На правах рукописи

Янович Юрий Александрович

# Асимптотические свойства процедур статистического оценивания на многообразиях

01.01.05 — Теория вероятностей и математическая статистика

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата физико-математических наук

Москва
2017

Работа выполнена в лаборатории №10 Института проблем передачи информации им. А.А. Харкевича Российской академии наук (ИППИ РАН).

Научный руководитель: доктор физико-математических наук,
главный научный сотрудник, профессор
БЕРНШТЕЙН Александр Владимирович
Центр по научным и инженерным
вычислительным технологиям
Сколковский институт науки и технологий (Сколтех)
(Москва)

Официальные оппоненты: доктор физико-математических наук,
проф. кафедры математической статистики, профессор
БЕНИНГ Владимир Евгеньевич,
Факультет вычислительной математики и кибернетики
ФГБОУ ВО «Московский государственный университет
имени М.В. Ломоносова» (Москва)

кандидат физико-математических наук,
научный сотрудник отдела теории вероятностей
и математической статистики
ЖИТЛУХИН Михаил Валентинович
Математический институт им. В.А. Стеклова
Российской академии наук (Москва)

Ведущая организация: Казанский федеральный университет (КФУ, Казань)

Защита диссертации состоится "5"сентября 2017 года в 17 часов на заседании совета Д002.077.03 при федеральном государственном бюджетном учереждении науки Институте проблем передачи информации им. А.А. Харкевича РАН (ИППИ РАН), расположенном по адресу: 127051, г. Москва, Большой Каретный переулок, д. 19, стр. 1.

С диссертацией можно ознакомиться в библиотеке ИППИ РАН.

Автореферат разослан "___"июля 2017 года.

Отзывы и замечания по автореферату в двух экземплярах, заверенные печатью, просьба высылать по вышеуказанному адресу на имя ученого секретаря диссертационного совета.

Ученый секретарь диссертационного совета,
д. ф.-м. н., профессор РАН Соболевский А.Н.

# Общая характеристика работы

## Актуальность темы

Развитие вычислительной техники и информационных технологий привело к возможности хранить, передавать по каналам связи и быстро обрабатывать большие массивы данных, осуществлять быстрый удаленный доступ к ним. Появление в результате таких возможностей "шквала данных", называемого обычно парадигмой *больших данных* (Big Data), и новые возможности для работы с ними, позволили ставить и эффективно решать новые научные и прикладные задачи в области промышленности и сельского хозяйства, биологии и медицины, обработки сигналов, речи, текстов и изображений, и др. Для решения таких задач были развиты методы работы с большими данными, научную основу которым составляет новая мультидисциплинарная область знаний, выделившаяся в XXI веке в отдельную академическую и университетскую дисциплину *наука о данных* (Data Science), в которой сконцентрированы методы математики и статистики, распознавания образов, визуализации и машинного обучения, информационных технологий и *интеллектуального анализа данных*.

Понятие *большие данные* включает в себя не только большой объем данных, но и их высокую размерность, так как реальные данные обычно имеют очень высокую размерность (например, размерность цифровой черно-белой фотографии равна числу ее пикселей и может достигать сотен тысяч; изображения головного мозга, получаемые ежесекундно с помощью функциональной магнито-резонансной томографии, имеют размерность порядка полутора миллионов). Однако многие традиционные методы и алгоритмы становятся неэффективными или просто неработоспособными для данных высокой размерности (не только в вычислительном, но и в статистическом смысле), и этот феномен назван *проклятием размерности*[1]. Известный статистик Д. Донохо сказал в 2000 году на конференции, посвященной математическим вызовам 21-го века: "мы можем с полной уверенностью сказать, что в наступающем веке анализ многомерных данных станет очень важным занятием, и совершенно новые методы многомерного анализа данных будут разработаны, просто мы еще не знаем, каковы они будут"[2].

Однако совокупность конкретных "реальных" данных, полученных из реальных источников, в силу наличия различных зависимостей между компонентами данных и ограничений на их возможные значения, занимает, как правило, малую часть высокоразмерного пространства наблюдений,

[1]Richard E. Bellman. Dynamic programming. Princeton University Press, 1957.

[2]David L. Donoho. High-dimensional data analysis: The curses and blessings of dimensionality. AMS conference on math challenges of 21st century, 2000.

имеющую невысокую внутреннюю размерность (например, множество всех черно-белых портретных изображений человеческих лиц с исходной размерностью порядка сотен тысяч, имеет внутреннюю размерность не выше ста). Следствием невысокой внутренней размерности является возможность построения низкоразмерной параметризации таких данных. Поэтому многие алгоритмы для работы с высокоразмерными данными начинаются с решения задачи снижения размерности, результатом которого являются низкоразмерные описания таких данных.

Традиционные методы математической статистики в задаче снижения размерности рассматривают, в основном, лишь случай, когда высокоразмерные данные сосредоточены вблизи неизвестного низкоразмерного аффинного подпространства, которое можно оценить, например, с помощью метода главных компонент К. Пирсона. Однако многие реальные данные имеют нелинейную, "искривленную" структуру, для нахождения которой в прошлом веке были предложены различные эвристические методы нелинейного снижения размерности (например, репликативные нейронные сети[3] или нелинейные методы многомерного шкалирования[4]), свойства которых невозможно было исследовать строгими методами в силу отсутствия математической модели для обрабатываемых данных. Имелись лишь отдельные работы (например, работы В.М. Бухштабера[5,6]) в которых строгие дифференциально-геометрические и статистические методы использовались при решении нелинейных статистических задач.

Лишь в 2000 году появилась первая математическая модель многомерных нелинейных данных, названная *моделью многообразия* (Manifold model[7]), в соответствии с которой, высокоразмерные данные расположены на (или вблизи) неизвестного низкоразмерного нелинейного многообразия (*многообразия данных*), вложенного в высокоразмерное пространство наблюдений, а наблюдаемая выборка случайно извлекается из многообразия данных в соответствии с неизвестным вероятностным распределением на нем. Размерность *многообразия данных* также может быть неизвестной и оцениваться по выборке.

Область анализа данных, в которой статистические задачи рассматриваются в рамках этой модели, получила название *моделирование многообразий* (Manifold Learning[8]) и в настоящее время является бурно раз-

[3]G.E. Hinton, R.R. Salakhutdinov. Reducing the dimensionality of data with neural networks. Science, 2000.

[4]T.F. Cox, M.A.A. Cox. Multidimensional Scaling. Chapman and Hall/CRC, London, 2001.

[5]В. М. Бухштабер, В. К. Маслов. Факторный анализ и экстремальные задачи на многообразиях Грассмана. Математические методы решения экономических задач, № 7. Наука, 1977.

[6]V. M. Buchstaber. Time series analysis and Grassmannians. Applied problems of Radon transform, Amer. Math. Soc. Transl. Ser. 2, 162, Amer. Math. Soc., Providence, RI, 1994.

[7]H.S. Seung, D.D. Lee. The Manifold Ways of Perception. Science, 2000.

[8]Y. Ma, Y. Fu (eds.). Manifold Learning Theory and Applications. CRC Press, London, 2011.

вивающимся разделом *науки о данных*. Изначально под моделированием многообразий понимались только задачи нелинейного снижения размерности[8,9], впоследствии в эту область были включены задачи регрессии на многообразии[10], оценки плотности на многообразии[11], и др.

Методы и алгоритмы моделирования многообразий существенно опираются на геометрическую структуру данных и позволяют эффективно решить большое количество прикладных задач анализа данных. Эти методы "эксплуатируют" различные дифференциально-геометрические свойства многообразий данных (например, возможность аппроксимации окрестности выбранной точки многообразия касательной плоскостью невысокой размерности, описание кратчайших расстояний на многообразии геодезическими линиями, и другие). По существу, появился новый раздел многомерного статистического анализа — анализ данных, лежащих на нелинейном многообразии меньшей размерности (manifold valued data), в котором существенную роль играют такие дифференциально-геометрические характеристики нелинейного многообразия данных как кривизна, риманова структура и др.

Большинство работ по моделированию многообразий содержит описание различных процедур, свойства которых исследуются с помощью вычислительных экспериментов. Лишь в небольшом числе работ[12,13] математически строго исследовались статистические свойства конкретных функций от данных (статистик). При этом известные фундаментальные методы анализа асимптотического поведения статистик, развитые во второй половине прошлого века, значительный вклад в которые внесли, в том числе, такие видные советские и российские исследователи как Л.Н. Большев, И.А. Ибрагимов, Р.З. Хасьминский, А.А. Боровков, Д.М. Чибисов и др., разработаны для данных на линейных подпространствах и непосредственно неприменимы для данных лежащих на нелинейных многообразиях.

В процедурах моделирования многообразий используются, как правило, три типа статистик:

1. локальные статистики, построенные по подвыборке, состоящей из точек попавших в малую окрестность выбранной точки, и оценивающие локальные и дифференциально-геометрические характеристики мно-

[9]X. Huo and A.K. Smith. A Survey of Manifold-Based Learning Methods. Journal of Machine Learning Research, 2008.

[10]A. Kuleshov and A. Bernstein. Nolinear Multi-Output Regression on Unknown Input Manifold. Annals of Mathematics and Artificial Intelligence, 2017.

[11]G. Henry, A. Mucoz, D. Rodrнguez. Locally adaptive density estimation on Riemannian manifolds. Statistics and Operations Research Transactions, 2013.

[12]A. Smith, H. Zha, X. Wu. Convergence and Rate of Convergence of a Manifold-Based Dimension Reduction Algorithm. Advances in Neural Information Processing Systems, 2009.

[13]L. Rosasco, M. Belkin, E. De Vito. On learning with integral operators. The Journal of Machine Learning Research, 2010.

гообразия (касательное пространство[14], Риманов метрический тензор[15], и др.) в выбранной точке; распределение таких статистик зависит, в том числе, от локальных и дифференциально-геометрических свойств многообразия в рассматриваемой точке;

2. глобальные (интегральные) статистики, являющиеся усреднением локальных статистик по точкам выборки и которые могут также зависеть от параметров (например, искомых неизвестных низкоразмерных представлений);

3. статистики, являющиеся результатом оптимизации глобальных статистик по этим параметрам.

Оптимизация выборочных интегральных статистики во многих случаях сводится к оптимизации выборочных квадратичных форм от этих параметров (алгоритмы IsoMap[16], Locally-linear Embedding[17], Local Tangent Space Alignment[18], Laplacian Eigenmaps[19], Hessian Eigenmaps[20], Grassmann&Stiefel Eigenmaps[21], и др.), решение которых основаны на спектральных методах[22] (обобщенные задачи на собственные векторы). Решение таких выборочных задач часто сводится к решению спектральных непрерывных задач на многообразии. Например, оптимизируемая выборочная квадратичная форма в алгоритме Laplacian Eigenmaps[14,19] является выборочным аналогом оператора Лапласа-Бельтрами на определенном классе функций, заданных на многообразии данных, а наилучшие низкоразмерные представления сходятся к конкретным собственным функциям этого оператора[13].

Однако до сих пор отсутствуют результаты об асимптотическом поведении статистик на многообразиях для достаточно общих классов, включающих в себя как ранее неисследованные статистики из различных алгоритмов, так и новые статистики для создаваемых алгоритмов с желаемы-

[14]A. Singer, H.-T Wu. Vector diffusion maps and the connection Laplacian. Communications on Pure and Applied Mathematics, 2012.

[15]D. Perrault-Joncas and M. Meila. Non-linear Dimensionality Reduction: Riemannian Metric Estimation and the Problem of Geometric Recovery. In: arXiv:1305.7255v1, 2013.

[16]J. B. Tenenbaum, V. de Silva, J. C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. Science, 2000.

[17]S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. Science, 2000.

[18]Z. Zhang and H. Zha. Principal Manifolds and Nonlinear Dimension Reduction via Local Tangent Space Alignment. SIAM Journal on Scientific Computing, 2004.

[19]M. Belkin, P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation, 2003.

[20]D. Donoho and C. Grimes. Hessian eigenmaps: locally linear embedding techniques for high-dimensional data. Proc. Natl. Acad. Sci. USA, 2003.

[21]A.V. Bernstein, A.P. Kuleshov. Manifold Learning: generalizing ability and tangent proximity. International Journal of Software and Informatics, 2013.

[22]L. K. Saul, K. Q. Weinberger, J. H. Ham, F. Sha, and D. D. Lee. Spectral methods for dimensionality reduction. Semisupervised Learning, MIT Press, 2006.

ми свойствами. Именно отсутствие таких общих результатов для классов статистик и явилось мотивацией исследования. Тем самым, целью диссертационной работы является изучение асимптотических свойств выбранных классов статистик каждого из трех типов.

## Основные задачи исследования

Для достижения сформулированной цели были поставлены и решены следующие взаимосвязанные асимптотические статистические задачи:

1. исследование статистических свойств случайных подвыборок, попавших в асимптотически малую окрестность выбранной точки многообразия;

2. исследование асимптотического поведения выбранного класса локальных статистик в окрестности рассматриваемой точки, зависящего, в том числе, от дифференциально-геометрических свойств многообразия в этой точке;

3. исследование асимптотического поведения глобальных (интегральных) статистик получаемых путем усреднения локальных статистик по точкам выборки;

4. исследование сходимости статистик, являющихся решениями выборочных спектральных оптимизационных задач, к решениям предельных непрерывных оптимизационных задач на многообразии.

## Основные выносимые на защиту результаты и их новизна

Все приведенные ниже результаты диссертации являются новыми и получены лично соискателем:

1. найдены асимптотическое распределение числа точек в медленно убывающей окрестности выбранной точки многообразия и условное распределение этих точек. Получена равномерная по многообразию верхняя оценка для вероятности больших уклонений числа точек в окрестности от своего среднего значения;

2. найдено асимптотическое распределение локальных статистик рассматриваемого класса, ранее известное лишь для конкретных статистик (оценки элементов локальной выборочной ковариационной матрицы[14] и оценки оператора Лапласа-Бельтрами специального вида[23]).

[23]E. Gine and V. Koltchinskii. Empirical graph Laplacian approximation of Laplace-Beltrami operators: large sample results.High Dimension Probability, 51:238-259, 2006.

Получена равномерная по многообразию верхняя оценка на вероятности больших уклонений локальных статистик от их средних значений;

3. найдены предельные значения глобальных статистик на многообразиях, принадлежащих рассматриваемому классу; для выбранного класса статистик получена верхняя оценка на вероятности больших уклонений глобальной статистики от своего предельного значения;

4. получена оценка для отклонения собственных чисел и собственных функций, являющихся решениями рассматриваемых выборочных спектральных оптимизационных задач для выбранных глобальных статистик от собственных чисел и собственных функций предельных непрерывных спектральных операторов.

## Основные методы исследования

В работе используются методы теории вероятностей и математической статистики для анализа асимптотического поведения случайных величин; методы математического анализа и дифференциальной геометрии для изучения локальной структуры многообразия — носителя данных; методы функционального анализа для получения равномерных по многообразию оценок и доказательства сходимости результата оптимизации выборочных функционалов к их предельным непрерывным аналогам.

## Теоретическая ценность и практическая значимость

Работа носит теоретический характер. Её результаты могут быть использованы для изучения свойств и совершенствования существующих алгоритмов анализа лежащих на многообразии данных, а также, для создания новых алгоритмов с желаемыми свойствами. В частности, результаты диссертации были использованы в работах по развитию новых алгоритмов моделирования многообразий и их использованию в прикладных задачах анализа данных [4-11], написанных в соавторстве с соискателем. В этих работах лично соискателю принадлежат результаты и выводы, основанные на полученных в диссертации результатах применительно к конкретным рассматриваемым в этих работах статистикам.

## Результаты апробации

Результаты диссертации были доложены на следующих мероприятиях:

1. семинар отдела теории вероятностей и математической статистики МИАН (2017, Москва, Россия);

2. семинар “Теория риска и смежные вопросы” кафедры математической статистики ВМК МГУ (2016, Москва, Россия);

3. математические и статистические семинары ИППИ РАН (Москва, Россия): Добрушинской математической лаборатории (2016), “Статистический кластерный анализ” (2016), сектора интеллектуального анализа данных (2014);

4. междисциплинарные школы-конференции “Информационные технологии и системы”: 40-я (2016, Санкт-Петербург, Россия), 39-я (2015, Сочи, Россия), 37-я (2013, Калининград, Россия);

5. International Conference on Algebra, Analysis and Geometry (2016, Kazan, Russia);

6. The 8th International Conference on Machine Vision, ICMV (2015, Barcelona, Spain);

7. The 14th International Conference on Machine Learning and Applications (2015, Miami, USA; ICMLA);

8. The third international symposium on statistical learning and data sciences, SLDS (2015, Royal Holloway, University of London, United Kingdom);

9. International IEEE Conference on Data Science and Advanced Analytics, DSAA (2015, Paris, France);

10. Yandex School of Data Analysis Conference “Machine Learning: Prospects and Applications” (2015, Berlin, Germany);

11. Premolab-WIAS Workshop “Advances in predictive modeling and optimization” (2014, WIAS, Berlin, Germany);

12. International Workshop on Statistical Learning (2013, Moscow, Russia);

13. The 29th European Meeting of Statisticians, EMS (2013, Budapest, Hungary);

14. Neural Information Processing Systems Conference (NIPS 2013, Lake Tahoe, USA).

## Публикации

Список работ автора по теме диссертации приведен в конце диссертации. Основные результаты диссертации содержатся в работах автора [1-3]. Работы [1-7] опубликованы в изданиях, индексируемых в системах Scopus или Web of Science.

## Структура и объем работы

Диссертация состоит из введения, четырех глав, заключения и списка литературы, включающего 38 наименований. Объем диссертации составляет 84 страницы.

# Краткое содержание диссертации

Во **введении** приведен обзор литературы по рассматриваемой тематике, формулируются цели работы и основные результаты.

**Первая глава** диссертации посвящена локальному поведению случайных выборок на многообразии.

**Постановка задачи.** В разделе **1.1** вводится используемая в диссертации модель многомерных данных и формулируются сделанные предположения. Модель включает в себя описание носителя данных и механизма извлечения выборки. Предполагается, что носителем данных является "достаточно хорошее" (предположения строго сформулированы как $M1 - M8$ в диссертации) неизвестное многообразие данных $\mathbb{M}$ с известной размерностью $q$, вложенным в $\mathbb{R}^p$, $p > q$. Предполагается, что на многообразии имеется "достаточно хорошая" (предположения строго сформулированы как $S1 - S3$ в диссертации) неизвестная мера $\mu$, абсолютно непрерывная относительно меры Римана на многообразии и обладающая плотностью $p_\mu$. Рассматривается выборка данных $\mathbb{X}_N = \{X_1, \ldots, X_N\}$ состоит из независимых одинаково распределенных величин распределенных в соответствии с мерой $\mu$.

Многообразие $\mathbb{M}$ в окрестности каждой точки $X$ может быть приближено своим $q$-мерным касательным пространством $T_X(\mathbb{M})$: в малой окрестности $U_X \subset \mathbb{M}$ точки $X \in \mathbb{M}$ существует взаимнооднозначное соответствие между точками этой окрестности и точками окрестности $TU_X$ нулевого вектора в касательном пространстве, представленных в виде вектора $t\theta$, где $t \geq 0$ и $\theta$ принадлежит единичной сфере $S_{q-1}$ в $\mathbb{R}^q$. Существует так называемое экспоненциальное отображение $\exp_X$ точек $TU_X$ в касательном пространстве в точки $X' = \exp_X(t\theta)$ окрестности $U_X$ (Рис. 1). Обратное отображение из $U_X$ подмножества касательного пространства $T_X(\mathbb{M})$ задает локальные римановы $q$-мерные координаты $t\theta$ точек $X' = \exp_X(t\theta) \in U_X(\mathbb{M})$ многообразия.

Различные методы построения и исследования локальных статистик на

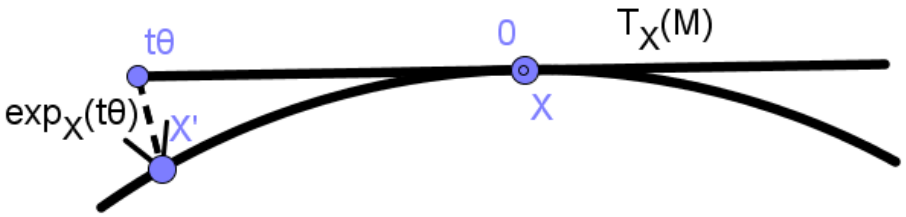


Рис. 1: Экспоненциальное отображение $\exp_X$ в окрестности точки $X$ многообразия $\mathbb{M}$ из касательного пространства $T_X(\mathbb{M})$ к многообразию $\mathbb{M}$ в точке $X$, $t \in [0, \infty)$, $\theta \in S_{q-1} \subset T_X(\mathbb{M})$ — вектор единичной длины.

многообразии, точки которого принадлежат высокоразмерному пространству $\mathbb{R}^p$, в той или иной форме используют локальное приближение многообразия в окрестности рассматриваемой точки его $q$-мерным касательным пространством $T_X(\mathbb{M})$ и порождаемые им $q$-мерные координаты точек окрестности. Рассмотрим в качестве окрестности $U_X$ пересечение $U_{X,\varepsilon}$ $\varepsilon$-шара в $\mathbb{R}^p$ с центром в $X \in \mathbb{M}$ с многообразием $\mathbb{M}$. Очевидно, что подобное приближение будет тем лучше, чем меньше радиус окрестности $\varepsilon$. Однако, для обеспечения состоятельности рассматриваемых локальных оценок, радиус $\varepsilon$ должен быть достаточно большим, чтобы в окрестность $U_{X,\varepsilon}$ попало достаточно много точек при стремлении размера выборки $N$ к бесконечности. Точные требования $P1-P3$ к поведению радиуса $\varepsilon = \varepsilon_N$ как функции от размера выборки $N$, в частности, требование $N \cdot \varepsilon^q \to \infty$, сформулированы в диссертации, этим требованиям удовлетворяет, например, функция

$$E(N) = C \cdot N^{-\frac{1}{q+2}}, \tag{1}$$

где $C > 0$ — константа.

Пусть $X$ — фиксированная точка многообразия $\mathbb{M}$, и $X' \in \mathbb{M}$ — случайная точка, имеющая распределение $\mu$. Обозначим $I_\varepsilon(X'|X)$ бернуллиевскую случайную величину являющуюся индикатором события $|X'-X| < \varepsilon$, с вероятностью успеха

$$P_\varepsilon = \mu(|X' - X| < \varepsilon) = \mathbb{E}_\mu I_\varepsilon(X'|X), \tag{2}$$

где $|\cdot|$ — обычная евклидова норма в $\mathbb{R}^p$. При малых $\varepsilon$ главный член вероятности успеха $P_\varepsilon$ ведет себя[24] (с точностью до членов большего порядка малости) как

$$P_\varepsilon \approx \varepsilon^q \cdot p_\mu(X) \cdot V_q, \tag{3}$$

[24]E. Levina, P. Bickel. Maximum Likelihood Estimation of Intrinsic Dimension. Advances in Neural Information Processing Systems. 2005.

где $V_q$ — объем $q$-мерного единичного шара. Обозначим

$$N_\varepsilon(X) = \sum_{n=1}^{N} I_\varepsilon(X_n|X) \tag{4}$$

биномиальную случайную величину равную числу точек выборки $\mathbb{X}_N$, попавших в окрестность $U_{X,\varepsilon}$ точки $X$. Для $N_\varepsilon(X)$ справедливо соотношение

$$\mathbb{E}N_\varepsilon(X) = N \cdot P_\varepsilon \approx N \cdot \varepsilon_N^q \cdot p_\mu(X) \cdot V_q. \tag{5}$$

В разделе **1.2** исследуется поведение случайной величины $N_\varepsilon(X)$ (4) в схеме серий ($N \to \infty, \varepsilon_N \to 0, N \cdot \varepsilon_N^q \to \infty$) с учетом приближений (3) и (5). Известные результаты[25] для евклидова случая (закон больших чисел, центральная предельная теорема, условная асимптотическая равномерность распределения точек, попавших в убывающую окрестность рассматриваемой точки), в которых вероятность успеха $P_\varepsilon$ (2) медленно стремится к нулю, с использованием стандартной техники[14,19] обобщены в диссертации на случай случайных величин, лежащих на многообразии, в полученном обобщении дополнительно исследовалось "смещение" математического ожидания, вызванное кривизной многообразия в точке $X$. Эти обобщенные результаты сформулированы как Утверждения 1-3 в виде, удобном для дальнейшего использования в диссертации (например, определяют главный член стохастического разложения $N_\varepsilon(X)$ по $\varepsilon$).

Однако, для дальнейшего использовании в диссертационной работе, требуются результаты о поведении случайной величины $N_\varepsilon(X)$ (4) для всех точек многообразия "сразу", оценивая, в частности, вероятность того, что в окрестности всех отделенных от границы точек попадет необходимое бесконечное число точек выборки. Поэтому в диссертации случайные величины $N_\varepsilon(X)$, $X \in \mathbb{M}$, рассматривается как "параметризованное" случайное поле на многообразии данных $\mathbb{M}$. В рассматриваемом нелинейном случае, ошибка аппроксимации в (3) вызвана не только тем, что плотность $p_\mu(X')$ меняется в окрестности $U_{X,\varepsilon}$, но и отличием $U_{X,\varepsilon}$ от $q$-мерного шара в касательной плоскости, вызванного кривизной многообразия в точке $X$. Если поведение членов более высокого порядка малости, обусловленное изменчивостью плотности $p_\mu(X')$ внутри окрестности $U_{X,\varepsilon}$, широко исследовалось и известно в литературе, то влияние кривизны носителя в точках $X$ мало изучено.

**Результаты.** В **Теореме 1** получены оценки для больших уклонений для случайной величины $N_\varepsilon(X)$ для произвольной фиксированной точеки $X$.

[25]P. Deheuvels, M.L. Puri, S.S. Ralescu. Asymptotic expansions for sums of nonidentically distributed Bernoulli random variables. Journal of Multivariate Analysis, 1989.

**Теорема 1** Для любой точки $X \in \mathbb{M}$, удаленной от границы многообразия не менее чем на $\varepsilon$, и произвольного числа $z \in (0, 1/16)$ выполнено:

$$P\left(\left|\frac{N_\varepsilon(X)}{N\varepsilon^q V_q p_\mu(X)} - 1\right| \geq z + \varepsilon^2 \cdot \frac{C_0}{V_q p_\mu(X)}\right) \leq 2 \cdot \exp\left(-4z^2 \cdot N\varepsilon^q V_q p_\mu(X)\right), \quad (6)$$

где $C_0 > 0$ — константа, зависящая только от многообразия $\mathbb{M}$.

Утверждение **Теоремы 1** показывает, что вероятность даже малых относительных уклонений $N_\varepsilon(X)$ от главного члена его разложения экспоненциально затухает с ростом размера выборки (например, в случае радиуса $E(N)$ (1), эта скорость равна $O(\exp(-N^{-2/(q+2)}))$). Аналогичную форму имеют теоремы о больших уклонениях в случае евклидова пространства, однако "прямой перенос" этих результатов на рассматриваемый нелинейный случай приводит к появлению слагаемого порядка $O(\varepsilon)$, связанного с кривизной многообразия в точке $X$, в правой части первого неравенства (под знаком вероятности) в (6). Однако более тонкая использованная в диссертации техника доказательства позволила исключить линейный по $\varepsilon$ член и в случае нелинейного многообразия.

Результат **Теоремы 1** записан в не совсем классическом виде: рассматриваемые уклонения $z + \varepsilon^2 \cdot \frac{C_0}{V_q p_\mu(X)}$ ограничены, так как $z < 1/16$ и $\varepsilon$ ограничено. Можно сформулировать результат и для $z \geq 1/16$, однако целью исследования было ограничение вероятности уклонений экспоненциально затухающей величиной.

В **Теореме 2** получен равномерный аналог **Теоремы 1**, который сформулирован в виде одностороннего неравенства и показывает, что вероятность того, что, равномерно по точкам многообразия, случайное число $N_\varepsilon(X)$ точек в окрестности $U_{X,\varepsilon}$ меньше по порядку чем его среднее значение $\mathbb{E}N_\varepsilon(X) = O(N\varepsilon^q)$ (5), стремится к нулю с экспоненциальной скоростью.

**Теорема 2** *(равномерные большие уклонения).* Для достаточно малого $\varepsilon \leq \varepsilon_0, \varepsilon_0 > 0$, произвольного положительного числа $z \in (0, 1/4)$ и всех точек $X \in \mathbb{M}_\varepsilon$, отдаленных на $\varepsilon$ от края многообразия, имеет место неравенство

$$P\left(\inf_{X \in \mathbb{M}_\varepsilon} \frac{N_\varepsilon(X)}{N\varepsilon^q V_q} \leq p_{\min} \cdot (1 - z) - \varepsilon^2 \cdot \frac{C_0 \cdot p_{\max}}{V_q p_{\min}}\right)$$
$$\leq \left(\frac{C\sqrt{p}}{\varepsilon}\right)^p \cdot \exp\left(-4z^2 \cdot N\varepsilon^q V_q p_{\min}^2/(9p_{\max})\right),$$

где $C_0 > 0$ — константа из Теоремы 1, $C > 0$ и $\varepsilon_0 > 0$ — константы, $p_{\min} > 0, p_{\max}$ — минимальное и максимальное значения плотности $p_\mu$ на многообразии $\mathbb{M}$.

Утверждение теоремы **Теоремы 2** гарантирует, что одновременно для всех отделенных от края точек многообразия $\mathbb{M}_\varepsilon$ число точек выборки $\mathbb{X}_N$, попавших в их окрестность, будет не меньше фиксированной неограниченной функции от размера выборки. Этот результат является новым. Он используется в работе для анализа асимптотических свойств глобальных процедур оценивания в задачах моделирования многообразий. Поведение статистик на краю многообразия в диссертации не изучается, так как их вклад в глобальные оценки асимптотически мал. Отметим, что в **Теореме 2** нелинейность многообразия влияет на константу $C_0$ в верхней оценке уклонения, а также на порядок и коэффициенты полиномиального множителя в правой части.

В разделе **1.3** вводятся некоторые используемые в работе определения из дифференциальной геометрии, а также формулируются в удобном для дальнейшего использования известные леммы, связывающие носитель $\mathbb{M}$ с его касательным пространством, и теоремы Муавра-Лапласа для медленно убывающей к нулю вероятности успеха. В разделе **1.4** доказаны Теоремы **1-2**. В приложении **A** даны развернутые комментарии к используемым конструкциям из дифференциальной геометрии. В приложении **B** доказаны Леммы из раздела **1.3**.

**Вторая глава** диссертации посвящена поведению непараметрических локальных и глобальных (интегральных) оценок специального вида на многообразии.

**Постановка задачи.** В разделе **2.1** вводится класс рассматриваемых в диссертации локальных и глобальных статистик и формулируются сделанные предположения.

Рассматриваемые локальные статистики для фиксированной точки $X$ имеют вид

$$F_N(X) = \frac{\sum_{n=1}^{N} K_\varepsilon(X, X_n) \cdot F(X, X_n)}{\varepsilon^d \cdot \sum_{n=1}^{N} K_\varepsilon(X, X_n)}, \tag{7}$$

где функция $F(X, X_n)$ определяется решаемой статистической задачей, а ядерная функция (ядро) $K_\varepsilon(X, X_n)$ играет роль веса для слагаемого $F(X, X_n)$ такого, $K_\varepsilon(X, X_n) > 0$ только для точек $X_n$ из $\varepsilon$-окрестности $U_{X,\varepsilon}$ точки $X$, здесь $d \in \mathbb{N} \cup \{0\}$ — числовой параметр, определяемый поведением функции $F(X, X_n)$ при $X_n \to X$.

В качестве глобальных статистик рассматриваются усреднения локальных статистик (7)

$$F_N = \frac{1}{N} \sum_{n=1}^{N} F_N(X_n). \tag{8}$$

по всем точкам выборки.

Примеры ядерных функций $K_\varepsilon(X, X')$, используемые в различных статистиках при *моделировании многообразий*:

- индикатор $I(X' \in U_{X,\varepsilon})$ в алгоритмах Local Linear Embedding[17] и Local Tangent Space Alignment[18];
- heat-ядро $\exp(-|X' - X|^2/T) \cdot I(X' \in U_{X,\varepsilon})$ в алгоритме Laplacian Eigenmaps[19], где $T > 0$ — параметр "температуры";
- ядро Епанечникова $\left(1 - |X' - X|^2/\varepsilon^2\right) \cdot I(X' \in U_{X,\varepsilon})$ в алгоритме Vector Diffusion Maps[14];
- ядро $K_{GSE}(X, X')$ в алгоритме Grassmann&Stiefel Eigenmaps[21], которое зависит не только от расстояния $|X - X'|$, но и от расстояния Бине-Коши между касательными пространствами $T_X(\mathbb{M})$ и $T_{X'}(\mathbb{M})$ к $\mathbb{M}$ в точках $X$ и $X'$.

При фиксированной точке $X$, представим точку $X' \in U_{X,\varepsilon}$ через её локально римановы координаты: $X' = \exp_X(t\theta)$, где $t \in [0, \infty)$ и $\theta \in S_{q-1} \subset T_X(\mathbb{M})$, где $S_{q-1}$ — единичная сфера в $q$-мерном пространстве. В работе введено параметрическое семейство гладких финитных ядер $\{K(X, \theta, t), t \geq 0, \theta \in S_{q-1}\}$, удовлетворяющих введенным в диссертации предположениям $K1-K4$, и рассматриваются ядерные функции $K_\varepsilon(X, X')$ представимые в виде

$$K_\varepsilon(X, X') = K_\varepsilon(X, \exp_X(t\theta)) = K(X, \theta, t/\varepsilon). \tag{9}$$

В диссертации, как и в других теоретических исследованиях статистик на многообразиях, предполагается, что ядра являются гладкими функциями своих аргументов. Однако в приведенных выше примерах ядра хотя и допускают представление (9) но являются разрывными. Однако эти ядра можно заменить их "сглаженными" аналогами, с сохранением всех других процедур в использующих их алгоритмах, что обычно мало влияет на основные характеристики алгоритмов[26]. Примером такого удовлетворяющего предположениям диссертации гладкого финитного ядра является ядро $K(X, \theta, t) = \exp\left(\frac{1}{t^2-1}\right) \cdot I\left(t \in (-1, 1)\right)$.

Примеры функций $F(X, X')$ используемых в локальных статистиках вида (7):

1. $F(X, X') = \psi(X')$, тогда статистика $F_N(X)$ (7) является непараметрической оценкой[27] значения $\psi(X)$ в точке $X$ неизвестной заданной на многообразии функции $\psi$ по её известным значениям $\psi(X')$ в точках выборки $X' \in \mathbb{X}_N$;

[26]Bishop C.M. Pattern Recognition and Machine Learning. Springer, 2006.

[27]Wasserman L. All of Nonparametric Statistics. Springer, 2006.

2. $F(X, X') = (X' - X) \cdot (X' - X)^T$, $X, X' \in \mathbb{R}^p$, тогда $p \times p$ матрица $F_N(X)$ (7) является оценкой ковариационной матрицы случайной точки $X' \in \mathbb{M}$ попавшей в окрестность $U_{X,\varepsilon}$ точки $X \in \mathbb{M}$. Эта оценка используется в локальном методе главных компонент для построения оценки касательного пространства $T_X(\mathbb{M})$ к многообразию $\mathbb{M}$ в точке $X$[14];

3. $F(X, X') = |\psi(X) - \psi(X')|^2$, где $\psi$ — скалярная или векторная функция, тогда локальная статистика $F_N(X)$ (7) используется в процедурах снижения размерности[19,21].

Параметр $d$ определяется порядком малости функции $F(X, X') = F(X, \exp_X(t\theta)) = \Phi(X, \theta, t) = O(t^d)$ при $t \to 0$:

$$\Phi(X, \theta, t) = t^d \cdot \phi(X, \theta, t),$$

где функция $\phi \not\equiv 0$ ограничена, предполагается что величина $d$ не зависит от точки $X$ и $\theta$.

**Результаты.** В разделе **2.2** приводятся основные результаты главы, устанавливающие асимптотическое поведение рассматриваемых статистик при неограниченном возрастании объема выборки $N$.

Все результаты главы получены в сформулированных в диссертации предположениях $M1 - M10$, $S1 - S2$, $P1 - P3$, $F1 - F2$, $K1 - K5$. В автореферате приведем лишь результаты для четного $d$ (в диссертации даны результаты для произвольного $d$).

Обозначим

$$\overline{F}(X) = \frac{\int_{S^{q-1}} \rho_{E,d}(X, \theta)\Phi(X, \theta, t)d\theta}{\int_{S^{q-1}} \rho_{E,0}(X, \theta)d\theta}, \tag{10}$$

и

$$\overline{F} = \int_{\mathbb{M}} \overline{F}(X)d\mu(X), \tag{11}$$

где функция $\rho_{E,m}(X, \theta)$ зависит от семейства ядер и целочисленного параметра $m$.

Доказано, что имеет место состоятельность статистик $F_N(X)$ (7) и $F_N$ (8): статистика $F_N(X)$ стремится к $\overline{F}(X)$ (10) для любой точки $X \in \mathbb{M}$ и статистика $F_N$ стремится к $\overline{F}$ (11) при $N \to \infty$ по вероятности. В литературе в явном виде эти утверждения не формулировались и не доказывались, поэтому для возможности их дальнейшего использования в диссертации приведены их доказательства.

**Теорема 3** При сформулированных в диссертации предположениях,

для любой точки $X \in \mathbb{M}$ при $N \to \infty$:

$$\begin{aligned} \mathbb{E}F_N(X) &\to \overline{F}(X); \\ N\varepsilon^q \cdot \mathtt{Var}F_N(X) &\to d(X); \\ \sqrt{N\varepsilon^q} \cdot \left(F_N(X) - \overline{F}(X)\right) &\to^D N(0, d(X)), \end{aligned}$$

где $d(X)$ — выписанная в диссертации функция, $\to^D$ — сходимость по распределению.

Новой также является **Теорема 3** о вероятностях больших уклонений статистики $F_N(X)$, ранее такой результат был известен только для одной конкретной статистики (оценки ковариационной матрицы[14]).

**Теорема 4** При выполнении сформулированных в диссертации предположений, существуют $N_0, \varepsilon_0 > 0$ для любого $N > N_0$ и $\varepsilon < \varepsilon_0$ и любой удаленной от его края не менее чем на $\varepsilon$ точки многообразия $X \in \mathbb{M}_\varepsilon$, выполнено:

$$\begin{aligned} P\left(\left|F_N(X) - \overline{F}(X)\right| \geq z + \varepsilon^2 \cdot C_1\right) \leq \\ \leq 4 \cdot \exp\left(-z^2 \cdot N\varepsilon^q \cdot C_2\right), \end{aligned}$$

где $N_0, \varepsilon_0, C_1, C_2$ — некоторые положительные константы.

Равномерный (по $X \in \mathbb{M}$) аналог **Теоремы 4** о вероятностях больших уклонений $F_N(X)$ является новым и сформулирован в **Теореме 5**.

**Теорема 5** При выполнении сформулированных в диссертации предположений, для любого $N > N_0$ и $\varepsilon < \varepsilon_0$ для любой удаленной от его края точки многообразия $X \in \mathbb{M}_\varepsilon$, выполнено:

$$\begin{aligned} P\left(\left|F_N(X) - \overline{F}(X)\right| \geq z + \varepsilon^2 \cdot C_1\right) \leq \\ \leq 4 \cdot \left(\frac{C\sqrt{p}}{\varepsilon^3}\right)^p \cdot \exp\left(-z^2 \cdot N\varepsilon^q \cdot C_2\right), \end{aligned} \tag{12}$$

где $N_0, \varepsilon_0, C_1, C_2, C$ — положительные константы.

Отметим, что "платой за равномерность" является наличие в правой части формулы (12) дополнительного множителя $\left(\frac{C\sqrt{p}}{\varepsilon^3}\right)^p$, который при $N \to \infty$ растет медленнее основного экспоненциально убывающего множителя $\exp\left(-z^2 \cdot N\varepsilon^q \cdot C_2\right)$.

Теорема 6 о вероятностях больших уклонений статистики $F_N$ (8) также является новой.

**Теорема 6** При выполнении сформулированных в диссертации предположений, найдутся положительные $N_1$, $C_3$, $C_4$ такие, что для любого $z \in [0, 1]$ и $N > N_1$:

$$P\left(\left|F_N - \overline{F}\right| \geq z + \varepsilon^2 \cdot C_3\right) \leq \exp\left(-z^2 \cdot N\varepsilon^q \cdot C_4\right).$$

Доказательства **Теорем 3-6** приведены в разделах **2.3** и **2.4**.

**Третья глава** диссертации посвящена изучению соответствия решений выборочных оптимизационных задач на многообразиях с их непрерывными интегральными аналогами.

**Постановка задачи** (раздел **3.1**). Пусть $\mathbf{F} = \{\phi\}$ — гильбертово пространство функций определенных на многообразии $\mathbb{M}$ со скалярным произведением $(\phi_1, \phi_2) = \int_{\mathbb{M}} \phi_1(X) \cdot \phi_2(X)\mu(dX)$, $\mu$ — вероятностная мера на $\mathbb{M}$, удовлетворяющие сформулированным в диссертации условиям регулярности. Пусть $L$ — действующий на $\mathbf{F}$ линейный самосопряженный оператор Гильберт-Шмидта с неположительными собственными значениями.

Рассмотрим задачу максимизации функционала

$$\int_{\mathbb{M}} \phi(X) \cdot L\phi(X) d\mu(X), \tag{13}$$

по нормированным функциям $\phi \in \mathbf{F}$ ($\|\phi\|_\mu^2 = \int_{\mathbb{M}} |\phi(X)|^2 d\mu(X) = 1$) ортогональным собственному подпространству $\mathbf{F}_0$ оператора $L$, например ядру $Ker(L) = \{\phi \in \mathbf{F} : L\phi = 0\}$ оператора $L$. Очевидно, что решение этой задачи даётся собственной функцией $\phi^*(X)$ оператора $L$ с наибольшим ненулевым собственным значением $\lambda^*$.

Такие задачи естественным образом возникают в моделировании многообразий. Например, задача минимизации функционала $\int_{\mathbb{M}} |\nabla\phi(X)|^2 d\mu(X)$ по нормированным функциям $\phi$, естественным образом возникающая при построении низкоразмерной параметризации многообразия данных (снижении размерности)[19,21], сводится к решению рассматриваемой оптимизационной задачи с оператором Лапласа-Бельтрами $L = \triangle_{LB}$, являющимся обобщением стандартного оператора Лапласа на случай многообразия.

В статистической постановке, многообразие $\mathbb{M}$ и мера $\mu$ неизвестны, и исследуется задача оценивания $\phi^*$ по конечной выборке $\mathbb{X}_N$, состоящей из точек многообразия $\mathbb{M}$ случайно и независимо друг от друга выбранных в $\mathbb{M}$ в соответствии с вероятностной мерой $\mu$.

Оптимизируемый функционал (13) квадратичен по $\phi$, поэтому его выборочный аналог естественно также строить в виде квадратичной формы от вектора $\overrightarrow{\phi}_N = (\phi(X_1), \ldots, \phi(X_N))^T$, состоящего из неизвестных значений функции $\phi$ в точках выборки $\mathbb{X}_N$: интеграл (13) заменяется квадратичной оценкой

$$\int_{\mathbb{M}} \phi(X) \cdot L\phi(X) d\mu(X) \approx \frac{1}{N} \sum_{n=1}^{N} \phi(X_n) \cdot L\phi(X_n) \approx \overrightarrow{\phi}_N^T \cdot W_N \cdot \overrightarrow{\phi}_N,$$

в последнем члене вектор, состоящий из величин $\{(L\phi)(X_n), n = 1, 2, \ldots, N\}$ заменен оценкой вида $W_N \cdot \overrightarrow{\phi}_N$, в котором $W_N$ — некоторая

явно построенная по оператору $L$ матрица размера $N \times N$. В диссертации рассматривается конкретный способ[13,19] построения такой матрицы, обеспечивающий состоятельность этой оценки.

Рассмотрим задачу максимизации квадратичной формы

$$\overrightarrow{\phi}_N^T \cdot W_N \cdot \overrightarrow{\phi}_N \tag{14}$$

по вектору $\overrightarrow{\phi}_N$ при условиях нормировки $\frac{1}{N}|\overrightarrow{\phi}_N|^2 = 1$ и ортогональности ядру матрицы $Ker(W_N) = \{\overrightarrow{\phi}_N \in \mathbb{F} : W_N \overrightarrow{\phi}_N = \mathbf{0}_N\}$, где $\mathbf{0}_N$ — вектор из $N$ нулей. Решение этой задачи даётся собственным вектором $\overrightarrow{\phi}^*_N = (\phi^*_N(X_1), \phi^*_N(X_2), \ldots, \phi^*_N(X_N))^T$ матрицы $W_N$ с наибольшим ненулевым собственным значением $\nu^*_N$.

По построенным оценкам $(\phi^*_N(X_1), \phi^*_N(X_2), \ldots, \phi^*_N(X_N))$ значений неизвестной функции $\phi^*(X)$ в точках выборки, с использованием стандартных методов ядерного непараметрического оценивания[29], можно построить оценку $\phi^*_N(X)$, функции $\phi^*(X)$ в произвольной точке $X$. В диссертации рассматривается конкретная оценка

$$\hat{\phi}_N(X) = \frac{\sum_{n=1}^N K_\varepsilon(X, X_n) \cdot \phi^*_N(X_n)}{\sum_{n=1}^N K_\varepsilon(X, X_n)}, \tag{15}$$

в которой ядро $K_\varepsilon(X, X')$ описано в Главе 2.

**Полученные результаты.** Сформулированы точные предположения, которым удовлетворяет оператор $L$, которым, в частности, удовлетворяет взвешенный оператор Лапласа-Бельтрами $a(X) \cdot \triangle_{LB}$ с ограничением на максимальные собственные занчения, где $a(X)$ — произвольная гладкая скалярная функция. Частному случаю $a(X) = 1$ соответствует оценка

$$\overrightarrow{\phi}_N^T \cdot W_N \cdot \overrightarrow{\phi}_N = -\frac{\sum_{n=1}^N K_\varepsilon(X, X_n) \cdot (\phi(X_n) - \phi(X))^2}{\sum_{n=1}^N K_\varepsilon(X, X_n)} \cdot \frac{q \cdot (q+1)}{2 \cdot V_{q-1} \cdot \varepsilon^2},$$

используемая в алгоритме спектральных вложений Лапласа[19].

**Основные результаты** главы приведены в разделе **3.2**, их доказательства содержатся в разделе **3.3**.

**Теорема 7** *(сходимость спектров).* Для собственных чисел $\lambda_m, m > 0$, оператора $L$, упорядоченных в порядке неубывания, и собственных чисел $\nu_{1,N} \leq \ldots \nu_{N,N}$ матрицы $W_N$, имеет место сходимость по вероятности при $N \to \infty$:

$$\sup_{m \leq N} |\lambda_m - \nu_{m,N}| \to^P 0.$$

**Теорема 8** *(сходимость собственных функций).* Для произвольного ненулевого собственной функции $\phi^*(X)$ оператора $L$ с собственным числом

$\lambda_m, m > 0$ и функций $\hat{\phi}_N(X)$ (15), построенных по собственным векторам матрицы $W_N$, соответствующим собственным числам $\nu_{m,N}$, при $N \to \infty$ равномерно по $m$ имеет место сходимость по вероятности:

$$\sup_{X \in \mathbb{M}} |\hat{\phi}_N(X) - \phi^*(X)| \to^P 0.$$

Аналог этих теорем, соответствующий случаю фиксированной (не стремящейся к нулю) ширины ядра $\varepsilon$, ранее был получен в работах[13,19].

**Теорема 9** *(большие уклонения для спектра).* Для собственных чисел $\lambda_m, m > 0$ оператора $L$, упорядоченных в порядке неубывания, и собственных чисел $\nu_{1,N} \geq \ldots \nu_{N,N}$ матрицы $W_N$, существуют числа $N_0$ и $\delta_0 > 0$ такие, что для всех $N > N_0$ и $\alpha \in [0, \delta_0]$ имеет место неравенство:

$$P\left(\sup_m |\lambda_m - \nu_m| > \alpha \cdot C_6/\sqrt{N\varepsilon^q} + C_7 \cdot \varepsilon^2\right) \leq Q(N) \cdot \exp(-C_8 \cdot \alpha^2 \cdot N\varepsilon^q),$$

где $C_6, C_7, C_8$ — положительные константы, а $Q(N)$ — полином от $N$.

**Теорема 10** *(большие уклонения для собственных функций).* Для всех ненулевых собственных чисел $\lambda_m$ оператора $L$ и соответствующих им собственных функций $\phi^*_m(X)$ и соответствующих последовательностей $\nu_{m,N}$ собственных чисел матриц $W_N$ и функций $\hat{\phi}_{m,N}(X)$ (15), для всех точек, отделенных на $\varepsilon$ от границы многообразия $X$ найдутся числа $N_0$ и $\delta_0 > 0$ такие, что для всех $N > N_0$:

$$P\left(\sup_m \sup_{X \in \mathbb{M}_\varepsilon} |\hat{\phi}_{m,N}(X) - \phi^*_m(X)| > C_9/\sqrt{N\varepsilon^q} + C_{10} \cdot \varepsilon^2\right) \leq$$
$$\leq Q(N) \cdot \exp(-C_{11} \cdot \alpha^2 \cdot N\varepsilon^q),$$

где $C_9, C_{10}, C_{11}$ — положительные константы, а $Q(N)$ — полином от $N$.

Ранее подобные задачи исследовались лишь при фиксированной ширине ядра $\varepsilon$ (в диссертации рассматривается случай $\varepsilon \to 0$) и ядерной функции $K_\varepsilon(X, X')$ строго отделенной от нуля, что позволяло доказать лишь сходимость решения статистической процедуры оценивания к решению сглаженного аналога исходной задачи (13) (проинтегрированного по $\varepsilon$-шару решения рассматриваемой в диссертации задачи), а не сходимость к исходному решению, доказанную в **Теремах 8 и 10**.

**Четвертая глава** диссертации содержит примеры конкретных статистических процедур моделирования многообразий[10,21], в которых были использованы получены в диссертации результаты. В частности, приведенные в диссертации результаты были использованы для выбора ширины ядра и исследования свойств процедур. Например, в диссертации доказана сходимость области определения алгоритма GSE к неизвестному многооб-

разию данных (**Теорема 11**), являющаяся следствием доказанной в **Главе 1 Теоремы 2**. Материалы данной главы (кроме **Теоремы 11**) не выносятся на защиту, и носят иллюстративный характер.

В **заключении** к диссертации представлены основные результаты работы и возможные дальнейшие темы исследований.

## Заключение

В диссертации получены следующие основные результаты:

1. Исследованы асимптотические свойства числа точек $N_\varepsilon(X)$ выборки, попавших в асимптотически малые $\varepsilon$-окрестности точек $X$ многообразия (состоятельность, асимптотическое разложение и большие уклонения биномиального случайного поля $N_\varepsilon(X)$), а также доказана асимптотическая равномерность этих точек в рассматриваемой асимптотически малой окрестности. Доказана равномерность по многообразию результатов о больших уклонениях биномиального случайного поля $N_\varepsilon(X)$ от своего предельного значения;

2. Доказана состоятельность, найдено асимптотическое разложение и получены теоремы о больших уклонениях, для рассмотренного класса локальных статистик $F_N(X)$, доказана равномерность этих результатов по точкам многообразия;

3. Рассмотрен класс глобальных статистик $F_N$, являющихся усреднениями локальных статистик по точкам выборки. Доказана их состоятельность и получены теоремы о больших уклонениях;

4. Для рассмотренного класса непрерывных функционалов от определенных на многообразии функций и их выборочных конечномерных матричных аналогов. Доказана сходимость собственных чисел и собственных функций матричных аналогов к собственным числам и собственным функциям предельных функционалов, получены равномерные теоремы о больших уклонениях.

## Публикации автора по теме диссертации

Публикации в журналах, входящих в перечень ВАК


1. *Yanovich Yu.* Asymptotic Properties of Local Sampling on Manifold // Journal of Mathematics and Statistics — 2016. Vol. 12. Issue 3. P. 157-175. DOI: 10.3844/jmssp.2016.157.175.

2. *Yanovich Yu.* Asymptotic Properties of Eigenvalues and Eigenfunctions of Linear Operators on Manifolds // Lobachevskii Journal of Mathematics, Issue 3, 2017.

3. *Yanovich Yu.* Asymptotic Properties of Nonparametric Estimation on Manifold // Proceedings of Machine Learning Research, Volume 60, 2017. P. 18-38.

4. *Bernstein A., Kuleshov A. P., Yanovich Y.* Information preserving and locally isometric&conformal embedding via Tangent Manifold Learning // Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on. P.: IEEE — 2015. P. 1-10. DOI: 10.1109/DSAA.2015.7344815.

5. *Bernstein A., Kuleshov A. P., Yanovich Y.* Locally isometric and conformal parameterization of image manifold // Proceedings of SPIE 9875, Eighth International Conference on Machine Vision (ICMV 2015). Barselona: SPIE — 2015. P. 1-7. DOI: 10.1117/12.2228741.

6. *Bernstein A., Kuleshov A. P., Yanovich Y.* Manifold Learning in Regression Tasks // Lecture Notes in Computer Science. Vol. 9047 — 2015. P. 414-423. 10.1007/978-3-319-17091-6_36.

7. *Bernstein A., Kuleshov A. P., Yanovich Y.* Statistical learning via manifold learning, in: Proceedings 2015 IEEE 14th International Conference on Machine Learning and Applications, ICMLA — 2015. P. 64-69. DOI: 10.1109/ICMLA.2015.26.

Другие публикации

8. *Bernstein A.V., Kuleshov A.P., Yanovich Yu.A.* Locally Isometric and Conformal Low-dimensional Data Representation in Data Analysis. Abstracts of Yandex School of Data Analysis Conference “Machine Learning: Prospects and Applications”, pp. 87–88, 2015. Available online.

9. *Bernstein A.V., Kuleshov A.P., Yanovich Yu.A.* Nonparametric algorithm for Tangent Bundle Manifold Learning problem. Twenty-Seventh Annual Conference on Neural Information Processing Systems (NIPS-2013), Workshop ’Modern Nonparametric Methods in Machine Learning’, USA, Nevada, Lake Tahoe, 5-10 of December, 2013. Available online.

10. *Bernstein A., Kuleshov A. P., Yanovich Y.* Asymptotically Optimal Method for Manifold Estimation Problem, in: Abstructs of the 29-th European Meeting of Statisticians. Budapest : Bernoulli Society — 2013. P. 325.

11. *Bernstein A., Kuleshov A. P., Yanovich Y.* Asymptotically Optimal Method for Manifold Estimation Problem. Abstracts of the International Workshop on Statistical Learning, June 26-28, Moscow, 2013, pp. 8-9, 2013.

12. *Янович Ю. А.* Состоятельность оценки области определения алгоритмом спектральных вложений Грассмана-Штифеля // Сборник статей конференции "Информационные технологии и системы"(ИТиС'16). М. : ИППИ РАН — 2016. С. 191-197.

13. *Янович Ю. А., Киселиус М.* Генерация последовательностей случайных точек с заданной плотностью на многообразиях // Сборник статей конференции "Информационные технологии и системы"(ИТиС'15). М : ИППИ РАН — 2015. С. 1036-1040.

14. *Янович Ю. А.* Равномерное оценивание касательного к многообразию пространства // В кн.: Сборник статей конференции "Информационные технологии и системы"(ИТиС'13). М.: ИППИ РАН — 2013. С. 371-375.