МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
имени М.В. ЛОМОНОСОВА

На правах рукописи

Палионная Софья Игоревна

Асимптотические свойства оценок риска в задачах множественной проверки гипотез

1.1.4— теория вероятностей и математическая статистика

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата физико-математических наук

Москва — 2023

Диссертация подготовлена на кафедре математической статистики факультета вычислительной математики и кибернетики МГУ имени М.В.Ломоносова.

Научный руководитель: **Шестаков Олег Владимирович**
доктор физико-математических наук, доцент

Официальные оппоненты: **Бурнаев Евгений Владимирович**
доктор физико-математических наук, доцент, Сколковский институт науки и технологий, профессор, руководитель исследовательского центра в сфере искусственного интеллекта по направлению оптимизации управленческих решений в целях снижения углеродного следа

Зорин Андрей Владимирович
доктор физико-математических наук, доцент, ННГУ имени Н.И. Лобачевского, заведующий кафедрой теории вероятностей и анализа данных института информационных технологий, математики и механики

Колчин Андрей Валентинович
кандидат физико-математических наук, доцент, Московский автомобильно-дорожный государственный технический университет (МАДИ), доцент кафедры «Прикладная математика» факультета управления

Защита диссертации состоится «28» июня 2023 г. в 15 часов 00 минут на заседании диссертационного совета МГУ.011.3 Московского государственного университета имени М.В.Ломоносова по адресу: Российская Федерация, 119991, ГСП-1, Москва, Ленинские горы, д. 1, МГУ, механико-математический факультет, аудитория 16-10.

E-mail: mexmat_disser85@mail.ru

С диссертацией можно ознакомиться в отделе диссертаций научной библиотеки МГУ имени М.В. Ломоносова (Ломоносовский просп., д. 27) и на портале: https://dissovet.msu.ru/dissertation/011.3/2518

Автореферат разослан «27» мая 2023 г.

Заместитель председателя диссертационного совета, доктор физико-математических наук, доцент — И.С. Ломов

Ученый секретарь диссертационного совета, доктор физико-математических наук, доцент — Н.А. Раутиан

Общая характеристика работы

Актуальность темы. В современном мире, где всё большую значимость приобретают задачи обработки больших данных, количество компонент-предикторов в моделях данных может быть очень велико, что существенно затрудняет работу с ними. Поэтому прежде чем перейти к задаче фильтрации шума в такой модели, необходимо преобразовать исходные данные так, чтобы выделить значимые признаки и удалить незначимые. Такое преобразование позволит привести наблюдаемые данные к «экономному» представлению. При этом важную роль при решении данной задачи играет универсальность выбранного метода, т.е. метод должен быть адаптивен к различным входным данным и моделям и в каждом случае давать результат, близкий к оптимальному.

Для того, чтобы осуществить сжатие данных можно прибегнуть к пороговой обработке, которая по сути эквивалентна задаче множественной проверки гипотез для набора коэффициентов разложения исходных данных: если нулевая гипотеза отклоняется, то соответствующая компонента разложения данных зануляется. В случае построения семейства статистических выводов возникает эффект множественных сравнений. Существуют различные методы решения задачи множественной проверки гипотез, устраняющие эффект множественных сравнений[1]. Суть этих методов заключается в контроле меры, обобщающей ошибку первого рода при множественной проверке гипотез. К этим мерам относятся FWER (Family-Wise Error Rate), FDR (False Discovery Rate) , pFDR (positive False Discovery Rate), HMP (harmonic mean p-value) (см. работы[2,3,4,5]). В данной диссертации будет использоваться FDR-мера, которая представляет собой долю ложных отклонений нулевых гипотез. Широкое распространение FDR-мера получила в случаях, когда количество проверяемых гипотез насколько велико, что предпочтительнее совершить небольшое количество ошибок первого рода для увеличения статистической мощности.

В рамках задачи пороговой обработки ключевую роль играет выбор порогового значения, который будет гарантировать построение в некотором смысле наилучшей модели, что приводит к необходимости вычислять

[1] *Miller R.* Simultaneous Statistical Inference. 2nd. New York : Springer Verlag, 1981. 272 p.

[2] *Storey J. D.* A direct approach to false discovery rates // Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2002. Vol. 64, no. 3. P. 479—498.

[3] *Benjamini Y., Hochberg Y.* Controlling the false discovery rate: a practical and powerful approach to multiple testing // Journal of the royal statistical society series b-methodological. 1995. Vol. 57. P. 289—300.

[4] *Wilson D. J.* The harmonic mean p-value for combining dependent tests // Proceedings of the National Academy of Sciences of the United States of America. 2019. Vol. 116. P. 1195—1200.

[5] *Wilson D. J.* Reply to Held: When is a harmonic mean p-value a Bayes factor? // Proceedings of the National Academy of Sciences. 2019. Vol. 116. P. 5857—5858.

погрешность модели после применения пороговой обработки. На практике получить значение самой погрешности (или риска) не представляется возможным, т.к. она зависит от неизвестных значений исходных данных. Однако можно рассмотреть несмещенные оценки риска, которые носят название SURE-оценок (Stein Unbiased Risk Estimator). Для высчисления таких оценок требуются только наблюдаемые данные, что позволяет вычислять эти оценки на практике. В работах[6,7,8] были предложены способы вычисления порогового значения, направленные на минимизацию среднеквадратичной погрешности (риска). В данной диссертации задача сжатия данных будет сведена к задаче пороговой обработки с использованием FDR-порога, контроль над которым будет осуществляться с помощью алгоритма Бенжамини-Хочберга[9].

Во многих областях знаний после применения пороговой обработки сигнала оказывается, что лишь небольшое количество компонент-предикторов исходной модели значимо отличны от нуля. В частности, такие модели рассматриваются в задачах из различных областей компьютреного зрения, аудио- и видео- обработки данных, обработки электроэнцефалограмм и т.д. Эти наблюдения приводят к необходимости рассматривать в некотором смысле разреженный сигнал исходных данных. В данной диссертации будут рассмотрены несколько вариантов определения разреженности сигнала.

Кроме того, зачастую информация, доступная для наблюдения, представляет собой некоторое преобразование исходных данных. Такие ситуации возникают, например, в астрофизических и томографических приложениях, физике плазмы и др.[10,11,12]. В этом случае дополнительно возникает задача обращения преобразования. Во второй главе данной диссертации будет рассмотрена постановка задачи в случае, когда исходный вектор данных подвергается действию линейного однородного оператора. При этом существование и ограниченность обратного оператора, вообще говоря, не гарантируется, что делает невозможным применение обратного оператора к наблюдаемым данным для получения исходного сигнала.

---

[6] *Donoho D., Johnstone I. M.* Ideal spatial adaptation via wavelet shrinkage // Biometrika. 1994. Vol. 81, no. 3. P. 425—455.

[7] *Donoho D., Johnstone I. M.* Adapting to unknown smoothness via wavelet shrinkage // J. Amer. Statist. Assoc. 1995. Vol. 90. P. 1200—1224.

[8] *Donoho D., Johnstone I. M.* Minimax estimation via wavelet shrinkage // The Annals of Statistics. 1998. Vol. 26, no. 3. P. 879—921.

[9] См. сноску 3 выше

[10] *Kalifa J., Laine A., Esser P. D.* Tomographic reconstruction with non-linear diagonal estimators // Proceedings of SPIE, the International Society for Optical Engineering. 2000. Vol. 4119. P. 576—586.

[11] *Kalifa J., Mallat S.* Thresholding estimators for linear inverse problems and deconvolutions // The Annals of Statistics. 2003. Vol. 31, no. 1. P. 58—109.

[12] *Kalifa J., Laine A., Esser P. D.* A wavelet shrinkage approach to tomographic image reconstruction // Journal of the American Statistical Association. 1996. Vol. 91, no. 435. P. 1079—1090.

Для решения этой задачи будет использоваться аппарат вейвлет-анализа, который представляет собой нелинейный метод обработки сигнала. Выбор вейвлет-анализа обусловлен рядом преимуществ этого метода. Во-первых, вейвлет-разложение приводит исходные данные к разреженному представлению, что является ключевым предположением в постановке задачи диссертации. Во-вторых, вейвлет-анализ, в отличие, например, от анализа Фурье, позволяет эффективнее работать с нестационарными сигналами. Наконец, реализация вейвлет-анализа в различных библиотеках программного обеспечения позволяет получать практические результаты с хорошей вычислительной скоростью.

В частности, в диссертации рассмотрен метод вейглет-вейвлет-разложения, который был предложен в статье[13]. Этот метод предполагает представление преобразованных оператором исходных данных в виде ряда из сдвигов и растяжений некоторой выбранной вейвлет-функции. Основная идея применения вейглет-вейвлет-метода заключается в том, чтобы представить исходные данные в виде разложения по вейглет-базису и выразить исходные данные через коэффициенты разложения их преобразования. Также стоит отметить, что при применении вейглет-вейвлет-метода преобразованные исходные данные раскладываются в ряд по ортонормированному базису, что сохраняет независимость шумовых коэффициентов разложения. Однако при применении этого метода возникает необходимость вводить корректировку значения среднеквадратичного риска для каждого масштаба разложения, что приводит к необходимости вычисления значения порога на каждом уровне.

Цель работы. Целью данной диссертации является исследование асимптотических свойств оценки среднеквадратичной погрешности множественной проверки гипотез с использованием FDR-порога в задаче оценивания математического ожидания гауссова вектора в случае рассмотрения векторов большой размерности и в случае, когда исходный сигнал данных был преобразован действием линейного однородного оператора.

Научная новизна. Результаты, изложенные в диссертации, являются новыми и заключаются в следующем.

Доказана сильная состоятельность и асимптотическая нормальность оценки риска в задаче множественной проверки гипотез с использованием FDR-порога. Также в диссертации оценивается скорость сходимости распределения данной оценки риска к нормальному закону.

Рассмотрен случай, когда исходный сигнал был подвержен линейному однородному преобразованию. Для такой постановки задачи доказывается сильная состоятельность и асимптотическая нормальность оценки риска, а также оценивается скорость сходимости распределения данной оценки риска к нормальному закону.

[13] *Abramovich F., Silverman B. W.* Wavelet decomposition approaches to statistical inverse problems // Biometrika. 1998. Vol. 85, no. 1. P. 115—129.

Методы исследования. В диссертации используются методы теории вероятностей и математической статистики, а также математического анализа. Основным методом является сведение задачи пороговой обработки к задаче множественной проверки гипотез с FDR-мерой, для контроля над которой используется алгоритм Бенжамини-Хочберга. Рассматриваются случаи использования мягкой и жесткой пороговой обработки. Также рассматривается задача обращения линейного однородного оператора, для решения которой используется аппарат вейвлет-анализа.

Основные положения, выносимые на защиту:

1. Сильная состоятельность оценки риска при множественной проверке гипотез с FDR-порогом.
2. Асимптотическая нормальность оценки риска при множественной проверке гипотез с FDR-порогом.
3. Оценка скорости сходимости распределения оценки риска к нормальному закону с использованием FDR-метода множественной проверки гипотез.
4. Сильная состоятельность оценки риска при множественной проверке гипотез с FDR-порогом при обращении линейных однородных операторов.
5. Асимптотическая нормальность оценки риска при множественной проверке гипотез с FDR-порогом при обращении линейных однородных операторов.
6. Оценка скорости сходимости распределения оценки риска к нормальному закону с использованием FDR-метода множественной проверки гипотез при обращении линейных однородных операторов.

Соответствие паспорту научной специальности. В диссертации исследуются асимптотические свойства оценки риска и скорость сходимости распределения оценки риска в задаче множественной проверки гипотез, в силу чего диссертация соответствует паспорту специальности 1.1.4 «Теория вероятностей и математическая статистика».

Апробация работы. Основные результаты работы докладывались автором на следующих конференциях:

1. Научная конференция «Тихоновские чтения 2020», МГУ имени М.В.Ломоносова, факультет ВМК, Москва, Россия, 26-31 октября 2020 г.
   Тема доклада: Свойства оценки риска при множественной проверке гипотез с использованием FDR-метода
2. Ломоносовские чтения 2021. Секция вычислительная математика и кибернетика, Москва, Россия, 20-29 апреля 2021 г.
   Тема доклада: Асимптотическое поведение оценки риска FDR-метода в задаче множественной проверки гипотез

3. XXXVI International Seminar on Stability Problems for Stochastic Models, г.Петрозаводск, Россия, 21-25 июня 2021 г.
   Тема доклада: Asymptotic behavior of a risk estimate for the FDR-method in the problem of multiple hypothesis testing
4. Ломоносовские чтения - 2022, Секция вычислительная математика и кибернетика, МГУ имени М.В.Ломоносова, факультет ВМК, Москва, Россия, 14-22 апреля 2022 г.
   Тема доклада: Свойства оценок риска в задачах обращения линейных операторов при использовании FDR-метода множественной проверки гипотез

Публикации. Основные результаты по теме диссертации изложены в 5 статьях [1—5] автора. Статья [5] (в соавторстве) опубликована в рецензируемом научном журнале, входящем в базы Scopus и Web of Science, статья [4] (в соавторстве) опубликована в рецензируемом научном журнале, входящем в базу Scopus, статьи [1—3] опубликованы в рецензируемых научных журналах, входящих в базу RSCI. Также работы автора представлены в материалах конференций [6—9]. Список работ автора приведен в конце автореферата и диссертации. В этих работах постановки задач принадлежат О.В. Шестакову, а все основные результаты, приведенные в статьях и диссертации, получены С.И. Палионной самостоятельно.

Объем и структура работы. Диссертация состоит из введения, двух глав, заключения и списка литературы, включающего в себя 74 наименований. Общий объем диссертации составляет 83 страниц. В диссертацию вошли результаты, полученные в рамках программы Московского центра фундаментальной и прикладной математики.

В первой главе рассматривается задача оценивания математического ожидания гауссова вектора в случае, когда наблюдаемые данные зашумлены, а неизвестный вектор исходных данных принадлежит некоторому классу разреженности. Приводятся условия, при которых имеет место сильная состоятельность и асимптотическая нормальность оценки риска. Оценивается скорость сходимости распределения оценки риска к нормальному закону.

Во второй главе рассматривается задача оценивания математического ожидания гауссова вектора в случае, когда наблюдаемые данные зашумлены и представляют собой преобразованные исходные данные, подвергнутые действию линейного однородного оператора, а неизвестный вектор исходных данных принадлежит некоторому классу разреженности. Рассматривается задача обращения линейного однородного оператора, приводятся условия, при которых имеет место сильная состоятельность и асимптотическая нормальность оценки риска. Оценивается скорость сходимости распределения оценки риска к нормальному закону.

В заключении кратко приведены основные результаты диссертации.

Краткое содержание диссертации

В первой главе рассматривается линейная регрессионная модель, у которой количество компонент-предикторов значительно велико. Классической задачей в таком случае будет являться выделение значимых признаков для модели и удаление из рассмотрения шумовых.

Для начала рассмотрим задачу множественной проверки статистических гипотез. Имеется n различных выборок, каждой их которых соответствует своя нулевая гипотеза $\{H_{0_i}, i = 1, \ldots, n\}$ и альтернатива $\{H_{1_i}, i = 1, \ldots, n\}$. Гипотезы проверяются статистиками T_i с заданными нулевыми распределениями и вычисляются достигаемые уровни значимости $\{p_i, i = 1, \ldots, n\}$. На основе получаемых значений p_i принимается решение об отвержении нулевой гипотезы $\{H_{0_i}\}$ для каждого i. Обозначим через M_0 множество индексов верных нулевых гипотез, а через R множество индексов отвергаемых гипотез. Тогда $V = |M_0 \cap R|$ — число ошибок первого рода. Задача заключается в минимизации числа ошибок первого рода за счет изменения параметра R.

Существует множество статистических процедур, предлагающих различные методы решения задачи множественной проверки гипотез. В данной диссертации будет использоваться мера, получившая название FDR (False Discovery Rate)[14]. FDR мера заключается в контроле ожидаемой доли ложных отклонений, т.е.

$$\mathrm{FDR} = \mathbb{E}\left(\frac{V}{\max(R,1)}\right).$$

Для контроля FDR-меры чаще всего используется метод Бенжамини–Хочберга, который позволяет ограничить сверху ожидаемую долю ложных отклонений параметром α:

$$\mathbb{E}\left(\frac{V}{\max(R,1)}\right) \leq \alpha.$$

Для гипотез H_{0_i} уровни значимости α_i меняются линейно и определяются следующим образом:

$$\alpha_i = \frac{\alpha \cdot i}{n}, i = 1, \ldots, n.$$

Затем строится вариационный ряд из достигаемых уровней значимости p_i:

$$p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(n)}.$$

[14] См. сноску 3 выше

По построенному вариационному ряду находится $k \in [1,n]$ — максимальный индекс такой, что для него выполнено условие

$$p_{(i)} \leq \alpha_i,$$

где $p_{(i)}$ — i-й член вариационного ряда из достигаемых уровней значимости. И отвергаются все гипотезы $H_{0_1}, \ldots, H_{0_k}$.

В диссертации будет использован метод FDR, т.к. этот метод хорошо адаптирован к работе с большими объемами данных. Также в работе[15] была доказана связь между FDR-мерой и байесовским подходом. Помимо этого в работе[16] было доказано, что при использовании FDR-меры можно получить асимптотически минимаксные оценки риска для различного вида разреженных данных.

Теперь перейдем к рассмотрению задачи оценивания математического ожидания гауссова вектора:

$$X_i = \mu_i + \varepsilon_i, \;\; i = 1, \ldots, n, \tag{1}$$

где X_i - наблюдаемые зашумленные данные, ε_i — независимые нормально распределенные случайные величины с нулевым математическим ожиданием и известной дисперсией σ^2, а $\mu = (\mu_1, \ldots, \mu_n)$ — неизвестный вектор истинных данных, принадлежащий некоторому заданному классу "разреженности". Для каждого $i \in [1, n]$ рассмотрим нулевую гипотезу H_{0_i}, включающую в себя предположение о распределении X_i, а также о равенстве нулю μ_i.

Предположение о разреженности вектора является ключевым в диссертации и предполагает, что лишь небольшое количество компонент вектора значительно отличается от нуля. Такая постановка задачи возникает в различных областях знаний, где имеет место обработка и анализ сигналов, содержащих шум. При этом разреженное представление сигнала обеспечивается за счет предварительной обработки сигнала, в частности, применения дискретного вейвлет-преобразования.

В диссертации будут рассмотрены несколько определений "разреженности". Пусть $\|\mu\|_0 = \#\{\mu_i : \mu_i \neq 0\}$ обозначает число компонент μ, отличных от нуля. Фиксируя η_n, определим класс

$$L_0(\eta_n) = \{\mu \in \mathbb{R}^n : \|\mu\|_0 \leq \eta_n n\}.$$

При малых значениях η_n лишь небольшое число компонент вектора из $L_0(\eta_n)$ отлично от нуля.

[15] *Storey J. D.* The positive false discovery rate: a Bayesian interpretation and the q-value // The Annals of Statistics. 2003. Vol. 31. P. 2013—2035.

[16] Adapting to unknown sparsity by controlling the false discovery rate / F. Abramovich [et al.] // The Annals of Statistics. 2006. Vol. 34, no. 2. P. 584—653.

Другой возможный способ определения "разреженности" заключается в ограничении абсолютных значений компонент μ. Для этого рассмотрим упорядоченные по абсолютной величине значения

$$|\mu|_{(1)} \geq \ldots \geq |\mu|_{(n)}$$

и при $0 < p < 2$ определим класс

$$L_p(\eta_n) = \{\mu \in \mathbb{R}^n : |\mu|_{(k)} \leq \eta_n n^{1/p} k^{-1/p} \text{ для всех } k = 1, \ldots, n\}.$$

Также "разреженность" можно моделировать с помощью ℓ_p-нормы

$$\|\mu\|_p = \left(\sum_{i=1}^{n} |\mu_i|^p \right)^{1/p}.$$

В этом случае "разреженный" класс определяется как

$$M_p(\eta_n) = \{\mu \in \mathbb{R}^n : \sum_{i=1}^{n} |\mu_i|^p \leq \eta_n^p n\}.$$

Между этими классами существует взаимосвязь, отмеченная в работе[17]. Также отметим, что при $p \to 0$ справедливо $\|\mu\|_p^p \to \|\mu\|_0$. Помимо этого для рассматриваемых классов разреженности справедливо вложение

$$M_p(\eta_n) \subset L_p(\eta_n) \not\subset M_{p'}(\eta_n),\ p' > p.$$

Для построения оценки вектора μ при наличии шума воспользуемся методом пороговой обработки. Рассмотрим жесткую (hard) пороговую обработку для каждой компоненты вектора:

$$\hat{\mu}_i = \rho_H(X_i, T) = \begin{cases} X_i & \text{при } |X_i| > T, \\ 0 & \text{при } |X_i| \leqslant T, \end{cases} \tag{2}$$

т. е. компонента вектора обнуляется, если ее абсолютное значение не превосходит критического порога T. При использовании FDR-метода пороговое значение T выбирается по следующему правилу: по абсолютным значениям величин исходного вектора строится вариационный ряд

$$|X|_{(1)} \geq \ldots \geq |X|_{(n)},$$

и затем $|X|_{(k)}$ сравнивается с квантилями Гауссова распределения $t_k = \sigma z(\alpha/2 \cdot k/n)$. Пусть k_F — наибольший индекс k, для которого

[17] См. сноску 16 выше

$|X|_{(k)} \geq t_k$, тогда выбирается порог $T^F = t_{k_F}$. Применение жесткой пороговой обработки для компонент вектора μ в задаче (1) эквивалентно процедуре множественной проверки гипотез, описанной выше.

В сочетании с методами проверки гипотез также широко используется метод штрафов, при котором минимизируется невязка с добавлением штрафной функции[18,19,20]. В частном случае этот метод приводит к так называемой мягкой (soft) пороговой обработке, при которой оценки компонент вектора вычисляются по правилу

$$\hat{\mu}_i = \rho_S(X_i,T) = \begin{cases} X_i - T & \text{при } X_i > T, \\ X_i + T & \text{при } X_i < -T, \\ 0 & \text{при } |X_i| \leqslant T. \end{cases} \tag{3}$$

Данный подход в некоторых случаях оказывается более адекватным, чем (2), поскольку функция ρ_S в (3) непрерывна по X_i.

Как в случае мягкой пороговой обработки, так и в случае жесткой пороговой обработки основной задачей становится выбор стратегии для определения порогового значения T, при котором полученная модель будет в некотором смысле наилучшей. Зачастую такую задачу сводят к поиску порогового значения, минимизирующего среднеквадратичную погрешность.

Среднеквадратичная погрешность (или риск) рассмотренных процедур определяется следующим образом:

$$R(T) = \sum_{i=1}^{n} \mathbb{E}\left(\hat{\mu}_i - \mu_i\right)^2. \tag{4}$$

Методы выбора порогового значения T, как правило, ориентированы на минимизацию риска (4) при условии принадлежности вектора μ заданному классу. "Идеальным" значением порога является

$$T^{min} : \; R(T^{min}) = \min_T R(T).$$

Заметим, что в выражении (4) присутствуют неизвестные величины μ_i и вычислить значения $R(T)$ и T^{min} на практике не представляется возможным. Поэтому рассмотрим оценку среднеквадратичного риска, используемую в книге[21], которая определяется выражением

$$\hat{R}(T) = \sum_{i=1}^{n} F[X_i,T], \tag{5}$$

[18] См. сноску 16 выше

[19] *Donoho D. L., Jin J.* Asymptotic minimaxity of false discovery rate thresholding for sparse exponential data // The Annals of Statistics. 2006. Vol. 34. P. 2980—3018.

[20] *Neuvial P., Roquain É.* On false discovery rate thresholding for classification under sparsity // The Annals of Statistics. 2012. Vol. 40. P. 2572—2600.

[21] *Mallat S.* A Wavelet Tour of Signal Processing. N. Y. : Academic Press, 1999. 857 p.

где в случае жесткой пороговой обработки

$$F[X_i, T] = \begin{cases} X_i^2 - \sigma^2 & \text{при } |X_i| \leqslant T, \\ \sigma^2 & \text{при } |X_i| > T \end{cases} \tag{6}$$

и в случае мягкой пороговой обработки

$$F[X_i, T] = \begin{cases} X_i^2 - \sigma^2 & \text{при } |X_i| \leqslant T, \\ \sigma^2 + T^2 & \text{при } |X_i| > T. \end{cases} \tag{7}$$

Благодаря тому, что при вычислении оценки риска используются только наблюдаемые данные, можно получать значения оценки риска на практике. Следовательно, оценки риска позволяют сравнивать между собой модели, полученные при применении различных порогов, основываясь исключительно на наблюдаемых данных.

Много важных результатов, связанных с адаптацией различных порогов к неизвестной разреженности входных данных, было приведено в работе[22]. В частности, в этой работе была установлена асимптотическая минимаксность FDR оценки и тот факт, что минимаксность оценки FDR адаптивна к различным классам разреженности. В то же время в этой работе (с использованием результатов, полученных ранее в статье[23]) было продемонстрировано, что универсальный порог $T^U = \sigma\sqrt{2\log n}$ [24,25,26,27] плохо адаптируется к неизвестной разреженности данных.

Таким образом, хорошие асимптотические свойства FDR-порога делают целесообразным его применение при работе с большими объемами данных.

Для удобства дальнейшего изложения введем ряд дополнительных обозначений для различных классов разреженности: для $L_0(\eta_n)$

$$\gamma_n = \frac{1}{\log\log n},\ \kappa_n = \frac{n\,\eta_n}{1-\alpha_n-\gamma_n}, T^1 = \sigma\big(2\log\eta_n^{-1}\big)^{1/2}, \tag{8}$$

и для $L_p(\eta_n)$

$$\gamma_n = \frac{1}{\log\log n},\ \tau_\eta = \sigma\big(2\log\eta_n^{-p}\big)^{1/2}, \kappa_n = \frac{n\,\eta_n^p\tau_\eta^{-p}}{1-\alpha_n-\gamma_n}, T^1 = \sigma\big(2\log\eta_n^{-p}\big)^{1/2}. \tag{9}$$

---

[22] См. сноску 16 выше

[23] *Donoho D., Johnstone I. M.* Minimax risk over lp-balls for lq-error // Probability Theory and Related Fields. 1994. Vol. 99. P. 277—303.

[24] См. сноску 6 выше

[25] *Donoho D.* De-noising by soft-thresholding // IEEE transactions on information theory. 1995. Vol. 41, no. 3. P. 613—627.

[26] *Donoho D., Johnstone I. M.* Neo-classical minimax problems, thresholding and adaptive function estimation // Bernoulli. 1996. Vol. 2, no. 1. P. 39—62.

[27] Wavelet shrinkage: asymptopia? / D. Donoho [et al.] // Journal of the Royal Statistical Society, series B. 1995. Vol. 57. P. 301—369.

В теореме 1.5 из работы [1] была доказана сильная состоятельность оценки риска $\hat{R}(T)$.

Теорема 1.5 *Пусть* $\mu \in L_0[\eta_n]$, *где* $\eta_n \in [n^{-1}(\log n)^5, n^{-\gamma}]$ *или* $\mu \in L_p(\eta_n)$, $0 < p < 2$, *где* $\eta_n^p \in [n^{-1}(\log n)^5, n^{-\gamma}]$ *соответственно,* $0 < \gamma < 1$. *Пусть* T^F – *FDR-порог с управляющим параметром* $\alpha_n \to 0$ *и* $\frac{\alpha_n k_n \gamma_n^2}{\log n} \to \infty$ *при* $n \to \infty$, *где* κ_n *и* γ_n *определены в* (8). *Тогда*

$$\frac{\hat{R}(T^F) - R(T^{\min})}{n} \to 0 \ \ п.в.$$

Теоремы 1.6 и 1.7, приведенные ниже, были доказаны в работе [5], в этих теоремах приводятся условия для асимптотической нормальности оценки (5) для класса разреженности $L_0(\eta_n)$ (теорема 1.6) и класса разреженности $L_p(\eta_n)$ (теорема 1.7).

Теорема 1.6. *Пусть* $\mu \in L_0(\eta_n)$, $\eta_n \in [n^{-1}(\log n)^5, n^{-\gamma}]$, $1/2 < \gamma < 1$. *Пусть* T_F – *FDR-порог с управляющим параметром* $\alpha_n \to 0$ *и* $\frac{\alpha_n \kappa_n \gamma_n^2}{\log n} \to \infty$ *при* $n \to \infty$, *где* κ_n *и* γ_n *определены в* (8). *Тогда*

$$\frac{\hat{R}(T^F) - R(T^{min})}{\sigma^2\sqrt{2n}} \Rightarrow \mathcal{N}(0,1).$$

Теорема 1.7 *Пусть* $\mu \in L_p(\eta_n)$, $0 < p < 2$, $\eta_n^p \in [n^{-1}(\log n)^5, n^{-\gamma}]$, $1/2 < \gamma < 1$. *Пусть* T_F – *FDR-порог с управляющим параметром* $\alpha_n \to 0$ *и* $\frac{\alpha_n \kappa_n \gamma_n^2}{\log n} \to \infty$ *при* $n \to \infty$, *где* κ_n *и* γ_n *определены в* (9). *Тогда*

$$\frac{\hat{R}(T^F) - R(T^{min})}{\sigma^2\sqrt{2n}} \Rightarrow \mathcal{N}(0,1).$$

Результаты теорем 1.6 и 1.7 позволяют строить асимптотические доверительные интервалы для оценки риска $R(T^{min})$ для классов разреженности $L_0(\eta_n)$ и $L_p(\eta_n)$. А именно, исходя из полученных результатов, что

$$\frac{\hat{R}(T^F) - R(T^{min})}{\sigma^2\sqrt{2n}} \Rightarrow \mathcal{N}(0,1),$$

можно записать следующее:

$$\mathbb{P}\left(-z_{\frac{1+\gamma}{2}} \le \frac{\hat{R}(T^F) - R(T^{min})}{\sigma^2\sqrt{2n}} \le z_{\frac{1+\gamma}{2}}\right) = \mathbb{P}\left(T_1 \le R(T^{min}) \le T_2\right) \to \gamma.$$

где $z_{\frac{1+\gamma}{2}}$ - квантиль стандартного нормального распределения уровня $\frac{1+\gamma}{2}$ и

$$T_1 = \hat{R}(T^F) - \sigma^2\sqrt{2n} \cdot z_{\frac{1+\gamma}{2}},$$

$$T_2 = \hat{R}(T^F) + \sigma^2\sqrt{2^n} \cdot z_{\frac{1+\gamma}{2}}.$$

Таким образом, границы доверительных интервалов минимального среднеквадратичного риска $R(T^{min})$ могут быть найдены через значения оценки риска при использовании FDR-порога.

В работе [2] были доказаны следующие теоремы 1.8 и 1.9 для различных классов разреженности данных.

Теорема 1.8 *Пусть* $\mu \in L_0(\eta_n)$, *где* $\eta_n \in [n^{-1}(\log n)^5, n^{-\gamma}]$, $1/2 < \gamma < 1$. *Пусть* T_F — *FDR-порог с управляющим параметром* $\alpha_n \to 0$ *и* $\frac{\alpha_n \kappa_n \gamma_n^2}{\log n} \to \infty$ *при* $n \to \infty$, *где* κ_n *и* γ_n *определены в* (8). *Тогда для* $x \in \mathbb{R}$ *выполнены следующие неравенства*:
для $\gamma \in \left(\frac{1}{2}, \frac{3}{4}\right]$

$$\sup_x \left| \mathbb{P}\Big(\frac{\hat{R}(T^F) - R(T^{min})}{\sigma^2\sqrt{2n}} < x \Big) - \Phi(x) \right| \le C \cdot n^{1/2-\gamma} \cdot \log n,$$

для $\gamma \in \left(\frac{3}{4}, 1\right)$

$$\sup_x \left| \mathbb{P}\Big(\frac{\hat{R}(T^F) - R(T^{min})}{\sigma^2\sqrt{2n}} < x \Big) - \Phi(x) \right| \le C \cdot n^{-1/4} \cdot \log n,$$

где в обоих случаях C *- это константы, вообще говоря, разные.*

Теорема 1.9 *Пусть* $\mu \in L_p(\eta_n)$, $0 < p < 2$, *где* $\eta_n^p \in [n^{-1}(\log n)^5, n^{-\gamma}]$, $1/2 < \gamma < 1$. *Пусть* T_F — *FDR-порог с управляющим параметром* $\alpha_n \to 0$ *и* $\frac{\alpha_n \kappa_n \gamma_n^2}{\log n} \to \infty$ *при* $n \to \infty$, *где* κ_n *и* γ_n *определены в* (9). *Тогда для* $x \in \mathbb{R}$ *выполнены следующие неравенства*:
для $\gamma \in \left(\frac{1}{2}, \frac{3}{4}\right]$

$$\sup_x \left| \mathbb{P}\Big(\frac{\hat{R}(T^F) - R(T^{min})}{\sigma^2\sqrt{2n}} < x \Big) - \Phi(x) \right| \le C \cdot n^{1/2-\gamma} \cdot \log n^{1+p/2},$$

для $\gamma \in \left(\frac{3}{4}, 1\right)$

$$\sup_x \left| \mathbb{P}\Big(\frac{\hat{R}(T^F) - R(T^{min})}{\sigma^2\sqrt{2n}} < x \Big) - \Phi(x) \right| \le C \cdot n^{-1/4} \cdot \log n,$$

где в обоих случаях C *- это константы, вообще говоря, разные.*

Полученные оценки скорости сходимости могут быть использованы в асимптотических методах статистики, например, при построении доверительных интервалов необходимо знать меру близости исходного и предельного распределений.

Вторая глава посвящена рассмотрению задачи, в которой необходимо оценить исходные данные $f \in L^2(\mathbb{R})$, но при этом для наблюдения доступны данные, подвергнутые действию линейного преобразователя K, помимо этого в наблюдаемых данных присутствует шум. Подобного рода задачи можно встретить при анализе результатов компьютерной томографии, физики плазмы и телекоммуникационного трафика. Модель можно записать формально в следующем виде:

$$X_i = (Kf)_i + \epsilon_i, \;\; i = 1, \ldots, n,$$

где $f \in L^2(\mathbb{R})$ — функция исходных данных, ϵ_i — независимые нормально распределенные случайные величины с нулевым математическим ожиданием и дисперсией σ^2, а K — линейное преобразование.

Для того, чтобы выразить исходные данные f, можно было бы применить обратный оператор K^{-1}, однако такой оператор может либо не существовать, либо не быть ограниченным, что приводит к так называемым некорректно поставленным статистическим задачам, для решения которых используются методы регуляризации в сочетании с различными методами разложения. Одним из видов разложений является применение аппарата вейвлет-анализа. В частности, рассмотрим метод вейглет-вейвлет-разложения, который был предложен в работе[28].

В диссертации будет рассматриваться случай, когда оператор K является однородным (о причинах выбора именно таких операторов будет сказано далее). Оператор K называется однородным с показателем $\beta \geq 0$, если для любого $a > 0$ выполнено

$$K[f(ax)] = a^{-\beta}(Kf)[ax]. \tag{10}$$

В случае, когда показатель $\beta \geq 0$ задача обращения оператора K будет являться некорректно поставленной статистической задачей. При $\beta < 0$ задача будет считаться корректной и такой случай не будет рассматриваться в рамках диссертации.

Вейвлет-разложение подразумевает под собой представление функции в виде колебаний, локализованных по частоте и времени. Это выгодно отличает вейвлет-преобразования от преобразований Фурье[29], которые не дают информации о том, на какой частоте в какой момент времени в исходных данных присутствовали колебания. Также популярность вейвлет-разложения обусловлена тем, что в отличие от рядов Фурье, вейвет-анализ хорошо применим не только к пространственно однородным функциям, но и к функциям, имеющим локальные сингулярности.

Аппарат вейвлет-анализа базируется на понятии «материнской» вейвлет-функции $\psi(x) \in L^2(R)$, которая представляет собой затухающие на

[28] См. сноску 13 выше
[29] См. сноску 21 выше

прямой колебания. Для того, чтобы ψ была вейвлет-функцией необходимо выполнение условия допустимости:

$$\int_{-\infty}^{\infty} \psi(t)dt = 0.$$

Так же, как и для преобразований Фурье, можно рассматривать интегральную форму (непрерывное вейвлет-преобразование) и разложение в ряд (двоичное вейвлет-преобразование). В диссертации будем рассматривать двоичное вейвлет-преобразование, которое зачастую в литературе называют дискретным. Пусть имеется исходная функция $f \in L^2(\mathbb{R})$, тогда ее вейвлет-разложение выглядит следующим образом:

$$f = \sum_{j,l\in\mathbb{Z}} \langle f, \psi_{j,l}\rangle \psi_{j,l}\,, \tag{11}$$

где $\psi_{j,l}(x) = 2^{j/2}\psi(2^j x - l)$ образуют ортонормированный базис в $L^2(\mathbb{R})$. При этом набор коэффициентов $\{\langle f, \psi_{j,l}\rangle\}$ называется двоичным вейвлет-преобразованием функции f и для любой функции $f \in L^2(\mathbb{R})$ можно записать представление

$$f(x) = \sum_{j=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} \langle f, \psi_{j,l}\rangle \psi_{j,l}(x).$$

В методе вейглет-вейвлет-разложения результат линейного преобразования функции f раскладывается в ряд из сдвигов и растяжений некоторой вейвлет-функции ψ:

$$Kf = \sum_{j,l\in\mathbb{Z}} \langle Kf, \psi_{j,l}\rangle \psi_{j,l}\,,$$

где $\psi_{j,l}(x) = 2^{j/2}\psi(2^j x - l)$, причем базис, по которому раскладывается преобразование Kf, получается ортонормированным. Этот важный факт позволяет сделать вывод о том, что зашумленные коэффициенты разложения по-прежнему будут являться независимыми случайными величинами, что существенно облегчает дальнейшие исследования.

Потребуем, чтобы для оператора K существовали такие константы $\beta_{j,l}$, чтобы последовательность функций $u_{j,l} = K^{-1}\psi_{j,l}/\beta_{j,l}$ образовывала устойчивый базис, т.е. существовали такие константы $0 < A \le B < \infty$, что

$$A\sum_{j,l} c_{j,l}^2 \le \left\| \sum_{j,l} c_{j,l} u_{j,l} \right\|_{L^2}^2 \le B\sum_{j,l} c_{j,l}^2. \tag{12}$$

Функции $u_{j,l}$, для которых выполнено $\langle f, u_{j,l}\rangle = \langle Kf, \psi_{j,l}\rangle$, называют «вейглетами». Очевидно, что не для всех опреаторов K найдется такое семейство вейглетов $u_{j,l}$, что будет выполнено условие (12). Поэтому в дальнейшем будем рассматривать только однородные преобразования K, для которых гарантируется выполнение (12) при соблюдении некоторых условий регулярности для материнской вейвлет-функции ψ[30].

Пусть $\beta_{j,l} = \left\|K^{-1}\psi_{j,l}\right\|_{L^2}$. В работе[31] было показано, что

$$\beta_{j,l} = 2^{\beta\cdot j}\beta_{0,0}.$$

Тогда разложение исходной функции f представимо в виде

$$F = \sum_{j,l\in\mathbb{Z}} 2^{\beta j}\beta_{0,0}\langle Kf, \psi_{j,l}\rangle u_{j,l}\,. \tag{13}$$

Таким образом, в методе вейглет-вейвлет-разложения происходит разложение наблюдаемого преобразованного сигнала Kf по ортонормированному базису, затем применяется пороговая обработка к коэффициентам разложения и наконец исходные незашумленные данные f выражаются через обновленные коэффициенты и вейглеты.

Аппроксимация функции Kf записывается в виде суммы из сдвигов и растяжений вейвлет-функции ψ:

$$Kf \approx \sum_{j=0}^{J-1}\sum_{l=0}^{2^j-1}\langle Kf, \psi_{j,l}\rangle\psi_{j,l}\,.$$

Следовательно, аппроксимация исходной функции f представима в виде

$$F = \sum_{j=0}^{J-1}\sum_{l=0}^{2^j-1} 2^{\beta j}\beta_{0,0}\langle Kf, \psi_{j,l}\rangle u_{j,l}\,, \tag{14}$$

где $u_{j,l} = K^{-1}\psi_{j,l}/\beta_{j,l}$.

Как правило, на практике функция f задана дискретно и имеет конечный носитель. Поэтому перейдем к рассмотрению функции f и ее линейного преобразования Kf в точках i/N, где $N = 2^J$ и $i = 1, \ldots, N$. Дискретное вейвлет-преобразование представляет собой умножение вектора значений функции на ортогональную матрицу, порождаемую вейвлет-функцией ψ[32]. Обозначим через $\mu_{j,l}$ соответствующие коэффициенты, полученные после дискретного вейвлет-преобразования. Коэффициенты $\mu_{j,l}$ образуют вектор $\boldsymbol{\mu} \in \mathbb{R}^d$, $d = 2^J$.

[30] См. сноску 13 выше
[31] См. сноску 13 выше
[32] См. сноску 21 выше

Для каждого уровня $j \in [0, J-1]$ рассмотрим упорядоченные по абсолютной величине значения $\mu_{j,(k)}$, $k = 1, \ldots, 2^j$:

$$\left|\mu_{j,(1)}\right| \geq \ldots \geq \left|\mu_{j,(2^j)}\right|.$$

Обозначим $\boldsymbol{\eta} = (\eta_0, \eta_1, ..., \eta_{J-1})$, где $\eta_j \in [0,1]$, $j \in [0, J-1]$. Как и в главе 1 будем рассматривать векторы $\boldsymbol{\mu}$, которые принадлежат классу разреженности L_p для $0 < p < 2$:

$$L_p(\boldsymbol{\eta}) = \{\boldsymbol{\mu} \in \mathbb{R}^d : \left|\mu_{j,(k)}\right| \leq \eta_j \cdot 2^{j/p} \cdot k^{-1/p} \text{ для всех } k = \overline{1, 2^j},\ j = \overline{0, J-1}\}.$$

Для пороговой обработки используются функции жесткой (2) и мягкой (3) пороговой обработки, определенные в главе 1. Однако в случае обращения линейного однородного оператора для каждого масштаба разложения возникает корректировка значения среднеквадратичного риска, поэтому необходимо вычислять значения порога на каждом уровне j. В таком случае агрументами для пороговых функций будут $X^W_{j,k}$ и T_j, для каждого $j \in [0, J-1]$, где $X^W_{j,k}$ — зашумленные вейвлет-коэффициенты из разложения (13) (см. работу[33]).

Пороговое значение T_j, вычисляемое для каждого уровня разложения j, в случае использования FDR-метода выбирается следующим образом: по исходной выборке строится вариационный ряд убывающих по абсолютному значению величин

$$|X|^{(1)}_j \geq |X|^{(2)}_j \geq \ldots \geq |X|^{(2^j)}_j,$$

и затем $|X|^{(k)}_j$ сравнивается с квантилями Гауссова распределения $t_k = \sigma z(\alpha/2 \cdot k/2^j)$. Пусть k_F — наибольший индекс k, для которого $|X|^{(k)}_j \geq t_k$, тогда выбирается порог $T^F_j = t_{k_F}$.

Определение среднеквадратичного риска в случае дискретного вейвлет-разложения несколько отличается от определения, приводимого в главе 1, а именно:

$$R(\boldsymbol{T}) = \sum_{j=0}^{J-1} \sum_{k=1}^{2^j} \beta^2_{j,k} \mathbb{E}\left(\rho(X^W_{j,k}, T_j) - \mu_{j,k}\right)^2, \tag{15}$$

где $\boldsymbol{T} = (T_0, T_1, ..., T_{J-1})$, а ρ — функция используемой пороговой обработки.

Обозначим значение порога, при котором риск достигает минимума, $\boldsymbol{T^{min}}$:

$$\boldsymbol{T^{min}} : \ R(\boldsymbol{T^{min}}) = \min_{\boldsymbol{T}} R(\boldsymbol{T}),$$

[33] *Abramovich F., Benjamini Y.* Adaptive thresholding of wavelet coefficients // Computational statistics and data analysis. 1996. Vol. 22, no. 4. P. 351—361.

где $\boldsymbol{T^{min}} = (T_0^{min}, T_1^{min}, ..., T_{J-1}^{min})$. Как уже отмечалось в главе 1 в выражении $R(\boldsymbol{T})$ присутствуют неизвестные величины $\mu_{j,k}$, из-за чего вычислить значения $R(\boldsymbol{T})$ и $\boldsymbol{T^{min}}$ на практике нельзя. Поэтому будем рассматривать оценку риска, которая вычисляется на основе только наблюдаемых данных и определяется следующим выражением:

$$\hat{R}(\boldsymbol{T}) = \sum_{j=0}^{J-1} \sum_{k=1}^{2^j} 2^{2\beta j} \beta_{0,0} F[X_{j,k}^W, T_j], \tag{16}$$

где в случае жесткой пороговой обработки

$$F[X_{j,k}^W, T_j] = \left((X_{j,k}^W)^2 - \sigma^2\right) \cdot \mathbf{1}(|X_{j,k}^W| \leqslant T_j) + \sigma^2 \cdot \mathbf{1}(|X_{j,k}^W| > T_j)$$

и в случае мягкой пороговой обработки

$$F[X_{j,k}^W, T_j] = \left((X_{j,k}^W)^2 - \sigma^2\right) \cdot \mathbf{1}(|X_{j,k}^W| \leqslant T_j) + (\sigma^2 + T_j^2) \cdot \mathbf{1}(|X_{j,k}^W| > T_j).$$

Как уже было отмечено в главе 1, в силу того, что на практике вычислить риск (15) и порог $\boldsymbol{T^{min}}$, минимизирующий этот риск, не представляется возможным из-за вхождения неизвестных компонент незашумленного сигнала, имеет смысл перейти к рассмотрению оценки риска. Оценка риска, определяемая выражением (16), в свою очередь не зависит от неизвестных компонент исходного сигнала, а значит может быть вычислена на практике. Более того, с помощью оценки риска можно сравнивать между собой различные значения порогов с целью нахождения в некотором смысле наилучшего порогового значения.

Для удобства дальнейшего изложения введем ряд обозначений:

$$\gamma_j = \frac{1}{\log\log(2^j)}, \quad \tau_j = \sigma \left(2 \log \eta_j^{-p}\right)^{1/2}, \quad \kappa_j = \frac{2^j \, \eta_j^p \tau_j^{-p}}{1 - \alpha_j - \gamma_j} \quad \text{для } L_p(\eta_j). \tag{17}$$

Следующая теорема о сильной состоятельности оценки риска в случае обращения линейных однородных операторов была доказана в статье [4].

Теорема 2.1. *Пусть* $\mu \in L_p(\boldsymbol{\eta})$, $\eta_j^p \in [2^{-j}(\log 2^j)^5, 2^{-j\cdot\gamma}]$, $0 < \gamma < 1$, $j \in [0, J-1]$. *Пусть* $\boldsymbol{T^F}$ *— вектор FDR-порогов с управляющими параметрами* $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, ..., \alpha_{J-1})$ *такими что,* $\alpha_j \to 0$, $\frac{\alpha_j \kappa_j \gamma_j^2}{\log 2^j} \to \infty$ *при* $j \geq tJ$ *для некоторого* $0 < t < 1$ *и* $J \to \infty$, *где* κ_j *и* γ_j *определены в* (17). *Тогда при* $J \to \infty$

$$\frac{\hat{R}(\boldsymbol{T^F}) - R(\boldsymbol{T^{min}})}{2^{J(1+2\beta)}} \to 0 \quad \textit{п. в.}$$

Из теоремы 2.1 следует, что для рассматриваемой оценки среднеквадратичного риска $\hat{R}(\boldsymbol{T^F})$ при использовании FDR-порога имеет место сходимость почти наверное к минимальному риску $R(\boldsymbol{T^{min}})$.

Далее рассмотрим асимптотическую нормальность оценки риска $\hat{R}(\boldsymbol{T^F})$ (это свойство было доказано в статье [4]).

Теорема 2.2. *Пусть* $\mu \in L_p(\boldsymbol{\eta})$, $\eta_j^p \in [2^{-j}(\log 2^j)^5, 2^{-j\cdot\gamma}]$, $1/2 < \gamma < 1$. *Пусть* $\boldsymbol{T^F}$ — *вектор FDR-порогов с управляющими параметрами* $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, ..., \alpha_{J-1})$ *такими что,* $\alpha_j \to 0$, $\frac{\alpha_j \kappa_j \gamma_j^2}{\log 2^j} \to \infty$ *при* $j \geq tJ$ *для некоторого* $0 < t < 1/2$ *и* $J \to \infty$, *где* κ_j *и* γ_j *определены в* (17). *Тогда*

$$\frac{\hat{R}(\boldsymbol{T^F}) - R(\boldsymbol{T^{min}})}{\sigma^2\sqrt{2^{J+4\beta J}\beta_{0,0}^4}} \Rightarrow \mathcal{N}(0,1).$$

Аналогично замечаниям, приведенным в главе 1, результат теоремы 2.2 позволяет строить доверительные интервалы для минимального среднеквадратичного риска $R(\boldsymbol{T^{min}})$. Далее приведем их вид.

$$\mathbb{P}\left(-z_{\frac{1+\gamma}{2}} \leq \frac{\hat{R}(\boldsymbol{T^F}) - R(\boldsymbol{T^{min}})}{\sigma^2\sqrt{2^{J+4\beta J}\beta_{0,0}^4}} \leq z_{\frac{1+\gamma}{2}}\right) = \mathbb{P}\left(T_1 \leq R(\boldsymbol{T^{min}}) \leq T_2\right) \to \gamma.$$

где $z_{\frac{1+\gamma}{2}}$ - квантиль стандартного нормального распределения уровня $\frac{1+\gamma}{2}$,

$$T_1 = \hat{R}(\boldsymbol{T^F}) - \sigma^2\sqrt{2^{J+4\beta J}\beta_{0,0}^4} \cdot z_{\frac{1+\gamma}{2}},$$

$$T_2 = \hat{R}(\boldsymbol{T^F}) + \sigma^2\sqrt{2^{J+4\beta J}\beta_{0,0}^4} \cdot z_{\frac{1+\gamma}{2}}.$$

В следующей теореме, доказанной в работе [3], приводится оценка скорости сходимости к нормальному закону оценки риска для случая обращения линейных однородных операторов.

Теорема 2.3 *Пусть* $\mu \in L_p(\boldsymbol{\eta})$, $\eta_j^p \in [2^{-j}(\log 2^j)^5, 2^{-j\cdot\gamma}]$, $1/2 < \gamma < 1$. *Пусть* $\boldsymbol{T^F}$ — *вектор FDR-порогов с управляющими параметрами* $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, ..., \alpha_{J-1})$ *такими что,* $\alpha_j \to 0$, $\frac{\alpha_j \kappa_j \gamma_j^2}{\log 2^j} \to \infty$ *при* $j \geq tJ$ *для некоторого* $0 < t < \frac{1-\gamma}{2\beta+1}$ *при* $J \to \infty$, *где* κ_j *и* γ_j *определены в* (17). *Тогда для* $x \in \mathbb{R}$ *выполнено неравенство*:

$$\sup_x \left| \mathbb{P}\left(\frac{\hat{R}(\boldsymbol{T^F}) - R(\boldsymbol{T^{min}})}{\sigma^2\sqrt{2^{J+4\beta J}\beta_{0,0}^4}} < x\right) - \Phi(x)\right| \leq C \cdot J^{2+p/2} \cdot 2^{-J(\gamma-1/2)}$$

$$\text{для } \gamma \in \left(\frac{1}{2}, \frac{3}{4}\right],$$

$$\sup_x \left| \mathbb{P}\left(\frac{\hat{R}(\boldsymbol{T^F}) - R(\boldsymbol{T^{min}})}{\sigma^2 \sqrt{2^{J+4\beta J}\beta_{0,0}^4}} < x \right) - \Phi(x) \right| \leq C \cdot J^2 \cdot 2^{-\frac{J}{4}} \text{ для } \gamma \in \left(\frac{3}{4}, 1\right),$$

где в обоих случаях C - это константы, вообще говоря, разные.

Оценка скорости сходимости, полученная в теореме 2.3, имеет порядок при $\gamma \in \left(\frac{1}{2}, \frac{3}{4}\right]$ равный $J^{2+p/2} \cdot 2^{-J(\gamma-1/2)}$, а при $\gamma \in \left(\frac{3}{4}, 1\right)$ — $J^2 \cdot 2^{-\frac{J}{4}}$. Вспомним, что во второй главе диссертации $2^J = N$, следовательно порядок при $\gamma \in \left(\frac{1}{2}, \frac{3}{4}\right]$ равный $(\log N)^{2+p/2} \cdot N^{1/2-\gamma}$, а при $\gamma \in \left(\frac{3}{4}, 1\right)$ — $(\log N)^2 \cdot N^{-\frac{1}{4}}$.

Как отмечалось ранее в главе 1, оценки скорости сходимости из теоремы 2.3 могут быть использованы для построения доверительных интервалов, а также, например, для уточнения уровня доверия в доверительных интервалах.

Заключение

1. Обзор проведенного исследования. Тематика диссертации относится к области математической статистики. В работе рассмотрены задачи множественной проверки гипотез при применении FDR-порога в случае рассмотрения векторов большой размерности (глава 1), а также в случае обращения линейного однородного оператора (глава 2). Также в работе исследуются асимптотические свойства оценки среднеквадратичного риска при пороговой обработке. Основные результаты диссертации состоят в следующем:

- **Доказана** сильная состоятельность оценки риска при применении FDR-порога в случае классов разреженности исходных данных $L_0(\eta_n)$ и $L_p(\eta_n)$, а также сильная состоятельность оценки риска в задаче обращения линейных однородных операторов для класса разреженности $L_p(\boldsymbol{\eta})$.
- **Доказана** асимптотическая нормальность оценки риска при применении FDR-порога в случае классов разреженности исходных данных $L_0(\eta_n)$ и $L_p(\eta_n)$, а также асимптотическая нормальность оценки риска в задаче обращения линейных однородных операторов для класса разреженности $L_p(\boldsymbol{\eta})$. Помимо этого приведены виды асимптотических доверительных интервалов для минимального риска в обоих случаях рассматриваемых постановок задачи.
- **Получены** оценки скорости сходимости распределения оценки риска к нормальному закону при применении FDR-порога в случае рассмотрения векторов большой размерности для классов разреженности $L_0(\eta_n)$ и $L_p(\eta_n)$, а также в случае рассмотрения задачи

обращения линейного однородного оператора для $L_p(\boldsymbol{\eta})$ класса разреженности исходных данных.

2. Рекомендации и перспективы по дальнейшей разработке темы диссертации.

- Исследовать возможность уточнения оценок скорости сходимости распределения оценки риска к нормальному закону при применении FDR-порога в случае рассмотрения векторов большой размерности для классов разреженности $L_0(\eta_n)$ и $L_p(\eta_n)$, а также в случае рассмотрения задачи обращения линейного однородного оператора для $L_p(\boldsymbol{\eta})$ класса разреженности исходных данных.
- Исследовать асимптотические свойства оценок среднеквадратичного риска при применении FDR-порога в случае класса разреженности исходных данных $M_p(\eta_n)$, а также асимптотические свойства оценок риска в задаче обращения линейных однородных операторов для класса разреженности $M_p(\boldsymbol{\eta})$.
- Получить оценки скорости сходимости распределения оценки риска к нормальному закону при применении FDR-порога в случае рассмотрения векторов большой размерности для класса разреженности $M_p(\eta_n)$, а также для задачи обращения линейного однородного оператора для $M_p(\boldsymbol{\eta})$ класса разреженности исходных данных.

Благодарность

Автор выражает искреннюю признательность и благодарность научному руководителю, профессору Шестакову Олегу Владимировичу за постановку задач, помощь и поддержку на всех этапах выполнения диссертации.

Публикации автора по теме диссертации

В изданиях из списка Web of Science, Scopus, RSCI

1. *Палионная С. И.* Сильная состоятельность оценки риска при множественной проверке гипотез с FDR-порогом // Вестник Московского университета. Серия 15: Вычислительная математика и кибернетика. — 2020. — № 4. — 34—39 / 0.375 п.л.
   Журнал индексируется в РИНЦ, RSCI (ИФ РИНЦ 0.077).
2. *Палионная С. И.* Скорость сходимости оценки риска к нормальному закону в задаче множественной проверки гипотез с использованием FDR-порога // Вестник Московского университета. Серия 15: Вычислительная математика и кибернетика. — 2021. — № 3. — 31—36 / 0.375 п.л.

Журнал индексируется в РИНЦ, RSCI (ИФ РИНЦ 0.077).

3. *Палионная С. И.* Скорость сходимости распределения оценки риска к нормальному закону с использованием FDR-метода множественной проверки гипотез при обращении линейных однородных операторов // Вестник Московского университета. Серия 15: Вычислительная математика и кибернетика. — 2022. — № 3. — 49—55 / 0.4375 п.л.
Журнал индексируется в РИНЦ, RSCI (ИФ РИНЦ 0.077).

4. *Палионная С. И., Шестаков О. В.* Использование FDR-метода множественной проверке гипотез при обращении линейных однородных операторов // Информатика и ее применения. — 2022. — Т. 16, № 2. — 44—51 / 0.5 п.л. / 0.3 п.л.
Журнал индексируется в Scopus, РИНЦ (ИФ Scopus 0.604).

5. *Palionnaya S. I., Shestakov O. V.* Asymptotic Properties of MSE Estimate for the False Discovery Rate Controlling Procedures in Multiple Hypothesis Testing // Mathematics. — 2020. — Vol. 8, no. 11. — 1913 / 0.0625 п.л. / 0.04 п.л.
Журнал индексируется в Scopus, РИНЦ, Web of Science (ИФ WoS 2.592).

В работах [4, 5] постановки задач принадлежат О. В. Шестакову, все результаты получены С. И. Палионной самостоятельно.

В сборниках трудов конференций

6. *Палионная С. И., Шестаков О. В.* Свойства оценки риска при множественной проверке гипотез с использованием FDR-метода // Тихоновские чтения. — 2020. — С. 31—31.

7. *Палионная С. И., Шестаков О. В.* Асимптотическое поведение оценки риска FDR-метода в задаче множественной проверки гипотез // Ломоносовские чтения-2021. — 2021. — С. 120—121.

8. *Палионная С. И., Шестаков О. В.* Свойства оценок риска в задачах обращения линейных операторов при использовании FDR-метода множественной проверки гипотез // Ломоносовские чтения-2022. — 2022. — С. 171—172.

9. *Palionnaya S. I., Shestakov O. V.* Asymptotic behavior of a risk estimate for the FDR-method in the problem of multiple hypothesis testing // XXXVI International Seminar on Stability Problems for Stochastic Models. — 2021. — Available at: http://isspsm2021.krc.karelia.ru/ru/conf/7/presentation.

Палионная Софья Игоревна

Асимптотические свойства оценок риска в задачах множественной проверки гипотез

Автореф. дис. на соискание ученой степени канд. физ.-мат. наук

Подписано в печать ___.___._____. Заказ № _____
Формат 60×90/16. Усл. печ. л. 1. Тираж 100 экз.
Типография _____