восстановленные области, $S = (S_1, S_2)$, $\hat{S} = (\hat{S}_1, \hat{S}_2)$, μ – мера Лебега на плоскости (площадь плоской фигуры).

Результаты работы алгоритмов приведены на рис. 1.

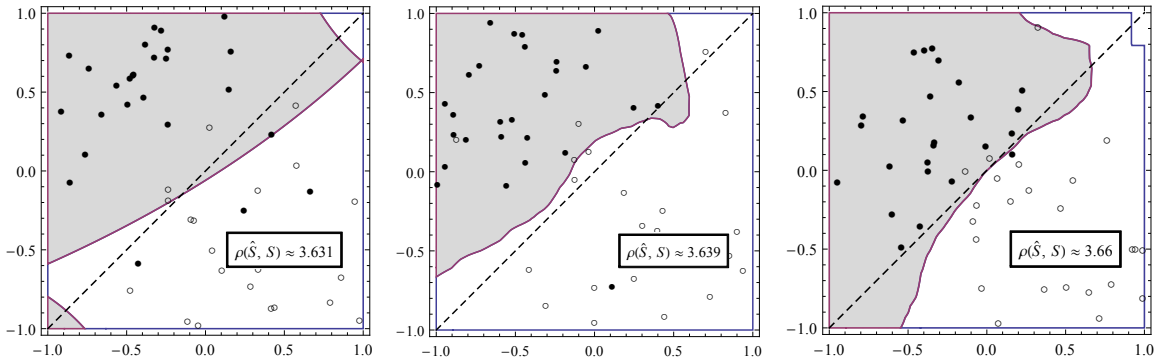

Рис. 1. Примеры решения задачи классификации.

Сравнение разработанных алгоритмов на тестовых задачах показало, что улучшение достигается при использовании алгоритма, основанного на оценке Розенблатта — Парзена (табл. 3.2).

Таблица 3.2. Результаты решения задачи классификации

Обозначение	**Оценка плотности**	**Оценка параметров**	$Q\{\hat{I}\}$
$\hat{\eta}_1$	проекционная (1.2), базис Лежандра	оценка Ченцова	3.55 ± 0.04
$\hat{\eta}_2$	Розенблатта — Парзена (1.6), параболическое ядро	минимизация оценки функционала качества	$\mathbf{3.56 \pm 0.02}$
$\hat{\eta}_3$	Розенблатта — Парзена (1.6), параболическое ядро	метод максимального правдоподобия	3.51 ± 0.02

В § 3.3 рассматривается задача оценивания количества информации. Для решения данной задачи было разработано 4 алгоритма, для которых было выполнено сравнение эффективности на тестовых задачах по критерию средней близости к истинному значению. Сравнительный анализ показал, что наибольшая точность достигается при использовании алгоритма, основанного на оценке Розенблатта — Парзена. При этом при настройке проекционной оценки эффективнее оказался предлагаемый алгоритм, основанный на методе

моментов (табл. 3.3).

Таблица 3.3. Результаты оценивания количества информации

Обозначение	**Оценка плотности**	**Оценка параметров**	$Q\{\hat{I}\}$
$\hat{I}_1$	проекционная (1.2), базис Эрмита	оценка Ченцова	0.143 ± 0.01
$\hat{I}_2$	проекционная (1.2), базис Эрмита	метод моментов	$\mathbf{0.135 \pm 0.01}$
$\hat{I}_3$	Розенблатта — Парзена (1.6), ядро Гаусса	минимизация оценки функционала качества	0.10 ± 0.01
$\hat{I}_4$	Розенблатта — Парзена (1.6), ядро Гаусса	метод максимального правдоподобия	$\mathbf{0.093 \pm 0.003}$

В **заключении** сформулированы основные результаты и выводы диссертационной работы.

Основные результаты и выводы

В ходе выполнения диссертационной работы были получены следующие результаты:

– показано, что весовое гильбертово пространство $L_{2,w}(\Omega)$ может быть использовано для построения проекционной оценки любой функции плотности вероятности (предл. 2.4);

– найден критерий на весовую функцию $w(\mathbf{x})$ для расширения пространства $L_2(\Omega)$ до пространства $L_{2,w}(\Omega)$, которое содержит более широкое множество функций плотности вероятности (теорема 2.4);

– предложен способ построения весовой функции $w(\mathbf{x})$, при котором соответствующее расширение $L_{2,w}(\Omega)$ пространства $L_2(\Omega)$ содержит оцениваемую функцию плотности вероятности $f(\mathbf{x})$ (формула (2.6));

– предложен новый метод настройки коэффициентов проекционной оценки функции плотности вероятности случайного вектора, являющийся обобщением метода моментов;

– доказано, что при определённых условиях частным случаем предложенного обобщения является традиционный метод оценивания коэффициентов;

– предложен новый метод оценивания длины ряда проекционной оценки, в которой коэффициенты настраиваются методом моментов или его обобщением;

– экспериментально установлено, что на малых выборках обобщение метода моментов позволяет повысить эффективность проекционной оценки (табл. 2.1);

– экспериментально установлено, что условиях малых выборок метод моментов является более предпочтительным при настройке проекционной оценки.

Публикации по теме диссертации

В изданиях, рекомендованных ВАК:

1. Branishti, V. V. On some Properties of Weighted Hilbert Spaces // Journal of Siberian Federal University. Mathematics & Physics. – 2017. – № 10 (4). – С. 410–421.

2. Браништи, В. В. Введение пространства $L_{2,w}$ при построении проекционной оценки плотности вероятности // Вестник СибГАУ. – 2016. – № 1. – С. 19–26.

3. Браништи, В. В. Некоторые обобщения метода моментов при оценивании плотности вероятности в виде ортогонального ряда // Вестник СибГАУ. – 2015. – Том 16, № 3. – С. 566–571.

4. Браништи, В. В. О параметрическом оценивании функции плотности вероятности // Научно-технический вестник Поволжья. – 2014. – № 1. – С. 13–16.

Свидетельства о регистрации программ для ЭВМ:

5. Браништи, В. В. Непараметрическое оценивание плотности распределения вероятности случайной величины / СибГАУ. – Свидетельство о государственной регистрации программы для ЭВМ №2008610808 от 15.02.2008 г.

В других изданиях:

6. Браништи, В. В. Один способ расчёта коэффициента размытости для непараметрической оценки Розенблатта — Парзена плотности распределения вероятности случайной величины // Наука. Технологии. Инновации: Материалы всероссийской научной конференции (Новосибирск, 6–9 декабря 2007 г.). Часть 1. – Новосибирск: Изд-во НГТУ, 2007. С. 18–19.

7. Браништи, В. В. Оптимизация алгоритмов настройки коэффициента размытости для непараметрических оценок // Молодежь и наука: сборник материалов всероссийской научно-технической конференции [Электронный ресурс] // Отв. ред. О. А. Краев. – Красноярск: Сиб. федер. ун-т, 2014. – Режим доступа: `http://conf.sfu-kras.ru/sites/mn2014/pdf/d02/s14/s14_002.pdf`

8. Браништи, В. В. Построение оценок плотности вероятности в виде суммы дельта-образных функций // Национальная ассоциация ученых. – 2015. – № 4 (9). Часть 7. – С. 10–13.

9. Браништи, В. В. Построение проекционных оценок для плотностей вероятности с неинтегрируемым квадратом // Решетнёвские чтения: материалы международной научно-практической конференции (Красноярск, 9–12 ноября 2016 г.). – Красноярск: СибГАУ, 2016. – С. 96–98.

10. Браништи, В. В. Применение метода моментов при оценивании функции плотности вероятности в виде линейных комбинаций ортогональных функций // Решетнёвские чтения: материалы международной научной

конференции (Красноярск, 11–13 ноября 2014 г.). – Красноярск: СибГАУ, 2014. С. 22–24.

11. Браништи, В. В. Сравнение двух алгоритмов настройки длины ряда для проекционной оценки плотности вероятности // Актуальные проблемы гуманитарных и естественных наук. – 2016. – № 9 (92), ч. 1. – С. 10–14.

12. Браништи, В. В. Сравнение проекционных оценок плотности вероятности // Актуальные проблемы авиации и космонавтики: сборник материалов (Красноярск, 14 апреля 2017 г.). – Том 2. – Красноярск: СибГАУ, 2017. – С. 262–264.