Московский государственный университет
имени М.В.Ломоносова
Механико-математический факультет

На правах рукописи

Кожевин Алексей Александрович

Вероятностные методы отбора значимых факторов

01.01.05 - теория вероятностей и математическая статистика

ДИССЕРТАЦИЯ
на соискание ученой степени
кандидата физико-математических наук

Научный руководитель:
доктор физико-математических наук, профессор
Булинский Александр Вадимович

Москва, 2021 г.

Содержание

Введение **2**

1 Модификация MDR метода для стратифицированных выборок **8**
1.1 Постановка задачи 8
1.2 Вспомогательные результаты 10
1.3 Основной результат 12
1.4 Стоимостной подход к проведению экспериментов и XOR-модель 23
1.5 Результаты численных экспериментов 30

2 Оценивание условной энтропии Шеннона **35**
2.1 Постановка задачи 35
2.2 Основной результат 38
2.3 Вспомогательные результаты 41
2.4 Доказательства новых результатов 43

3 Оценивание взаимной информации в смешанной модели **86**
3.1 Постановка задачи 86
3.2 Основной результат 89
3.3 Сравнение оценок взаимной информации 96
3.4 Сравниваемые модели 99
3.5 Результаты компьютерных симуляций 101
3.6 Вспомогательные результаты 104

4 Выбор значимых факторов **111**
4.1 Основные определения 111
4.2 Метод идентификации значимых признаков 112
4.3 Основные результаты 112
4.4 Численные эксперименты 116

Заключение **118**

Список литературы **120**

Введение

Актуальность темы. Диссертация посвящена некоторым методам отбора значимых факторов (признаков), влияющих на изучаемый случайный отклик, а также новым оценкам условной энтропии и взаимной информации в рамках смешанной модели. В таких моделях вектор объясняющих переменных является абсолютно-непрерывным, а переменная отклика имеет дискретное распределение. Эти результаты также важны для идентификации набора значимых факторов.

В настоящее время необходимость применения методов анализа данных и отбора значимых факторов возникает в самых различных областях: от медицины и биологии до финансов и маркетинга (см., например, [50, 55, 63, 64, 81, 86] и там же ссылки). С развитием технологий увеличивается и количество данных, которые необходимо анализировать. За последние десятилетия был предложен целый ряд новых методов для решения задач классификации и регрессии (см., например, [4, 8, 35, 42]). Многие методы могут давать более точные прогнозы, если из данных удалить шумовые и избыточные признаки. Для этого стали разрабатываться новые методы отбора значимых признаков. Кроме того, отбор значимых признаков помогает создавать интерпретируемые модели, что особо важно в медико-биологических исследованиях, например, для определения генетических факторов, влияющих на возникновение того или иного заболевания.

Методам отбора значимых признаков посвящено огромное количество работ (см., например, [10, 23, 56, 66, 82, 89]). Множество стохастических и эвристических методов, направленных на выявление эпистаза (в генетике) рассматривается в [24, 71, 84].

В главе 1 диссертации предлагается модификация MDR метода (*multi-factor-dimensionality reduction*), основанная на использовании стратифицированной выборки вместо выборки из независимых одинаково распределённых наблюдений.

MDR метод впервые был предложен в работе [81] и разработан для обнаружения значимых (в определенном смысле) факторов, существенно влияющих на значение бинарной переменной отклика. Обзор, представленный в работе [45], демонстрирует огромную популярность этого метода. Только в период с 2001 по 2014 год было опубликовано около 800 работ, посвящённых его модификациям, обобщениям и практическим применениям. В работах [3, 29, 30] был предложен новый подход к введению MDR метода на основе функционала ошибки предсказания функции отклика. Кроме того, в этих статьях был доказан ряд асимптотических свойств возникающих статисти-

ческих оценок. Однако в случае, если анализируемая выборка из независимых одинаково распределенных наблюдений несбалансирована, т.е. количество наблюдений одного из классов в выборке намного больше, чем другого, то базовый MDR метод может быть недостаточно точен для выборок малого размера, поэтому в нашей работе [100] была предложена модификация метода для статифицированных выборок. В стратифицированной выборке наблюдения уже не являются независимыми в совокупности, поэтому используемая в MDR методе оценка функционала ошибки требует изменений. Для рассматриваемой модификации доказана сильная состоятельность оценки функционала ошибки, а также состоятельность процедуры отбора значимых факторов при условии, что их число известно. Также в первой главе дается новый стоимостной подход для сравнения предложенной модификации с базовым MDR-EFE методом, проведены компьютерные симуляции, продемонстрировавшие преимущества данной модификации.

Глава 2 посвящена новой оценке условной энтропии дискретной случайной величины, принимающей конечное число значений, при условии, задаваемом абсолютно непрерывным вектором. Понятие энтропии играет важную роль в физике и математике, см., напр., [16]. Следует отметить вклад в развитие понятия работы Л.Больцмана, Дж.Гиббса, М.Планка, К.Шеннона, А.Н.Колмогорова, Я.Г.Синая, А.Реньи, К. Цаллиса, А.С.Холево. В различных задачах возникает необходимость использования статистических оценок энтропии, построенных по независимым одинаково распределённым наблюдениям. Например, такие оценки важны при выборе значимых переменных [79] и выявлении неоднородностей в материалах [11], а также применяются в различных областях физики, в задачах теории информации и машинного обучения [9, 43, 47, 48]. О прочих задачах, в которых используются статистические оценки энтропии можно прочитать, например, в [76]. Разработано множество различных подходов к оцениванию энтропии в различных моделях, с которыми можно ознакомиться, например, в следующих работах: [12, 15, 31, 51, 69, 74, 77, 80, 90, 91, 93, 97].

Главная цель данной главы — ввести новую статистическую оценку условной энтропии Шеннона для смешанной модели, в которой дискретная переменная отклика со значениями в произвольном конечном множестве зависит от вектора, состоящего из факторов (признаков), имеющих плотность относительно меры Лебега на пространстве $\mathbb{R}^d$. Смешанной моделью является, например, известная модель логистической регрессии (см., например, [53, 58]). Предлагаемая оценка основана на статистиках k-ближайших соседей, где $k = k_n$ зависит от числа наблюдений n (со статистиками k-ближайших соседей можно ознакомиться, например, в книге [19]). Отметим, что предлагаемая оценка не использует известные стати-

стики Козаченко-Леоненко [5, 37, 44, 46, 88]. Для новой оценки устанавливается асимптотическая несмещённость и L_2-состоятельность при стремлении размера выборки к бесконечности. Исследование оценок условной энтропии интересно по следующей причине. Взаимная информация двух случайных векторов представляется в виде разности безусловной энтропии одной случайной величины и условной энтропии этой случайной величины при условии второй. Эта информационная характеристика двух случайных величин может быть использована для нахождения значимых факторов, существенно влияющих на значение рассматриваемой переменной отклика. Подобная задача возникает в медицинских и биологических приложениях. Таким образом, статистические оценки взаимной информации, предложенные в диссертации, могут быть использованы для отбора значимых факторов. Это направление исследований развивается в следующей главе.

В главе 3 предлагается оценка взаимной информации абсолютно непрерывного случайного вектора и дискретной случайной величины, принимающей произвольное конечное число значений, основанная на описанной в главе 2 оценке условной энтропии. Взаимная (совместная) информация (*mutual information*) двух случайных векторов широко используется в различных задачах, например, для отбора значимых факторов и классификации, см. [14, 17, 38, 39, 41, 65, 75, 79, 87, 95, 96]. Для предлагаемой новой оценки и оценки взаимной информации, введенной в работе [32], доказаны асимптотическая несмещённость и L_2-состоятельность. Также приводятся результаты компьютерных симуляций, проведенных для сравнения нескольких известных оценок взаимной информации в рамках смешанной модели. Они продемонстрировали преимущества предложенной новой оценки.

В главе 4 описывается процедура отбора значимых признаков, использующая разработанную оценку взаимной информации для смешанной модели. В предположении, что число значимых признаков известно, доказывается состоятельность описанной процедуры. Также в главе приводятся результаты численных экспериментов, оценивающих точность предлагаемого метода.

Цель работы. Цель работы — разработка новых методов идентификации значимых признаков, влияющих на изучаемый случайный отклик. А именно, предложено обобщение MDR метода на случай стратифицированной выборки и построены новые оценки взаимной информации и условной энтропии в смешанной модели, когда вектор объясняющих переменных является абсолютно-непрерывным, а переменная отклика имеет дискретное распределение. Установлена состоятельность введенной процедуры отбора

значимых признаков.

Структура и объем работы. Диссертация, объемом 129 страниц, состоит из введения, четырех глав, заключения и списка литературы, насчитывающего 106 наименований. В заключении к диссертации сформулированы возможные направления дальнейших исследований. В диссертацию вошли результаты, полученные при работе над проектом 14-21-00162 Российского научного фонда (руководитель гранта - академик РАН А.Н.Ширяев).

Научная новизна. Результаты диссертации являются новыми и состоят в следующем.

В главе 1 вводится модификация оценки MDR-EFE ([3]) для случая стратифицированной выборки, которая используется для обнаружения значимых факторов среди дискретных случайных величин $X_1, \ldots, X_n$ в модели с дискретной переменной отклика Y. Для данной модифицированной оценки доказывается её сильная состоятельность. Вводится обобщённая XOR-модель. Дается стоимостной подход, позволяющий сравнивать предложенную оценку для стратифицированной выборки с оценкой MDR-EFE для выборки из независимых одинаково распределённых случайных величин.

В главе 2 определяется новая оценка условной энтропии дискретной случайной величины Y при условии случайного вектора X, имеющего абсолютно непрерывное распределение. Оценка основана на использовании статистик ближайших соседей (см., например, [19]). При широких условиях доказывается асимптотическая несмещённость и L_2-состоятельность упомянутой оценки.

В главе 3 доказывается асимптотическая несмещенность и L_2-состоятельность оценки взаимной информации, введённой в работе [32], а также определяется новая оценка и доказываются ее аналогичные свойства.

В главе 4 на основе предложенной оценки взаимной информации предлагается процедура отбора значимых факторов, и в предположении, что число значимых признаков известно, доказывается ее состоятельность.

Положения, выносимые на защиту.

1. Оценка функционала ошибки MDR-EFE метода в случае стратифицированных выборок является сильно состоятельной, а предложенная на ее основе модификация метода позволяет добиться более точных результатов (в смысле стоимостного подхода) по сравнению с базовым MDR методом в случае несбалансированных выборок.

2. Введенная оценка условной энтропии в смешанной модели является асимптотически несмещенной и L_2-состоятельной при выполнении весьма широких условий.

3. Оценка взаимной информации, построенная на основе предложенной оценки условной энтропии, является асимптотически несмещенной и L_2-состоятельной, а также более точной по сравнению с рядом других оценок в рамках смешанных моделей.

4. Предложенная оценка взаимной информации используется для построения процедуры отбора значимых признаков и при этом позволяет получить состоятельную оценку набора значимых признаков.

Методы исследования. В диссертации использованы методы теории вероятностей и математической статистики, а также математического анализа. В **главе 1** мы применяем закон больших чисел для треугольных массивов и метод кросс-валидации (перекрестной валидации) для оценивания ошибки предсказания. В **главе 2** используются свойства условного математического ожидания, различные виды неравенств концентрации. Существенную роль играет аппарат теории меры. **Главы 3 и 4** основаны на результатах второй главы и используют предельные теоремы для статистических оценок дифференциальной энтропии Шеннона.

Теоретическая и практическая ценность. Работа носит теоретический характер. Ее результаты могут быть использованы для дальнейшего развития теории методов отбора значимых признаков, а также для практического применения при анализе данных наблюдений.

Апробация. Основные результаты диссертационной работы были доложены на трех международных конференциях и всероссийской конференции:

- 3rd International Workshop «Analysis, Geometry and Probability», Ульм, Германия, 28 сентября - 3 октября 2015 г.
  Тема доклада: Variable selection for overlapping groups.
- 9th International Workshop on Applied Probability (IWAP 2018), Будапешт, Венгрия, 18-21 июня 2018 г.
  Тема доклада: Feature selection and the mutual information estimation.
- 4th International Conference on Stochastic Methods, г. Новороссийск, Россия, 2-9 июня 2019 г.
  Тема доклада: An information theory-based approach to feature selection
- Третья Санкт-Петербургская зимняя молодежная конференция по теории вероятностей и математической физике, Санкт-Петербург, Россия, 16-18 декабря 2019 г.
  Тема доклада: Критерии остановки для процедуры отбора значимых признаков.

По теме диссертации автором были сделаны следующие доклады на научно-исследовательских семинарах:

- Большой семинар кафедры теории вероятностей МГУ под руководством академика РАН, проф. А.Н. Ширяева (МГУ, Москва, 2019 г.),
- «Forschungsseminar Stochastische Geometrie und räumliche Statistik» под руководством Prof. E.Spodarev (Institut für Stochastik, Ulm University, Germany, 2016 г.),
- Аспирантский коллоквиум кафедры теории вероятностей под руководством академика РАН, проф. А.Н. Ширяева (МГУ, Москва, 2018 и 2019 гг.),
- «Современные проблемы фундаментальной математики и механики» под руководством академика РАН, проф. В.А.Садовничего (МГУ, Москва, 2019 г.),
- Городской семинар по теории вероятностей и математической статистике под руководством академика РАН, проф. И.А.Ибрагимова (ПОМИ РАН, Санкт-Петербург, 2019 г.),
- «Асимптотический анализ случайных процессов и полей» под руководством проф. А.В. Булинского (МГУ, Москва, 2015-2019 гг.).

Публикации. Основные результаты диссертации изложены в 7 публикациях автора [100]—[106]. Из них 4 статьи (см. [100]—[103], работы [100]—[102] в соавторстве), которые опубликованы в рецензируемых научных журналах, входящих в базы SCOPUS и Web of Science. В материалах международных конференций представлены 3 публикации [104]—[106].

Личный вклад автора Диссертантом совместно с научным руководителем проводился выбор темы, а также осуществлялось планирование всей работы. Профессору А.В.Булинскому принадлежит постановка задач и общий подход к их решению, им также доказаны леммы 1.1, 2.6, 2.8 и 3.9. Автору диссертации принадлежит доказательство теорем, следствий, остальных лемм, а также проведение компьютерных симуляций. В начале каждой главы диссертации также приводится список соответствующих публикаций с долей участия авторов.

Благодарность. Работа выполнена под научным руководством профессора Александра Вадимовича Булинского, которому автор выражает искреннюю благодарность.

1 Модификация MDR метода для стратифицированных выборок

При подготовке данной главы диссертации использован материал публикации [100], выполненной автором в соавторстве с профессором А.В.Булинским. В публикации А.В.Булинскому принадлежит постановка задач и общий подход к их решению, им также доказана лемма 1 (лемма 1.1 в диссертации). Все остальные результаты доказаны автором диссертации, им проведены все компьютерные симуляции.

1.1 Постановка задачи

Пусть $X_1, \ldots, X_d$ — набор случайных переменных (факторов) и Y — (случайная) переменная отклика, некоторым образом зависящая от $X := (X_1, \ldots, X_d)$. Мы предполагаем, что все рассматриваемые случайные элементы в данной работе определены на одном вероятностном пространстве $(\Omega, \mathcal{F}, \mathsf{P})$. Пусть X_i принимает значения в конечном множестве $\mathbb{X}_i$, $i = 1, \ldots, d$, а переменная отклика Y принимает значения -1 и 1 (в качестве области значений можно использовать множество $\{0, 1\}$, однако приводимые ниже выкладки станут чуть более громоздкими). Таким образом, $X : \Omega \to \mathbb{X}$, где $\mathbb{X} := \mathbb{X}_1 \times \ldots \times \mathbb{X}_d$. В эту схему укладывается ряд важных медицинских и биологических моделей, в которых X сочетает в себе генетические и негенетические факторы, а Y характеризует состояние здоровья пациента. Например, $Y = 1$ означает, что пациент болен *(case)*, $Y = -1$ соответствует случаю, когда пациент здоров *(control)*. Интерес представляет задача предсказания риска заданного заболевания по набору факторов X, а также идентификация набора значимых факторов $(X_{i_1}, \ldots, X_{i_r})$, отвечающих за его возникновение, где $\{i_1, \ldots, i_r\} \subset \{1, \ldots, d\}$, $r < d$.

Для любой функции $f \colon \mathbb{X} \to \{-1, 1\}$ качество прогноза Y с помощью $f(X)$ может быть оценено с помощью *функционала ошибки* (см. [3, 26])

$$Err(f) = \mathsf{E}|Y - f(X)|\psi(Y),$$

где *штрафная функция* $\psi \colon \{-1, 1\} \to \mathbb{R}_+$ вводится для того, чтобы поразному учитывать ошибки предсказания различных значений Y. Тривиальные случаи $\psi \equiv 0$, $Y \equiv 0$ или $Y \equiv 1$ исключаются из рассмотрения. Очевидно, что

$$Err(f) = 2 \sum_{y \in \{-1,1\}} \psi(y)\mathsf{P}(Y = y, f(X) \neq y). \tag{1}$$

Класс *оптимальных функций* f_{opt}, т.е. все такие f, на которых достигается минимум $Err(f)$, были описаны в работе [3]. Удобно из класса оптимальных выбрать функцию $f^*(x) = \mathbb{I}\{A\}(x) - \mathbb{I}\{\overline{A}\}(x)$, $x \in \mathbb{X}$, где

$$A = \left\{ x \in M \colon \mathsf{P}(Y = 1 | X = x) > \frac{\psi(-1)}{\psi(-1) + \psi(1)} \right\}, \tag{2}$$

$M = \{x \in \mathbb{X} \colon \mathsf{P}(X = x) > 0\}$, $\mathbb{I}\{A\}$ — индикатор множества A и $\overline{A}$ означает дополнение множества A. Отметим, что любой оптимальной функции f_{opt} соответствует одно и то же значение функционала ошибки $Err(f^*)$. В [3] объясняется, почему выбор штрафной функции

$$\psi(y) = \frac{1}{\mathsf{P}(Y = y)}, \quad y \in \{-1, 1\}, \tag{3}$$

предложенной в [94], является естественным.

Поскольку совместное распределение X и Y неизвестно, то невозможно вычислить значение функционала $Err(f)$ и найти оптимальную функцию f^*. Однако можно статистически оценить $Err(f)$ на основе выборки из независимых случайных векторов $(X^1, Y^1), \ldots,$ (X^n, Y^n), имеющих то же самое распределение, что и (X, Y). В работах [3, 29, 30] показано, как такие оценки могут быть использованы для определения набора значимых факторов. Этот метод называется MDR-EFE (*multifactor dimensionality reduction - error function estimation*). В работах [25, 29, 30] устанавливается асимптотическая нормальность введённых оценок, а в двух последних статьях исследуется случай небинарной переменной отклика.

Мы сконцентрируемся на следующей задаче. Пусть вероятность $\mathsf{P}(Y{=}1)$ мала (такая ситуация характерна для ряда заболеваний). Тогда в выборке $(X^1, Y^1), \ldots, (X^n, Y^n)$ для недостаточно большого n будет мало наблюдений с положительным значением переменной отклика (т.е. наблюдений с $Y^i = 1$) для того, чтобы сделать какие-нибудь статистически значимые выводы. Далее мы обсудим несколько способов обойти эту сложность, а именно, модификации некоторых предыдущих результатов, основанные на использовании стратифицированных выборок, а также рассмотрим ситуацию, в которой выборка состоит из случайного числа случайных наблюдений.

В разделе 1.2 приводятся вспомогательные результаты о законе распределения наблюдений, имеющих одно и тоже значение переменной отклика. В разделе 1.3 приводится основной результат: критерий сильной состоятельности статистик, оценивающих ошибку аппроксимации переменной отклика Y с помощью функции, зависящей от $X_1, \ldots, X_d$ для случая

стратифицированной выборки. Для построения данных статистик использовалась процедура кросс-валидации. Помимо обсуждения установленного результата, в данном разделе приводится пример приложения метода, а также показано, как применить основной результат к отбору значимых факторов. Раздел 1.4 посвящён новому стоимостному подходу к проведению экспериментов и обобщению XOR-модели (см. [98]). Метод MDR-EFE для независимых одинаково распределённых наблюдений сравнивается с модификацией метода для стратифицированных выборок. Результаты численных экспериментов (раздел 1.5) показали, что даже для выборок малого размера предложенная модификация имеет заметные преимущества.

1.2 Вспомогательные результаты

Пусть $(X, Y), (X^1, Y^1), (X^2, Y^2), \ldots$ — последовательность независимых одинаково распределённых случайных векторов, определённых на вероятностном пространстве $(\Omega, \mathcal{F}, \mathsf{P})$. Для каждого элементарного исхода $\omega \in \Omega$ рассмотрим последовательность $Y^1(\omega), Y^2(\omega), \ldots$ и выберем все индексы $1 \leqslant j_{-1}^1(\omega) < j_{-1}^2(\omega) < \ldots$, для которых $Y^{j_{-1}^k(\omega)}(\omega) = -1$, $k \in \mathbb{N}$. Аналогично все наблюдения $Y^i(\omega)$ со значением 1 обозначим $\{Y^{j_1^m(\omega)}(\omega)\}_{m\in\mathbb{N}}$, где $1 \leqslant j_1^1(\omega) < j_1^2(\omega) < \ldots$. Напомним, что для испытаний Бернулли с вероятностью успеха p случайная величина $U_{r,p}$, имеющая отрицательное биномиальное распределение, вводится как число успехов, необходимое для получения r неудач, где $r \in \mathbb{N}$ (в таком случае пишут $U_{r,p} \sim NB(r, p)$). Таким образом,

$$\mathsf{P}(U_{r,p} = k) = \binom{k+r-1}{k} p^k (1-p)^r, \quad k = 0, 1, \ldots.$$

Если мы рассмотрим события $\{Y^i = 1\}$ и $\{Y^i = -1\}$ как успехи и неудачи, соответственно (с вероятностью успеха $p = \mathsf{P}(Y = 1)$), то случайная величина j_{-1}^r будет иметь такое же распределение как $U_{r,p} + r$. Следовательно,

$$\mathsf{P}(j_{-1}^r = m) = \begin{cases} \binom{m-1}{m-r} p^{m-r}(1-p)^r, & m = r, r+1, \ldots, \\ 0, & m = 1, \ldots, r-1. \end{cases} \tag{4}$$

При $r = 1$ мы рассматриваем только первую строку в (4), поскольку в этом случае $\{1, \ldots, r-1\} = \varnothing$. Аналогично можно показать, что j_1^r имеет такое же распределение, что и $U_{r,1-p} + r$ где $U_{r,1-p} \sim NB(r, 1-p)$. Другими словами, j_1^r и $G_p^1 + \ldots + G_p^r$ одинаково распределены, где $G_p^1, \ldots, G_p^r$ — независимые случайные величины, распределённые по геометрическому

закону с параметром p (т.е. $\mathsf{P}(G_p^1 = k) = p(1-p)^{k-1}$, $k = 1, 2, \ldots$). Таким образом, $\mathsf{E}j_1^r = \frac{r}{p}$ и $j_1^r < \infty$ п.н. для всех $r \in \mathbb{N}$ (аналогично $\mathsf{E}j_{-1}^r = \frac{r}{1-p}$ для всех $r \in \mathbb{N}$). Отметим также, что существуют различные определения для отрицательного биномиального и геометрического распределения, поэтому мы приводим точные формулы.

Положим $Z^k := X^{j_1^k}$ для всех $k \in \mathbb{N}$. Определим σ-алгебру $\mathcal{B}$ всех подмножеств $\mathbb{X}$. Важную роль играет следующий простой результат.

Лемма 1.1 *Для всех $m \in \mathbb{N}$ случайные переменные $Z^1, \ldots, Z^m$ независимы и распределены так же, как X при условии $Y = 1$ (будем писать $X|Y = 1$), т.е. для всех $B \in \mathcal{B}$ и $k = 1, \ldots, m$*

$$\mathsf{P}(Z^k \in B) = \mathsf{P}(X \in B|Y = 1).$$

Доказательство. Сначала покажем, что для всех $B_i \in \mathcal{B}$ и $i = 1, \ldots, m$ выполняется тождество

$$\mathsf{P}(Z^1 \in B_1, \ldots, Z^m \in B_m) = \mathsf{P}(Z^1 \in B_1) \ldots \mathsf{P}(Z^m \in B_m). \tag{5}$$

По формуле полной вероятности

$$\begin{aligned}\mathsf{P}(Z^1 \in B_1, \ldots, Z^m \in B_m) &= \mathsf{P}(X^{j_1^1} \in B_1, \ldots, X^{j_1^m} \in B_m) \\ &= \sum_{\substack{(k_1,\ldots,k_m)\in\mathbb{N}^m: \\ k_1<\cdots<k_m}} \mathsf{P}(X^{j_1^1} \in B_1, \ldots, X^{j_1^m} \in B_m, j_1^1 = k_1, \ldots, j_1^m = k_m).\end{aligned}$$

Отметим, что для произвольных положительных $k_1 < k_2 < \cdots < k_m$

$$\{j_1^1 = k_1, \ldots, j_1^m = k_m\} = \bigcap_{i=1}^m \{Y^{k_i} = 1\} \cap \bigcap_{r \in T_m} \{Y^r = -1\},$$

где $T_m = T_m(k_1, k_2, \ldots, k_m) := \{1, \ldots, k_m\} \setminus \{k_1, \ldots, k_m\}$. В силу независи-

мости случайных векторов $(X^1,Y^1),(X^2,Y^2),\ldots$, получаем

$$\begin{aligned}
&\mathsf{P}(Z^1\in B_1,\ldots,Z^m\in B_m)\\
&\quad=\sum_{\substack{(k_1,\ldots,k_m)\in\mathbb{N}^m:\\ k_1<\cdots<k_m}}\mathsf{P}\Bigg(\{X^{k_1}\in B_1,\ldots,X^{k_m}\in B_m,Y^{k_1}=1,\ldots,Y^{k_m}=1\}\\
&\qquad\qquad\cap\Bigg\{\bigcap_{r\in T_m}\{Y^r=-1\}\Bigg\}\Bigg)\\
&\quad=\sum_{\substack{(k_1,\ldots,k_m)\in\mathbb{N}^m:\\ k_1<\cdots<k_m}}\mathsf{P}(X^{k_1}\in B_1,Y^{k_1}=1)\ldots\mathsf{P}(X^{k_m}\in B_m,Y^{k_m}=1)\\
&\qquad\qquad\times\mathsf{P}\left(\bigcap_{r\in T_m}\{Y^r=-1\}\right)\\
&=\sum_{\substack{(k_1,\ldots,k_m)\in\mathbb{N}^m:\ q=1\\ k_1<\cdots<k_m}}\prod^m\frac{\mathsf{P}(X^{k_q}\in B_q,Y^{k_q}=1)}{\mathsf{P}(Y^{k_q}=1)}\mathsf{P}\left(\bigcap_{i=1}^m\{Y^{k_i}=1\}\cap\bigcap_{r\in T_m}\{Y^r=-1\}\right)\\
&\quad=\prod_{q=1}^m\mathsf{P}(X\in B_q|Y=1)\sum_{\substack{(k_1,\ldots,k_m)\in\mathbb{N}^m:\\ k_1<\cdots<k_m}}\mathsf{P}(j_1^1=k_1,\ldots,j_1^m=k_m)\\
&\qquad\qquad=\prod_{q=1}^m\mathsf{P}(X\in B_q|Y=1).
\end{aligned}$$

Если положить $B_1=\cdots=B_{k-1}=B_{k+1}=\cdots=B_m=\mathbb{X}$, то

$$\mathsf{P}(Z^k\in B_k)=\mathsf{P}(X\in B_k|Y=1),\ \ k=1,\ldots,m.$$

Следовательно, случайные переменные $Z^1,\ldots,Z^m$ одинаково распределены и равенство (5) справедливо. □

Замечание 1.2 . *Лемма 1.1 также выполняется для выборки* (X,Y), $(X^1,Y^1),\ldots,(X^n,Y^n)$, *когда* X *принимает значения в любом пространстве* S, *снабжённом* σ*-алгеброй* $\mathcal{B}$ *(случайный вектор* X *измерим относительно* σ*-алгебр* $\mathcal{F}$ *и* $\mathcal{B}$*). Тем же способом можно показать, что* $X^{j_{-1}^1},X^{j_{-1}^2},\ldots$ *независимы и распределены также как* $X|Y=-1$.

1.3 Основной результат

Пусть $(X^1,Y^1),(X^2,Y^2),\ldots$ — последовательность из независимых векторов с тем же распределением, что и (X,Y). Предположим, что имеется

возможность сформировать выборку с заданным неслучайным числом наблюдений обеих групп (n_1 и n_{-1} соответственно). Более точно, положим

$$\zeta_{n_1}^1 := \{(X^{j_1^1}, 1), \ldots, (X^{j_1^{n_1}}, 1)\}, \;\; \zeta_{n_{-1}}^{-1} := \{(X^{j_{-1}^1}, -1), \ldots, (X^{j_{-1}^{n_{-1}}}, -1)\},$$

случайные индексы j_1^k и j_{-1}^k, $k \in \mathbb{N}$ определены в разделе 1.2. Пусть $n := n_1 + n_{-1}$ — размер стратифицированной выборки $\zeta_n := \zeta_{n_1}^1 \cup \zeta_{n_{-1}}^{-1}$. Тогда имеется не одна выборка $\xi_n := \{(X^1, Y^1), \ldots, (X^n, Y^n)\}$, состоящая из независимых одинаково распределённых случайных векторов с распределением $Law(X, Y)$, а две подвыборки $\zeta_{n_1}^1$ и $\zeta_{n_{-1}}^{-1}$, которые согласно лемме 1.1 состоят из наблюдений, распределенных как $X|Y = 1$ и $X|Y = -1$, соответственно. Отметим, что по выборке ζ_n нельзя построить частотные оценки для $\mathsf{P}(Y = 1)$ или $\mathsf{P}(X \in B, Y = 1)$, где $B \subset \mathbb{X}$. Далее мы будем полагать, что $n_1 = \max\{[an], 1\}$ и $n_{-1} = n - n_1$, где параметр $a \in (0, 1)$, $n \in \mathbb{N}$, а $[\cdot]$ означает целую часть вещественного числа. Пусть имеются некоторые оценки $\widehat{\mathsf{P}}_n^y$ вероятностей $\mathsf{P}(Y = y)$ такие, что

$$\widehat{\mathsf{P}}_n^y \to \mathsf{P}(Y = y) \text{ п.н., } n \to \infty, \;\; y \in \{-1, 1\}. \tag{6}$$

Например, такая оценка $\widehat{\mathsf{P}}_n^y$ может быть построена по наблюдениям

$$\{(X^k, Y^k), 1 \leqslant k \leqslant \max\{j_1^{n_1}, j_{-1}^{n_{-1}}\}\}.$$

В таком случае, построенные по этим наблюдениям частотные оценки вероятностей $\mathsf{P}(Y = 1)$ или $\mathsf{P}(X \in B, Y = 1)$, где $B \subset \mathbb{X}$, являются сильно состоятельными, поскольку

$$\max\{j_1^{n_1}, j_{-1}^{n_{-1}}\} \to \infty \text{ п.н. при } n \to \infty.$$

Достоинства и недостатки использования стратифицированных выборок обсуждаются в разделе 1.4. Введём вектор оценок $\widehat{\mathsf{P}}_n := (\widehat{\mathsf{P}}_n^{-1}, \widehat{\mathsf{P}}_n^1)$. Напомним, что тривиальные случаи $\mathsf{P}(Y = -1) = 0$ или $\mathsf{P}(Y = 1) = 0$ исключаются из рассмотрения.

Пусть $f_{PA}(x, \zeta_n, \widehat{\mathsf{P}}_n)$ является функцией, определяющей алгоритм предсказания, т.е. функцией со значениями во множестве $\{-1, 1\}$, построенной по значению $x \in \mathbb{X}$, выборке ζ_n и $\widehat{\mathsf{P}}_n$. В действительности рассматривается семейство функций, где вместо ζ_n используются подвыборки, поэтому мы будем писать $f_{PA}(x, \zeta_n(S), \widehat{\mathsf{P}}_n)$ для $\zeta_n(S) := \{(X^j, Y^j), j \in S\}$, $S \subset (\{j_1^1, \ldots, j_1^{n_1}\} \cup \{j_{-1}^1, \ldots, j_{-1}^{n_{-1}}\})$. Для каждой $f \colon \mathbb{X} \to \{-1, 1\}$ найдем оценку f_{PA}, которая в некотором смысле близка к f, и будем использовать f_{PA} для оценивания $Err(f)$. Для этих целей мы также применяем процедуру кросс-валидации или, более точно, метод сэмплирования с усреднением. Вначале для некоторого фиксированного $K \in \mathbb{N}$ и всех

$y \in \{-1,1\}$ рассматривается разбиение $\{j_y^1,\dots,j_y^{n_y}\}$ на K подмножеств $S_k^y(n_y,\omega)$, $k=1,\dots,K$, следующим образом:

$$S_k^y(n_y,\omega) := \left\{ j_y^i(\omega) \colon i \in \left\{ (k-1)\left[\frac{n_y}{K}\right]+1,\dots,k\left[\frac{n_y}{K}\right]\mathbb{I}\{k<K\}+n_y\mathbb{I}\{k=K\} \right\} \right\}. \quad (7)$$

Теперь определим случайное множество

$$S_k(n) = S_k(n,\omega) := S_k^1(n_1,\omega) \cup S_k^2(n_2,\omega)$$

и введем оценку функционала ошибки

$$\widehat{Err}_K(f_{PA},\zeta_n,\widehat{\mathsf{P}}_n) := \frac{2}{K}\sum_{y\in\{-1,1\}}\sum_{k=1}^{K}\sum_{j\in S_k^y(n_y)}\frac{\widehat{\psi}(y,\zeta_n(\overline{S_k(n)}),\widehat{\mathsf{P}}_n)\mathbb{I}\{f_{PA}^j(n,k)\neq y\}\widehat{\mathsf{P}}_n^y}{\sharp S_k^y(n_y)}, \quad (8)$$

где $f_{PA}^j(n,k) := f_{PA}(X^j,\zeta_n(\overline{S_k(n)}),\widehat{\mathsf{P}}_n)$ и $\widehat{\psi}(y,\zeta_n(\overline{S_k(n)}),\widehat{\mathsf{P}}_n)$ — оценка $\psi(y)$, $y\in\{-1,1\}$, построенная по наблюдениям $\zeta_n(\overline{S_k(n)})$ и $\widehat{\mathsf{P}}_n$, $\sharp$ обозначает мощность конечного множества и

$$\overline{S_k(n)} = \{j_1^1,\dots,j_1^{n_1}\} \cup \{j_{-1}^1,\dots,j_{-1}^{n_{-1}}\} \setminus S_k(n).$$

Мы предполагаем, что для всех $k=1,\dots,n$,

$$\widehat{\psi}(y,\zeta_n(\overline{S_k(n)}),\widehat{\mathsf{P}}_n) \to \psi(y) \text{ п.н., } n\to\infty, \;\; y\in\{-1,1\}. \quad (9)$$

Введём $L(x) := \psi(1)\mathsf{P}(X=x,Y=1) - \psi(-1)\mathsf{P}(X=x,Y=-1)$, $x\in\mathbb{X}$. Следующий результат является аналогом теоремы 1 [3] для стратифицированных выборок.

Теорема 1.3 *Пусть ζ_n — выборка, описанная выше, ψ — штрафная функция, $f\colon \mathbb{X}\to\{-1,1\}$ — произвольная функция и f_{PA} определяет алгоритм предсказания. Предположим, что (9) выполняется и существует непустое множество $U\subset\mathbb{X}$ такое, что для всех $x\in U$ и $k=1,\dots,K$, соотношение*

$$f_{PA}(x,\zeta_n(\overline{S_k(n)}),\widehat{\mathsf{P}}_n) \to f(x) \text{ п.н., } n\to\infty, \quad (10)$$

выполняется. Тогда для всех $a\in(0,1)$ ($n_1=\max\{[an],1\}$, $n_{-1}=n-n_1$),

$$\widehat{Err}_K(f_{PA},\zeta_n,\widehat{\mathsf{P}}_n) \to Err(f) \text{ п.н., } n\to\infty, \quad (11)$$

тогда и только тогда, когда

$$\sum_{k=1}^{K}\sum_{y\in\{-1,1\}}\sum_{x\in\mathbb{X}_y} y\mathbb{I}\{f_{PA}(x,\zeta_n(\overline{S_k(n)}),\widehat{\mathsf{P}}_n)=-y\}L(x)\to 0 \text{ п.н.}, \quad n\to\infty, \tag{12}$$

где

$$\mathbb{X}_y=(\mathbb{X}\setminus U)\cap\{x\in\mathbb{X}: f(x)=y\}, \ \ y\in\{-1,1\}. \tag{13}$$

Доказательство. Достаточно рассмотреть отношение

$$\frac{2}{K}\sum_{k=1}^{K}\sum_{y\in\{-1,1\}}\psi(y)\sum_{j\in S_k^y(n_y)}\frac{\mathbb{I}\{f_{PA}^j(n,k)\neq y\}\mathsf{P}(Y=y)}{\sharp S_k^y(n_y)}\to Err(f) \text{ п.н., } n\to\infty, \tag{14}$$

поскольку разность между $\widehat{Err}_K(f_{PA},\zeta_n,\widehat{\mathsf{P}}_n)$ и левой частью (14) стремится к нулю п.н. при $n\to\infty$, а также в силу (6), (9) и неравенства

$$\frac{1}{\sharp S_k^y(n_y)}\sum_{j\in S_k^y(n_y)}\mathbb{I}\{f_{PA}^j(n,k)\neq y\}\leqslant 1,$$

которое выполняется для каждого $y\in\{-1,1\}$, любого $k=1,\ldots,K$ и всех n.

Теперь отметим, что в силу (1) соотношение (11) эквивалентно следующему:

$$\frac{2}{K}\sum_{k=1}^{K}\sum_{y\in\{-1,1\}}\psi(y)\left(\sum_{j\in S_k^y(n_y)}\frac{\mathbb{I}\{f_{PA}^j(n,k)\neq y\}\mathsf{P}(Y=y)}{\sharp S_k^y(n_y)}-\mathsf{P}(f(X)\neq y,Y=y)\right)$$
$$\to 0 \quad \text{п.н. } n\to\infty.$$

Для каждого $y\in\{-1,1\}$ и произвольного $m\in\mathbb{N}$, положим

$$\widetilde{X}^{j_y^m}:=\mathbb{I}\{f(X^{j_y^m})\neq y\}-\mathsf{P}(f(X^{j_y^m})\neq y).$$

Тогда $|\widetilde{X}^{j_y^m}|\leqslant 2$, $\mathsf{E}\tilde{X}^{j_y^m}=0$ для таких y и m. Зафиксируем некоторое $y\in\{-1,1\}$ и $k\in\{1,\ldots,K\}$. Рассмотрим массив

$$\mathcal{A}(y,k):=\{\widetilde{X}^{j_y^m}, j_y^m\in S_k^y(n_y), n_y=n_y(n)\},$$

где $n\in\mathbb{N}$. Усиленный закон больших чисел для массивов, сформулированный в теореме 2.1 [92], позволяет утверждать, что когда $p=2$, $\psi(x)=|x|^{5/2}$,

$x \in \mathbb{R}$, и $k \in \mathbb{N}$ (здесь мы сохраняем обозначения p, ψ и k, используемые в работе [92] для других объектов), выполняется

$$\frac{1}{\sharp S_k^y(n_y)} \sum_{j \in S_k^y(n_y)} \mathbb{I}\{f(X^j) \neq y\} \to \mathsf{P}(f(X^{j_y^1}) \neq y) \text{ п.н., } n \to \infty. \tag{15}$$

Более точно, в [92] рассматривается случай, когда наблюдения представлены в виде треугольных массивов. Подобные структуры изучаются во множестве работ, например, [6]. Изучение $\mathcal{A}(y,k)$ может быть сведено к рассмотрению последовательности треугольных массивов. Действительно, в силу (7) вклад $\frac{1}{\sharp S_K^y(n_y)} \sum_{j \in S_K^y(n_y)} \mathbb{I}\{f(X^j) \neq y\}$ в асимптотическое поведение последних слагаемых, входящих в $S_K^y(n_y)$, с индексами, большими чем $[n_y/K]K$, пренебрежимо мал. Следовательно, для каждого k, принадлежащего множеству $\{1, \dots, K\}$, мы получаем массив $\widetilde{\mathcal{A}}(y,k)$ со строками, содержащими $r_1, r_2, \dots$ элементов ($(r_i)_{i \in \mathbb{N}}$ зависит от $(n_y(n))_{n \in \mathbb{N}}$), где $r_1 \leqslant r_2 \leqslant \dots$ ($\widetilde{\mathcal{A}}(y,k) = \mathcal{A}(y,k)$ для $k = 1, \dots, K-1$). Кроме того, если $r_m = \dots = r_q$, то $q - m \leqslant I$, где $I = [Ka]$. Теперь можем рассматривать отдельно каждый из I массивов (беря по одной строке из I строк одинаковой длины) со строго увеличивающимся числом элементов в строке. Тогда, если необходимо, введём вспомогательные строки с элементами $\mathsf{P}(f(X^{j_y^1}) \neq y)$, чтобы получить треугольный массив и применить усиленный закон больших чисел из [92]. Лемма 1.1 показывает, что $\mathsf{P}(f(X^{j_y^1}) \neq y) = \mathsf{P}(f(X) \neq y | Y = y)$. Следовательно,

$$\frac{2}{K} \sum_{k=1}^{K} \sum_{y \in \{-1,1\}} \psi(y) \sum_{j \in S_k^y(n_y)} \frac{\mathbb{I}\{f(X^j) \neq y\} \mathsf{P}(Y = y)}{\sharp S_k^y(n_y)} \to Err(f) \text{ п.н., } n \to \infty.$$

Таким образом, соотношение (14) выполняется тогда и только тогда, когда

$$\sum_{k=1}^{K} \sum_{y \in \{-1,1\}} \psi(y) \frac{1}{\sharp S_k^y(n_y)} \sum_{j \in S_k^y(n_y)} Z_{n,k}^j(y) \to 0 \text{ п.н., } n \to \infty,$$

где для $y \in \{-1, 1\}$ и

$$Z_{n,k}^j(y) = (\mathbb{I}\{f_{PA}(X^j, \zeta_n(\overline{S_k(n)}), \widehat{\mathsf{P}}_n) \neq y\} - \mathbb{I}\{f(X^j) \neq y\}) \mathsf{P}(Y = y).$$

Положим $F_{n,k}(x,y) := \mathbb{I}\{f_{PA}(x, \zeta_n(\overline{S_k(n)}), \widehat{\mathsf{P}}_n) \neq y\} - \mathbb{I}\{f(x) \neq y\}$. Для каждого $y \in \{-1, 1\}$, $n \in \mathbb{N}$ и $k = 1, \dots, K$ введём случайные переменные

$$Q_{n,k,y} = \frac{1}{\sharp S_k^y(n_y)} \sum_{j \in S_k^y(n_y)} F_{n,k}(X^j, y).$$

Тогда (14) эквивалентно соотношению

$$\sum_{k=1}^{K}\sum_{y\in\{-1,1\}}\psi(y)\mathsf{P}(Y=y)Q_{n,k,y}\to 0 \text{ п.н., } n\to\infty.$$

Заметим, что $Q_{n,k,y}=Q^{(1)}_{n,k,y}+Q^{(2)}_{n,k,y}$, где

$$Q^{(1)}_{n,k,y}=\frac{1}{\sharp S^y_k(n_y)}\sum_{j\in S^y_k(n_y)}\mathbb{I}\{X^j\in U\}F_{n,k}(X^j,y),$$

$$Q^{(2)}_{n,k,y}=\frac{1}{\sharp S^y_k(n_y)}\sum_{j\in S^y_k(n_y)}\mathbb{I}\{X^j\notin U\}F_{n,k}(X^j,y).$$

Имеем

$$|Q^{(1)}_{n,k,y}|\leqslant\sum_{x\in U}\frac{1}{\sharp S^y_k(n_y)}\sum_{j\in S^y_k(n_y)}|\mathbb{I}\{f_{PA}(x,\zeta_n(\overline{S_k(n)}),\widehat{\mathsf{P}}_n)\neq y\}-\mathbb{I}\{f(x)\neq y\}|.$$

Условие (10) влечёт выполнение соотношения

$$\sum_{k=1}^{K}\sum_{y\in\{-1,1\}}\psi(y)\mathsf{P}(Y=y)Q^{(1)}_{n,k,y}\to 0 \text{ п.н., } n\to\infty.$$

Если $U=\mathbb{X}$, то $Q^{(2)}_{n,k,y}=0$ для всех n,k,y и (14) выполняется. Пусть $U\neq\mathbb{X}$. Тогда

$$V_{n,k}:=\sum_{y\in\{-1,1\}}\psi(y)\mathsf{P}(Y=y)Q^{(2)}_{n,k,y}=\sum_{x\in\mathbb{X}_{-1}\cup\mathbb{X}_1}\sum_{y\in\{-1,1\}}G_{n,k,y}(x),$$

где $n\in\mathbb{N}$, $k=1,\dots,K$, множества $\mathbb{X}_y$ вводились в (13) и

$$G_{n,k,y}(x):=\frac{\psi(y)\mathsf{P}(Y=y)}{\sharp S^y_k(n_y)}\sum_{j\in S^y_k(n_y)}\mathbb{I}\{X^j=x\}\Big(\mathbb{I}\{f_{PA}(x,\zeta_n(\overline{S_k(n)}),\widehat{\mathsf{P}}_n)\neq y\}-\mathbb{I}\{f(x)\neq y\}\Big).$$

Пусть $R^y_j(x):=\mathbb{I}\{X^j=x\}\psi(y)\mathsf{P}(Y=y)$, $y\in\{-1,1\}$, $j\in\mathbb{N}$. Имеют место следующие равенства:

$$\sum_{x\in\mathbb{X}_1}\sum_{y\in\{-1,1\}}G_{n,k,y}(x)$$
$$=\sum_{x\in\mathbb{X}_1}\mathbb{I}\{f_{PA}(x,\zeta_n(\overline{S_k(n)}),\widehat{\mathsf{P}}_n)=-1\}\sum_{y\in\{-1,1\}}\frac{y}{\sharp S^y_k(n_y)}\sum_{j\in S^y_k(n_y)}R^y_j(x),$$

$$\sum_{x\in\mathbb{X}_{-1}}\sum_{y\in\{-1,1\}} G_{n,k,y}(x)$$
$$=-\sum_{x\in\mathbb{X}_{-1}} \mathbb{I}\{f_{PA}(x,\zeta_n(\overline{S_k(n)}),\widehat{\mathsf{P}}_n)=1\}\sum_{y\in\{-1,1\}}\frac{y}{\sharp S_k^y(n_y)}\sum_{j\in S_k^y(n_y)} R_j^y(x).$$

Аналогично выводу (15) может быть показано, что для всех $y\in\{-1,1\}$, $k=1,\dots,K$ и $x\in\mathbb{X}$, соотношение

$$\frac{1}{\sharp S_k^y(n_y)}\sum_{j\in S_k^y(n_y)} R_j^y(x)\to J(x,y) \text{ п.н., } n\to\infty,$$

выполняется. Здесь $J(x,y):=\psi(y)\mathsf{P}(X=x,Y=y)$.

Отсюда для всех $\varepsilon>0$, $x\in\mathbb{X}$, $y\in\{-1,1\}$ и почти всех $\omega\in\Omega$, существует $n_1(\omega,\varepsilon,k,y)$ такое, что

$$\left|\frac{1}{\sharp S_k^y(n_y)}\sum_{j\in S_k^y(n_y)}\big(R_j^y(x)-J(x,y)\big)\right|<\varepsilon \tag{16}$$

для $n>n_1(\omega,\varepsilon,k,y)$. Заметим, что $L(x)=\sum\limits_{y\in\{-1,1\}} yJ(x,y)$. Таким образом,

$$V_{n,k}=\sum_{x\in\mathbb{X}_1}\mathbb{I}\{f_{PA}(x,\zeta_n(\overline{S_k(n)}),\widehat{\mathsf{P}}_n)=-1\}L(x)$$
$$+\sum_{x\in\mathbb{X}_1}\mathbb{I}\{f_{PA}(x,\zeta_n(\overline{S_k(n)}),\widehat{\mathsf{P}}_n)=-1\}\sum_{y\in\{-1,1\}}\frac{1}{\sharp S_k^y(n_y)}\sum_{j\in S_k^y(n_y)} y(R_j^y(x)-J(x,y))$$
$$-\sum_{x\in\mathbb{X}_{-1}}\mathbb{I}\{f_{PA}(x,\zeta_n(\overline{S_k(n)}),\widehat{\mathsf{P}}_n)=1\}L(x)$$
$$-\sum_{x\in\mathbb{X}_{-1}}\mathbb{I}\{f_{PA}(x,\zeta_n(\overline{S_k(n)}),\widehat{\mathsf{P}}_n)=1\}\sum_{y\in\{-1,1\}}\frac{1}{\sharp S_k^y(n_y)}\sum_{j\in S_k^y(n_y)} y(R_j^y(x)-J(x,y)).$$

Учитывая (16), для произвольного $\varepsilon > 0$ мы получаем неравенство

$$\Bigg| \sum_{k=1}^{K} \Big(\sum_{x \in \mathbb{X}_1} \mathbb{I}\{f_{PA}(x, \zeta_n(\overline{S_k(n)}), \widehat{\mathsf{P}}_n) = -1\} \times \sum_{y \in \{-1,1\}} \frac{1}{\sharp S_k^y(n_y)} \sum_{j \in S_k^y(n_y)} y(R_j^y(x) - J(x,y)) - \sum_{x \in \mathbb{X}_{-1}} \mathbb{I}\{f_{PA}(x, \zeta_n(\overline{S_k(n)}), \widehat{\mathsf{P}}_n) = 1\} \times \sum_{y \in \{-1,1\}} \frac{1}{\sharp S_k^y(n_y)} \sum_{j \in S_k^y(n_y)} y(R_j^y(x) - J(x,y)) \Big) \Bigg| < 2K\varepsilon \sharp \mathbb{X},$$

когда $n > n_1(\omega, \varepsilon) := \max\limits_{k=1,\dots,K, y \in \{-1,1\}} n_1(\omega, \varepsilon, k, y)$. Таким образом, $\sum\limits_{k=1}^{K} V_{n,k} \to 0$ п.н. при $n \to \infty$ тогда и только тогда, когда условие (12) выполняется. $\square$

Замечание 1.4 *Теорема 1.3 показывает, какие именно условия следует проверять вне «хорошего» множества U (где f_{PA} приближает f поточечно), чтобы гарантировать выполнение* (11). *Чтобы пояснить условие* (12), *мы перепишем его (ср. с еоремой 1 [29]) следующим образом:*

$$\sum_{k=1}^{K} \sum_{y \in \{-1,1\}} \sum_{x \in \mathbb{X}_y} y \mathbb{I}\{f_{PA}(x, \zeta_n(\overline{S_k(n)}), \widehat{\mathsf{P}}_n) \neq f(x)\} L(x) \to 0 \ \textit{п.н.}, \quad n \to \infty.$$

Отметим также, что условие (12) *совпадает с соответствующим условием для случая выборки, состоящей из независимых одинаково распределённых случайный величин. (ср. с теоремой 1 [3]).*

Замечание 1.5 *Согласно Следствию 1 [3], выбор множества*

$$U = \left\{ x \in M \colon \mathsf{P}(Y = 1 | X = x) \neq \frac{\psi(-1)}{\psi(-1) + \psi(1)} \right\}, \tag{17}$$

где M определено в (2), *влечёт выполнение* (12), *а значит задача состоит в том, чтобы показать, что* (10) *выполняется на этом U для рассматриваемого f и подходящего f_{PA}.*

Замечание 1.6 *Существует множество различных способов получить сильно состоятельные оценки для $\widehat{\mathsf{P}}_n^y$, $y \in \{-1, 1\}$. Первый способ заключается в том, чтобы оценить $\mathsf{P}(Y = y)$, $y \in \{-1, 1\}$, с помощью другой*

выборки. В этом случае необходимо, чтобы наблюдения выборок подчинялись одному и тому же закону распределения. Второй подход подразумевает построение оценок $\mathsf{P}(Y=y)$ *одновременно с формированием выборки* ζ_n. *Более точно,* $\widehat{\mathsf{P}}_n^y$ *является частотной оценкой* $\mathsf{P}(Y=y)$, *построенной по случайному числу наблюдений* $Y^1, Y^2, \ldots, Y^{\widetilde{n}}$, *где* $\widetilde{n} = \max\{j_1^{n_1}, j_{-1}^{n_{-1}}\}$. *Такая оценка является сильно состоятельной, поскольку* $n \to \infty$ *и* $\widetilde{n} \geqslant n$ *п.н. В разделе 1.1 упоминалось, что выбор* ψ, *приведённый в* (3), *является естественным. Более того, для такой функции* ψ *мы можем упростить выражение* (8), *а именно, в данном случае оно не будет содержать* $\widehat{\psi}(y, \zeta_n(S_k(n)), \widehat{\mathsf{P}}_n)$ *и* $\widehat{\mathsf{P}}_n^y$.

Оценки $\mathsf{P}(X \in B, Y = y)$, где $B \subset \mathbb{X}$ и $y \in \{-1, 1\}$ нужны в том случае, когда $\psi(y)$ зависит от совместного распределения X и Y. Для оценивания такой функции выборка ζ_n не подойдёт, однако формула Байеса поможет построить желаемые оценки. Такой подход в задаче классификации применялся в работе [78].

Пример. Мы показали, что как с помощью стратифицированной выборки ζ_n и оптимальной функции f^* можно построить f_{PA}, чтобы получить результат, описанный в теореме 1.3.

Теорема Байеса утверждает, что для $B \subset \mathbb{X}$

$$\mathsf{P}(Y=1|X\in B) = \frac{\mathsf{P}(X\in B|Y=1)\mathsf{P}(Y=1)}{\mathsf{P}(X\in B|Y=1)\mathsf{P}(Y=1)+\mathsf{P}(X\in B|Y=-1)\mathsf{P}(Y=-1)}.$$

Следовательно, оценки для $\mathsf{P}(Y=1|X\in B)$ можно построить с помощью ζ_n и $\widehat{\mathsf{P}}_n$. Принцип подстановки (*plug-in*) объясняет, как получить подходящую аппроксимацию f_{PA} для оптимальной функции f^*. Согласно этому принципу нужно построить оценки $\mathsf{P}(Y=1|X=x)$ и $\frac{\psi(-1)}{\psi(-1)+\psi(1)}$. Тогда остаётся подставить их в определение A, введённое в (2).

Для $y \in \{-1,1\}$, $k = 1, \ldots, K$ и $N \in \mathbb{N}$, положим $W_k^y(n_y) := \cup_{m\neq k} S_m^y(n_y)$ и предположим, что (9) выполняется. Рассмотрим следующие оценки для $\mathsf{P}(Y=1|X=x)$ и $\frac{\psi(-1)}{\psi(-1)+\psi(1)}$, соответственно

$$g(x, \zeta_n(\overline{S_k(n)}), \widehat{\mathsf{P}}_n) := \frac{\widehat{\mathsf{P}}_N^1 I_1(x, \zeta_n(\overline{S_k(n)}), \widehat{\mathsf{P}}_n)}{\widehat{\mathsf{P}}_n^{-1} I_{-1}(x, \zeta_n(\overline{S_k(n)}), \widehat{\mathsf{P}}_n) + \widehat{\mathsf{P}}_N^1 I_1(x, \zeta_n(\overline{S_k(n)}), \widehat{\mathsf{P}}_n)},$$

$$h(\zeta_n(\overline{S_k(n)}), \widehat{\mathsf{P}}_n) := \frac{\widehat{\psi}(-1, \zeta_n(\overline{S_k(n)}), \widehat{\mathsf{P}}_n)}{\widehat{\psi}(-1, \zeta_n(\overline{S_k(n)}), \widehat{\mathsf{P}}_n) + \widehat{\psi}(1, \zeta_n(\overline{S_k(n)}), \widehat{\mathsf{P}}_n)} \qquad (18)$$

где

$$I_y(x,\zeta_n(\overline{S_k(n)}),\widehat{\mathsf{P}}_n) := \frac{1}{\sharp W_k^y(n_y)} \sum_{j\in W_k^y(n_y)} \mathbb{I}(X^j = x),\ \ y\in\{-1,1\},\ \ x\in\mathbb{X}.$$

Как обычно, полагаем $0/0 := 0$. Определим

$$A_n = \{x\in\mathbb{X}\colon g(x,\zeta_n(\overline{S_k(n)}),\widehat{\mathsf{P}}_n) > h(\zeta_n(\overline{S_k(n)}),\widehat{\mathsf{P}}_n)\}$$

и

$$f_{PA}(x,\zeta_n(\overline{S_k(n)}),\widehat{\mathsf{P}}_n) := \mathbb{I}\{A_n\}(x) - \mathbb{I}\{\overline{A_n}\}(x). \tag{19}$$

Аналогично доказательству (15) можно установить, что для $y\in\{-1,1\}$ и $x\in\mathbb{X}$

$$I_y(x,\zeta_n(\overline{S_k(n)}),\widehat{\mathsf{P}}_n) \to \mathsf{P}(X=x|Y=y) \text{ п.н., } n\to\infty.$$

Используя (6) и (9), приходим к следующим соотношениям:

$$g(x,\zeta_n(\overline{S_k(n)}),\widehat{\mathsf{P}}_n) \to \mathsf{P}(Y=1|X=x) \text{ п.н., } n\to\infty,$$
$$h(\zeta_n(\overline{S_k(n)}),\widehat{\mathsf{P}}_n) \to \frac{\psi(-1)}{\psi(-1)+\psi(1)} \text{ п.н., } n\to\infty.$$

Для всех $x\in A$ имеем $\mathsf{P}(Y=1|X=x) > \frac{\psi(-1)}{\psi(-1)+\psi(1)}$. Следовательно, для любых $\varepsilon>0$ таких, что

$$\mathsf{P}(Y=1|X=x) - \varepsilon > \frac{\psi(-1)}{\psi(-1)+\psi(1)} + \varepsilon,$$

почти всех $\omega\in\Omega$ и всех $x\in A$ существуют $n_2(\omega)$ и $n_3(\omega)$ такие, что

$$g(x,\zeta_n(\overline{S_k(n)}),\widehat{\mathsf{P}}_n) > \mathsf{P}(Y=1|X=x) - \varepsilon$$

при $n > n_2(\omega)$ и

$$h(\zeta_n(\overline{S_k(n)}),\widehat{\mathsf{P}}_n) < \frac{\psi(-1)}{\psi(-1)+\psi(1)} + \varepsilon$$

для $n > n_3(\omega)$. Таким образом, для почти всех $\omega\in\Omega$, любых $x\in A$ и $n>\max\{n_2(\omega),n_3(\omega)\}$ равенства $f(x)=1$ и $f_{PA}(x,\zeta_n(\overline{S_k(n)}),\widehat{\mathsf{P}}_n)=1$ верны. Похожим образом, для

$$x\in U\cap\overline{A} = \left\{x\in M\colon \mathsf{P}(Y=1|X=x) < \frac{\psi(-1)}{\psi(-1)+\psi(1)}\right\}$$

можно показать, что для почти всех $\omega \in \Omega$ существует такое $n_4(\omega) \in \mathbb{N}$, что для $n > n_4(\omega)$, равенства $f(x) = -1$ и $f_{PA}(x, \zeta_n(\overline{S_k(n)}), \widehat{\mathsf{P}_N}) = -1$ справедливы. Значит, условие (10) выполняется для U, определенного формулой (17). Следовательно, оценка $Err(f^*)$, фигурировавшая в теореме 1.3 сильно состоятельна, если $f_{PA}(x, \zeta_n(\overline{S_k(n)}), \widehat{\mathsf{P}}_n)$ определяется формулой (19). Желаемый алгоритм f_{PA} построен. $\square$

Если ψ вводится формулой (3), то U и A имеют следующий вид:

$$U = \{x \in M \colon \mathsf{P}(Y = 1|X = x) \neq \mathsf{P}(Y = 1)\},$$
$$A = \{x \in M \colon \mathsf{P}(Y = 1|X = x) > \mathsf{P}(Y = 1)\}.$$

Следовательно, вместо (18) мы можем положить $h(\zeta_n(\overline{S_k(n)}), \widehat{\mathsf{P}}_n) := \widehat{\mathsf{P}}_n^1$ и упростить наш пример

Теперь рассмотрим ситуацию, когда переменная отклика Y зависит только от части факторов. Пусть $\{k_1, \ldots, k_r\} \subset \{1, \ldots, d\}$, где $r < d$. Будем говорить (ср. [21], глава 2), что факторы $X_{k_1}, \ldots, X_{k_r}$ являются значимыми (или набор индексов $k_1, \ldots, k_r$ является значимым), если для каждого $x = (x_1, \ldots, x_d) \in M$

$$\mathsf{P}(Y = 1|X_1 = x_1, \ldots, X_d = x_d) = \mathsf{P}(Y = 1|X_{k_1} = x_{k_1}, \ldots, X_{k_r} = x_{k_r}). \quad (20)$$

Для произвольного набора индексов $\{m_1, \ldots, m_r\} \subset \{1, \ldots, n\}$ положим

$$f^{m_1,\ldots,m_r}(x)$$
$$= \begin{cases} 1, & \text{если } \mathsf{P}(Y = 1|X_{m_1} = x_{m_1}, \ldots, X_{m_r} = x_{m_r}) > \frac{\psi(-1)}{\psi(-1)+\psi(1)}, \quad x \in M, \\ -1, & \text{иначе.} \end{cases}$$

Если $\{k_1, \ldots, k_r\}$ является значимым набором, то оптимальная функцией f^* будет являться $f^{k_1,\ldots,k_r}$. Следовательно, для любого значимого набора $\{k_1, \ldots, k_r\} \subset \{1, \ldots, d\}$ и произвольного $\{m_1, \ldots, m_r\} \subset \{1, \ldots, d\}$

$$Err(f^{k_1,\ldots,k_r}) \leqslant Err(f^{m_1,\ldots,m_r}).$$

Для любого подмножества $\{m_1, \ldots, m_r\} \subset \{1, \ldots, n\}$, каждого $x \in \mathbb{X}$ и произвольной функции ψ, рассмотрим алгоритм предсказания $f_{PA}^{m_1,\ldots,m_r}$ такой, что

$$f_{PA}^{m_1,\ldots,m_r}(x, \zeta_n(\overline{S_k(n)}), \widehat{\mathsf{P}}_n)$$
$$= \begin{cases} 1, & \text{если } g^{m_1,\ldots,m_r}(x, \zeta_n(\overline{S_k(n)}), \widehat{\mathsf{P}}_n) > h(\zeta_n(\overline{S_k(n)}), \widehat{\mathsf{P}}_n), \\ -1, & \text{иначе,} \end{cases}$$

где $k = 1, \ldots, K$, h определяется формулой (18) и

$$g^{m_1,\ldots,m_r}(x, \zeta_n(\overline{S_k(n)}), \widehat{\mathsf{P}}_n) = \frac{\widehat{\mathsf{P}}_N^1 I_1^{m_1,\ldots,m_r}(x, \zeta_n(\overline{S_k(n)}), \widehat{\mathsf{P}}_n)}{\widehat{\mathsf{P}}_N^{-1} I_{-1}^{m_1,\ldots,m_r}(x, \zeta_n(\overline{S_k(n)}), \widehat{\mathsf{P}}_n) + \widehat{\mathsf{P}}_N^1 I_1^{m_1,\ldots,m_r}(x, \zeta_n(\overline{S_k(n)}), \widehat{\mathsf{P}}_n)}.$$

Здесь для $y \in \{-1, 1\}$ и $x \in \mathbb{X}$,

$$I_y^{m_1,\ldots,m_r}(x, \zeta_n(\overline{S_k(n)}), \widehat{\mathsf{P}}_n) = \frac{1}{\sharp W_k^y(n_y)} \sum_{j \in W_k^y(n_y)} \mathbb{I}(X_{m_1}^j = x_{m_1}, \ldots, X_{m_r}^j = x_{m_r}).$$

Положим

$$U := \left\{ x \in M_{m_1,\ldots,m_r} \colon \mathsf{P}(Y = 1 | X_{m_1} = x_{m_1}, \ldots, X_{m_r} = x_{m_r}) \neq \frac{\psi(-1)}{\psi(-1) + \psi(1)} \right\},$$

где $M_{m_1,\ldots,m_r} = \{x \in \mathbb{X} \colon \mathsf{P}(X_{m_1} = x_{m_1}, \ldots, X_{m_r} = x_{m_r}) > 0\}$. Аналогично [3], можно показать, что для значимого набора $\{k_1, \ldots, k_r\} \subset \{1, \ldots, d\}$, произвольного набора $\{m_1, \ldots, m_r\} \subset \{1, \ldots, d\}$, любого $\varepsilon > 0$ и почти всех $\omega \in \Omega$,

$$\widehat{Err}_K(f_{PA}^{k_1,\ldots,k_r}, \zeta_n, \widehat{\mathsf{P}}_n) \leqslant \widehat{Err}_K(f_{PA}^{m_1,\ldots,m_r}, \zeta_n, \widehat{\mathsf{P}}_n) + \varepsilon \text{ п.н.},$$

если n достаточно велико. Таким образом, естественно выбрать в качестве значимых факторов такой набор $X_{k_1}, \ldots, X_{k_r}$, на котором достигается минимум ошибки предсказания $\widehat{Err}_K(f_{PA}^{k_1,\ldots,k_r}, \zeta_n, \widehat{\mathsf{P}}_n)$.

1.4 Стоимостной подход к проведению экспериментов и XOR-модель

В данном разделе мы сравним два подхода к применению MDR-EFE метода для различных схем проведения экспериментов. Первый подход описан в работе [3] и заключается в использовании неслучайного числа независимых одинаково распределённых случайных наблюдений. Второй подход был описан в предыдущем разделе данной диссертации и основан на использовании стратифицированных выборок. Более точно, стратификация означает разделение наблюдений на основе значения исследуемой переменной отклика. Мы будем обозначать эти подходы iMDR-EFE и sMDR-EFE, соответственно. Может показаться, что главным недостатком рассматриваемого в разделе 1.3 подхода является слишком большое количество независимых наблюдений, необходимое для формирования стратифицированной

выборки из N элементов с фиксированным отношением размеров групп, когда вероятность $\mathsf{P}(Y = 1)$ очень мала.

Кроме того, по построению метода мы пропускаем большое количество наблюдений $Y^i = -1$. Однако стоит отметить, что на стадии формирования выборки не требуется никакой информации о значениях предикторов X^i, поэтому для пропущенных наблюдений Y^i не нужно измерять X^i. Это существенно в ситуации, когда измерение Y дешевле, чем измерение X (поскольку X состоит из множества компонент $X_1, \ldots, X_n$). Таким образом, сравнение iMDR-EFE и sMDR-EFE методов не представляет интереса для выборок равного размера и должно учитывать наблюдения, пропущенные при формировании выборки для sMDR-EFE.

Мы предлагаем сравнивать два метода в смысле общей стоимости эксперимента. Предположим, что имеется фиксированная сумма денег C ($C \in \mathbb{N}$), выделенная на проведение исследования. Пусть каждое наблюдение (X^i, Y^i) стоит 1 и пусть отношение стоимости измерения Y к стоимости измерения X равно $w \in \mathbb{R}_+$. Мы рассмотрим максимальные размеры выборок $s_{ind}(C, w)$ и $s_{str}(C, w)$, которые доступны для использования методов iMDR-EFE и sMDR-EFE, соответственно. Пусть $C_{ind}(n, w)$ — общая стоимость выборки размера n, состоящей из независимых наблюдений. Пусть $C_{str}(n, w)$ — общая стоимость стратифицированной выборки размера n. Тогда

$$s_{ind}(C, w) = \max\{n \in \mathbb{N} \colon C_{ind}(n, w) \leqslant C\},$$
$$s_{str}(C, w) = \max\{n \in \mathbb{N} \colon C_{str}(n, w) \leqslant C\}.$$

Очевидно, $C_{ind}(n, w) = n$ и, следовательно, $s_{ind}(C, w) = C$. Отметим, что это число неслучайно и известно ещё до проведения эксперимента.

Напомним, что согласно замечанию 1.6, $\widetilde{n} = \widetilde{n}(n, a, \omega) = \max\{j_{-1}^{n_{-1}}, j_1^{n_1}\}$. Значит,

$$C_{str}(n, w, \omega) = \frac{1}{w+1} n + \frac{w}{w+1} \widetilde{n}(n, a, \omega),$$

потому что мы должны измерить X в n наблюдениях (которые включаются в выборку) и измерить Y в $\widetilde{n}$ наблюдениях. Таким образом, $C_{str}(n, w)$ является случайной переменной, значение которой неизвестно перед проведением эксперимента. В общем случае $s_{str}(C, w)$ также является случайной переменной, однако мы можем выбрать такое n, что $C_{str}(n, w)$ не превзойдёт C с вероятностью не менее $1 - \alpha$, $\alpha \in (0, 1)$. Предполагается, что в случае, когда стоимость эксперимента превышает C с заданной малой вероятностью, то дополнительные расходы могут быть покрыты за счёт некоторого резервного фонда. Теперь найдём такое $s'_{str}(C, w, \alpha)$, что

$$s'_{str}(C, w, \alpha) = \max\left\{n \in \mathbb{N} \colon \mathsf{P}(C_{str}(n, w) \leqslant C) \geqslant 1 - \alpha\right\}.$$

Очевидно, что $\widetilde{n}(n,a,\omega) \geqslant n$ для всех $\omega \in \Omega$. Следовательно для всех $n \in \mathbb{N}$ и $\omega \in \Omega$

$$C_{str}(n,w,\omega) = n + \frac{w}{w+1}(\tilde{n}(n,a,\omega) - n) \geqslant n.$$

Следовательно, для всех $n > C$,

$$\mathsf{P}(C_{str}(n,w) \leqslant C) \leqslant \mathsf{P}(C_{str}(n,w) < n) = 0.$$

Таким образом, для произвольного $C \in \mathbb{N}$, заданных $\alpha \in (0,1)$ и $w > 0$, мы получаем, что

$$s'_{str}(C,w,\alpha) = \max\left\{n \in \{1,2,\ldots,C\} \colon \mathsf{P}(C_{str}(n,w) \leqslant C) \geqslant 1-\alpha\right\}. \quad (21)$$

Стоит отметить, что $s'_{str}(C,w,\alpha)$ является неслучайной величиной, зависящей от параметров и распределения $\widetilde{n}$. Сформулируем следующий результат:

Лемма 1.7 *Для всех* $m \in \{0,1,2\ldots\}$

$$\mathsf{P}(\widetilde{n} = m) = \begin{cases} 0, & \text{если } m < n, \\ \mathsf{P}(\eta_{-1} = m - n_{-1}) + \mathsf{P}(\eta_1 = m - n_1) & \text{иначе,} \end{cases} \quad (22)$$

где $\eta_{-1} \sim NB(n_{-1}, \mathsf{P}(Y=1))$, $\eta_1 \sim NB(n_1, \mathsf{P}(Y=-1))$.

Доказательство. $\mathsf{P}(\widetilde{n} = m) = 0$ для всех $m \in \{0,1,2,\ldots,n-1\}$, поскольку $\widetilde{n}(n,a,\omega) \geq n$ для всех $\omega \in \Omega$. Если $m \in \{n, n+1, \ldots\}$, то

$$\begin{aligned}\mathsf{P}(\widetilde{n} = m) &= \mathsf{P}\left(\max\{j_{-1}^{n_{-1}}, j_1^{n_1}\} = m\right) \\ &= \mathsf{P}\left(j_{-1}^{n_{-1}} = m, j_1^{n_1} < m\right) + \mathsf{P}\left(j_{-1}^{n_{-1}} < m, j_1^{n_1} = m\right) \\ &\qquad + \mathsf{P}\left(j_{-1}^{n_{-1}} = m, j_1^{n_1} = m\right).\end{aligned}$$

Пусть $\widetilde{S}_y(k) = \sharp\{i \in \{1,\ldots,k\} \colon Y^i = y\}$. Заметим, что для всех $k \in \mathbb{N}$ $\widetilde{S}_y(k) + \widetilde{S}_{-y}(k) = k$. Таким образом, для всех $m \in \{n, n+1, \ldots\}$ и $y \in \{-1, 1\}$,

$$\begin{aligned}\left\{j_y^{n_y} = m\right\} &= \left\{\widetilde{S}_y(m) = n_y\right\} \cap \left\{\widetilde{S}_y(m-1) = n_y - 1\right\} \\ &\subset \left\{\widetilde{S}_y(m-1) = n_y - 1\right\} = \left\{\widetilde{S}_{-y}(m-1) = m - n_y\right\} \\ &\subset \left\{\widetilde{S}_{-y}(m-1) \geqslant n_{-y}\right\} \subset \left\{j_{-y}^{n_{-y}} < m\right\}.\end{aligned}$$

Кроме того, $j_{-1}^{n_{-1}}(\omega) \neq j_1^{n_1}(\omega)$ для всех $\omega \in \Omega$. Значит,

$$\mathsf{P}(\widetilde{n} = m) = \mathsf{P}(j_{-1}^{n_{-1}} = m) + \mathsf{P}(j_1^{n_1} = m).$$

В разделе 1.2 было показано, что $j_y^{n_y}$ имеет такое же распределение, как и $n_y + \eta_y$. Таким образом, (22) выполнено. □

Замечание 1.8 *Лемма 1.7 показывает, что $\widetilde{n}$ зависит от известных параметров n_{-1}, n_1 и, в общем случае, неизвестной вероятности $\mathsf{P}(Y{=}1)$. Однако, если вероятность $\mathsf{P}(Y = 1)$ известна или мы имеем оценки, построенные по выборке $\xi_{\widetilde{n}}$ или другой выборке (см. замечание 1.6), то $s'_{str}(C, w, \alpha)$ может быть вычислено или оценено.*

С помощью леммы 1.7 мы можем переписать $\mathsf{P}\left(C_{str}(n, w) \leqslant C\right)$ следующим образом:

$$\begin{aligned}\mathsf{P}\left(C_{str}(n, w) \leqslant C\right) &= \mathsf{P}\left(\widetilde{n}(n, a, \omega) \leqslant \frac{w+1}{w}C - \frac{n}{w}\right) \\ &= \mathsf{P}\left(n_1 \leqslant \eta_{-1} \leqslant \frac{w+1}{w}C - \frac{n}{w} - n_{-1}\right) \\ &\qquad + \mathsf{P}\left(n_{-1} \leqslant \eta_1 \leqslant \frac{w+1}{w}C - \frac{n}{w} - n_1\right), \quad (23)\end{aligned}$$

где η_{-1} и η_1 определены выше. Таким образом, для некоторого n вероятность $\mathsf{P}\left(C_{str}(n, w) \leqslant C\right)$ может быть вычислена или оценена (см. замечание 1.8). Кроме того,

$$\mathsf{P}\left(C_{str}(n, w) \leqslant C\right) \geqslant \mathsf{P}\left(C_{str}(n+1, \omega) \leqslant C\right) \quad (24)$$

для всех $n \in \mathbb{N}$ поскольку для каждого $\omega \in \Omega$,

$$\frac{1}{w+1}n + \frac{w}{w+1}\tilde{n}(n, a, \omega) \leqslant \frac{1}{w+1}(n+1) + \frac{w}{w+1}\widetilde{n}(n+1, a, \omega).$$

На использовании формул (23) и (24) основан следующий алгоритм. Если вероятность $p := \mathsf{P}(Y = 1)$ известна, то мы вычисляем $\mathsf{P}\left(C_{str}(n, w) \leqslant C\right)$ по формуле (23) для $n \in \mathbb{N}$ и $C \in \mathbb{N}$ и определяем максимальное значение n из множества $\{1, \ldots, C\}$ такое, что (21) выполняется. Так мы определим $s'_{str}(C, w, \alpha)$. Если значение p неизвестно, но может быть оценено с помощью оценки $\hat{p}_{n'}$ такой, что $\hat{p}_{n'} \to p$ почти наверное при $n' \to \infty$ (например, это может быть оценка, построенная на основе другой выборки), то для

всех фиксированных C, n и w мы можем оценить $\mathsf{P}\left(C_{str}(n,w) \leqslant C\right)$ с помощью равенства

$$\begin{aligned}&\nu_{n'}(n,w,C)\\&= \sum_{n_1 \leqslant m \leqslant L_1} \binom{n_1+m-1}{m} (1-\hat{p}_{n'})^{n_{-1}} \hat{p}_{n'}^{m} + \sum_{n_{-1} \leqslant m \leqslant L_{-1}} \binom{n_{-1}+m-1}{m} \hat{p}_{n'}^{n_1} (1-\hat{p}_{n'})^{m},\end{aligned}$$

где $L_y := \frac{w+1}{w} C - \frac{n}{w} - n_{-y}$, $y \in \{-1, 1\}$. Действительно, $\nu_{n'}(n,w,C)$ стремится к правой части выражения (23) почти наверное при $n' \to \infty$. Тогда можно ввести

$$\widehat{s'}_{str,n'}(C,w,\alpha) := \max\left\{n \in \{1,2,\ldots,C\} \colon \nu_{n'}(n,w,C) \geqslant 1-\alpha\right\}$$

и применить подход, предложенный для случая, когда p известно.

Чтобы сравнить iMDR-EFE и sMDR-EFE мы обратились к известной XOR-модели (см., например, [98]). Эта модель используется в генетике для объяснения эффекта эпистазиса без главного эффекта. Пусть $\mathbb{X} = \{0,1,2\}^n$ ($0, 1, 2$ соответствует числу минорных аллелей в конкретном гене). Предположим, что компоненты случайного вектора $X = (X_1, \ldots, X_n)$ независимы и для каждого $i \in \{1, \ldots, n\}$ существует такое $p_i \in (0, 0.5]$, что

$$\mathsf{P}(X_i = 0) = (1-p_i)^2, \;\; \mathsf{P}(X_i = 1) = 2p_i(1-p_i), \;\; \mathsf{P}(X_i = 2) = p_i^2. \qquad (25)$$

Такая ситуация типична для полногеномного поиска ассоциаций (GWAS), где каждый X_i соответствует одиночному нуклеотидному полиморфизму (SNP) и p_i есть частота появления минорной аллели (MAF). Предположим, что $X_{k_1}, \ldots, X_{k_r}$, где $\{k_1, \ldots, k_r\} \subset \{1, \ldots, d\}$ — набор значимых факторов, описывающих бинарную переменную отклика Y.

В наших численных экспериментах мы используем XOR-модель зависимости между предикторами и переменной отклика. Эта модель является обобщением XOR-модели, описанной в [98] для случая двух значимых факторов. Для каждого $x = (x_1, \ldots, x_d) \in \mathbb{X}$ пусть

$$\mathsf{P}(Y = 1 | X = x) = \begin{cases} \gamma, & \text{если } (x_{k_1} + \ldots + x_{k_r}) \bmod 2 = 1, \\ 0 & \text{иначе,} \end{cases} \qquad (26)$$

где $\gamma \in (0,1)$ и $p_{k_1} = \cdots = p_{k_r} = 0.5$. Таким образом, $(X_{k_1}, \ldots, X_{k_r})$ является набором значимых факторов согласно (20). XOR-модель представляет интерес в силу своего свойства, описанного в лемме 1.9 ниже.

Для $\overline{s} := \{s_1, \ldots, s_q\} \subset \{1, \ldots, d\}$ положим $X_{\overline{s}} := (X_{s_1}, \ldots, X_{s_q})$. Будем писать, что $X_{\overline{s}} = x_{\overline{s}}$, где $x_{\overline{s}} = (x_{s_1}, \ldots, x_{s_q}) \in \{0,1,2\}^q$, если $X_{s_i} = x_{s_i}$ для всех $i = 1, \ldots, q$.

Лемма 1.9 *Пусть* $\mathbb{X}, X, Y$ *и* $\overline{k}$ *определены выше, зависимость между* X *и* Y *описана формулой* (26). *Тогда для любого набора*

$$\overline{m} = \{m_1, \ldots, m_l\} \subset \{1, \ldots, d\}$$

переменная отклика Y *является зависимой с* $X_{\overline{m}}$ *тогда и только тогда, когда* $\overline{k} \subset \overline{m}$.

Доказательство. Формула (26) показывает, что Y и $X_{\overline{m}}$ зависимы, если $\overline{k} \subset \overline{m}$. Пусть теперь $\overline{m} = \overline{u} \cup \overline{v}$, где $\overline{u} \subsetneq \overline{k}$ и $\overline{v} \subsetneq \{1, \ldots, d\} \setminus \overline{k}$. Введём $t := \sharp\overline{u}$. Поскольку $\mathsf{P}(X_{\overline{m}} = x_{\overline{m}}) \neq 0$ для всех $x_{\overline{m}} \in \{0, 1, 2\}^l$, остаётся установить, что

$$\mathsf{P}(Y = 1 | X_{\overline{m}} = x_{\overline{m}}) = \mathsf{P}(Y = 1). \tag{27}$$

Очевидно, что

$$\begin{aligned}
&\mathsf{P}(Y = 1 | X_{\overline{m}} = x_{\overline{m}}) \\
&\quad = \sum_{x_{\overline{k}\setminus\overline{u}} \in \{0,1,2\}^{r-t}} \mathsf{P}(Y = 1 | X_{\overline{u}} = x_{\overline{u}}, X_{\overline{k}\setminus\overline{u}} = x_{\overline{k}\setminus\overline{u}}, X_{\overline{v}} = x_{\overline{v}}) \mathsf{P}(X_{\overline{k}\setminus\overline{u}} = x_{\overline{k}\setminus\overline{u}}) \\
&\quad = \gamma \sum_{x_{\overline{k}\setminus\overline{u}} \in \{0,1,2\}^{r-t}} \mathbb{I}\left\{ \left(\sum_{i \in \overline{k}\setminus\overline{u}} x_i + \sum_{i \in \overline{u}} x_i \right) \mod 2 = 1 \right\} \mathsf{P}(X_{\overline{k}\setminus\overline{u}} = x_{\overline{k}\setminus\overline{u}}) \\
&\quad = \gamma \mathsf{P}\left(\sum_{i \in \overline{k}\setminus\overline{u}} X_i \bmod 2 \neq \sum_{i \in \overline{u}} x_i \bmod 2 \right). \qquad (28)
\end{aligned}$$

В силу (25) закон распределения $(X_1, \ldots, X_d)$ совпадает с законом для $(X'_1, \ldots, X'_d)$, где $X'_i = \sum_{j=1}^{2} B_i^j$ и $B_1^1, B_1^2, \ldots, B_d^1, B_d^2$ являются независимыми бернулиевскими случайными величинами с вероятностью успеха $\mathsf{P}(B_i^k = 1) = p_i$, $k = 1, 2$, $i = 1, \ldots, d$. Значит,

$$Law\left(\sum_{i \in \overline{s}} X_i \right) = Law\left(\sum_{i \in \overline{s}} (B_i^1 + B_i^2) \right)$$

для всех $\overline{s} \subset \{1, \ldots, n\}$. Таким образом, для $\overline{s} := \{s_1, \ldots, s_q\} \subset \overline{k}$ $(q \leqslant r)$, имеем

$$\mathsf{P}\left(\sum_{i \in \overline{s}} X_i \bmod 2 = 0 \right) = \mathsf{P}\left(\sum_{i=1}^{2q} \tilde{B}_i \bmod 2 = 0 \right) = \sum_{j=0}^{q} C_{2q}^{2j} 2^{-2q} = \frac{1}{2},$$

где $\widetilde{B}_1, \ldots, \widetilde{B}_{2r}$ — независимые бернулиевские случайные величины с вероятностью успеха 0.5. Следовательно,

$$\mathsf{P}\left(\sum_{i\in\overline{s}} X_i \bmod 2 = 1\right) = \frac{1}{2}.$$

Учитывая (28), получаем, что $\mathsf{P}(Y = 1|X_{\overline{m}} = x_{\overline{m}}) = \frac{\gamma}{2}$ для любого $x_{\overline{m}} \in \{0,1,2\}^l$. Остаётся доказать, что

$$\mathsf{P}(Y=1) = \sum_{x\in\mathbb{X}} \mathsf{P}(Y=1|X=x)\mathsf{P}(X=x)$$

$$= \sum_{x_{\overline{k}}\in\{0,1,2\}^r} \mathsf{P}(Y=1|X_{\overline{k}}=x_{\overline{k}})\mathsf{P}(X_{\overline{k}}=x_{\overline{k}}) = \gamma\mathsf{P}\left(\sum_{i\in\overline{k}} X_i \bmod 2 = 1\right) = \frac{\gamma}{2}.$$

Таким образом, (27) выполняется. □

Замечание 1.10 *Лемма 1.9 показывает, что только весь набор значимых факторов $X_{k_1}, \ldots, X_{k_r}$ является зависимым с переменной отклика Y в XOR-модели, при этом Y не зависит от любого его поднабора.*

Чтобы завершить этот раздел, обсудим некоторые асимптотические свойства случайной выборки при $C \to \infty$. Для заданного C iMDR-EFE метод включает C наблюдений, в то время как sMDR-EFE (при том же количестве денег C) приводит к n наблюдениям и

$$\frac{1}{w+1}n + \frac{w}{w+1}\max\{j_{-1}^{n-([an]\vee 1)}, j_1^{[an]\vee 1}\} \leqslant C, \tag{29}$$

где $a \in (0,1)$. Понятно, что (29) влечёт $n \leqslant C$. Рассмотрим ситуацию, когда $n = \lambda C$ для некоторого $\lambda \in (0,1)$ такого, что с вероятностью близкой к 1 неравенство (29) выполнено.

Лемма 1.11 *Для произвольных (фиксированных) $a \in (0,1)$ и w неравенство (29) выполняется с вероятностью, стремящейся к 1 при $C \to \infty$ тогда и только когда, когда*

$$\lambda < \lambda_0 := (1+w)\left(1 + w\max\left\{\frac{a}{p}, \frac{1-a}{1-p}\right\}\right)^{-1}. \tag{30}$$

Доказательство. Мы можем переписать (29) следующим образом:

$$\max\{j_{-1}^{[\lambda C]-[a[\lambda C]]}, j_1^{[a[\lambda C]]}\} \leqslant \frac{C(w+1)-[\lambda C]}{w}.$$

Заметим, что

$$\begin{aligned}&\mathsf{P}\left(j_1^{[a[\lambda C]]} \leqslant \frac{C(w+1)-[\lambda C]}{w}\right)\\ &\quad = \mathsf{P}\left(\frac{j_1^{[a[\lambda C]]} - \frac{[a[\lambda C]]}{p}}{\sigma\sqrt{[a[\lambda C]]}} \leqslant \frac{1}{\sigma\sqrt{[a[\lambda C]]}}\left(\frac{C(w+1)-[\lambda C]}{w} - \frac{[a[\lambda C]]}{p}\right)\right),\end{aligned}$$

где $\sigma^2 := \frac{1-p}{p^2}$ (дисперсия случайной переменной, имеющей геометрическое распределение с параметром p). Центральная предельная теорема для независимых одинаково распределённых случайных величин, имеющих конечную дисперсию, влечёт неравенство $j_1^{[a[\lambda C]]} \leqslant \frac{C(w+1)-[\lambda C]}{w}$, которое выполняется с вероятностью, стремящейся к 1 при $C \to \infty$ тогда и только тогда, когда $(w+1-\lambda)/w - a\lambda/p > 0$, поэтому

$$\lambda < (1+w)\left(1+\frac{aw}{p}\right)^{-1}. \tag{31}$$

Аналогично получаем, что

$$\mathsf{P}\left(j_{-1}^{[\lambda C]-[a[\lambda C]]} \leqslant \frac{C(w+1)-[\lambda C]}{w}\right) \to 1, \;\; C \to \infty,$$

тогда и только тогда, когда

$$\lambda < (1+w)\left(1+\frac{(1-a)w}{1-p}\right)^{-1}. \tag{32}$$

Таким образом, утверждение леммы 1.1 выполнено тогда и только тогда, когда (31) и (32) выполнены одновременно, т.е. когда выполнено (30). □

Лемма 1.11 обеспечивает оптимальную границу для λ_0.

1.5 Результаты численных экспериментов

Рассмотрим различные значения C и параметра γ XOR-модели. Отметим, что теперь $\mathsf{P}(Y=1) = \frac{\gamma}{2}$, т.е. $p = \frac{\gamma}{2}$. Мы использовали два уровня значений параметра w. Значение $w = 0$ соответствует случаю, когда

измерения переменной отклика бесплатны. Эти эксперименты могут рассматриваться как предельный случай ситуации, когда стоимость измерения Y низка по сравнению со стоимостью измерения X. Согласно [94] применение MDR метода оправданно для сбалансированных выборок, например, с одинаковым количеством больных и здоровых пациентов. Поэтому мы рассматриваем стратифицированные выборки со значением параметра $a = 0.5$ и только если n удовлетворяет условию из (21). Для каждой комбинации параметров мы генерируем D независимых выборок для применения iMDR-EFE и sMDR-EFE, чтобы оценить их качество с помощью метода Монте Карло.

Значения параметров наших экспериментов приводятся в Таблице 1. Для значимых факторов $X_{k_1}, \ldots, X_{k_r}$ мы полагаем $p_{k_i} = 0.5$, $i = 1, \ldots, r$ (согласно XOR-модели), а для остальных (незначимых) факторов X_i соответствующие значения p_i генерируются из равномерного распределения $U(0.05, 0.5)$ для получения каждой из выборок. Для каждой выборки p_i генерируются независимо. Такая схема генерации p_i была предложена в [36]. Интервал $(0.05, 0.5)$ для равномерного закона был взят, поскольку такие p_i соответствуют значениям MAF.

Для генерации Y при условии X мы использовали бернулиевские случайные величины с вероятностью успеха 0, если $X_{k_1} + \cdots + X_{k_r} \bmod 2 = 0$, и вероятностью γ, если $X_{k_1} + \cdots + X_{k_r} \bmod 2 = 1$.

Выбирая $w = 0.1$, мы предполагаем, что p известно, чтобы использовать детерминированное значение $s'_{str}(C, w, \alpha)$, введённое в (21). Для $w = 0$ размер выборки $s'_{str}(C, 0, \alpha)$ равен C, как и в случае независимых наблюдений. Поэтому это значение неслучайно и фиксировано до проведения эксперимента, и мы оцениваем $\mathsf{P}(Y = 1)$ с помощью $Y^1, \ldots, Y^{\widetilde{C}}$. Здесь $\widetilde{C} = \widetilde{C}(C, a, \omega) = \max\{j_{-1}^{C_{-1}}, j_1^{C_1}\}$, $C_1 = [aC]$ и $C_{-1} = C - [aC]$. Чтобы сравнить качество методов, мы используем метрику TMR *(true model rate)*, которая определяется как

$$TMR := \frac{1}{D} \sum_{d=1}^{D} T_d,$$

где

$$T_d = \begin{cases} 1, & \text{если для } d\text{-й выборки все факторы правильно обнаружены,} \\ 0, & \text{иначе.} \end{cases}$$

Далее фраза «применение iMDR-EFE (sMDR-EFE) к выборке» означает, что мы выбрали такой набор из r факторов, который имеет наименьшее значение $\widehat{Err}_K$.

Параметр	**Значение**
$\psi(y)$	$(\mathsf{P}(Y=y))^{-1}$
D	100
n	100
K	5
r	3
$(k_1,\ldots,k_r)$	(2,3,5)
a	0.5
C	5 значений: $\{100,200,300,400,500\}$
w	2 значения: $\{0,0.1\}$
γ	3 значения: $\{0.05,0.1,0.2\}$
α	0.05

Таблица 1: Параметры экспериментов

Процедура генерации данных может быть описана следующим образом:

Шаг 1. Фиксируем некоторые значения γ и C из таблицы 1, другие параметры кроме w также фиксированы.

Шаг 2. Вычислим $s'_{str}(C,w,\alpha)$, где $w=0.1$ и $\alpha=0.5$, предполагая, что $p=\mathsf{P}(Y=1)$ известно (и равно $\gamma/2$).

Шаг 3. Далее D раз

1. генерируем независимые $p_i \sim U(0.05,0.5)$ для незначимых факторов;
2. генерируем независимые X^j, имеющие закон распределения такой, как описан в (25), и соответствующие Y^j, чтобы получить 3 выборки:
   (a) выборку ξ_C , состоящую из C независимых наблюдений;
   (b) стратифицированную выборку ζ_n, которая состоит из $n = s'_{str}(C,w,\alpha)$ наблюдений (для $w=0.1$);
   (c) стратифицированную выборку ζ_C, которая состоит из C наблюдений (для $w=0$). Одновременно с генерацией наблюдений (X^j,Y^j) для выборки ζ_C оцениваем p (будем говорить «p оценено по наблюдениям $Y^1,\ldots,Y^{\widetilde{C}}$»);

3. применяем iMDR-EFE к ξ_C, p неизвестно (как и $\psi(y)$);
4. применяем iMDR-EFE к ξ_C, p известно, следовательно $\psi(y)$ также известно и

$$\widehat{\psi}(y, \xi_C(\overline{S_k(C)})) = 1/\mathsf{P}(Y = y);$$

5. применяем sMDR-EFE к ζ_n, где $n = s'_{str}(C, w, \alpha)$, p известно, следовательно $\widehat{\mathsf{P}}_n = (1 - p, p)$ и

$$\widehat{\psi}(y, \zeta_n(\overline{S_k(n)}), \widehat{\mathsf{P}}_n) = 1/\mathsf{P}(Y = y);$$

6. применяем sMDR-EFE к ζ_C, p неизвестно и оценивается по наблюдениям $Y^1, \ldots, Y^{\widetilde{C}}$;
7. применяем sMDR-EFE к ζ_C, p известно, поэтому $\widehat{\mathsf{P}}_C = (1 - p, p)$ и

$$\widehat{\psi}(y, \zeta_C(\overline{S_k(C)}), \widehat{\mathsf{P}}_C) = 1/\mathsf{P}(Y = y);$$

8. вычисляем T_d для каждого рассматриваемого метода.

Шаг 4. Оцениваем TMR для каждого подхода.

Результаты экспериментов показаны на рисунках 1 и 2. Для фиксированной стоимости эксперимента C использование стратифицированных выборок позволяет достичь лучших результатов по сравнению с независимыми выборками. Несмотря на то, что $s_{ind}(C, \alpha) > s'_{str}(C, w, \alpha)$ (когда $w > 0$, см. рисунок 2), sMDR-EFE показывает лучшие результаты чем iMDR-EFE во всех 3 моделях для каждого значения C и разных сценариях, касающихся p. Кроме того, маленькие значения w позволяют включить больше наблюдений в стратифицированную выборку. Рисунок 1 показывает, что это приводит к улучшению качества sMDR-EFE метода.

Пусть $a = 0.5$ и $w = 0.1$. Тогда для γ принимающего значения 0.05, 0.1 и 0.2 ($p = \gamma/2$), Лемма 1.11 позволяет вычислить значения λ_0, которые равны 0.366(6), 0.55 и 0.733(3), соответственно. Для любого $a \in (0, 1)$, всех $p \in (0, 1)$ и $w = 0$ мы получаем $\lambda_0 = 1$. Очевидно, что если w близко к 0, то λ_0 близко к 1.

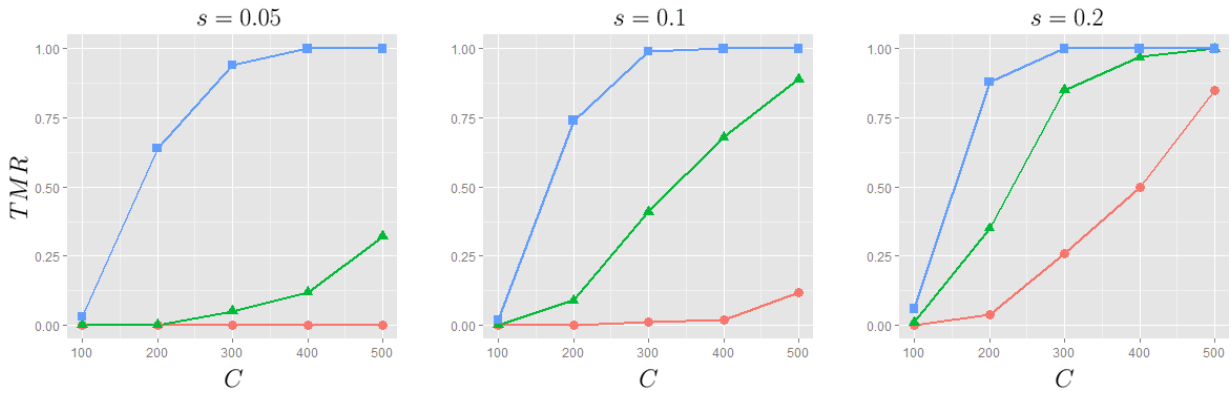

Рис. 1: Качество (TMR) iMDR-EFE и sMDR-EFE для разных значений $\gamma = 2p$.

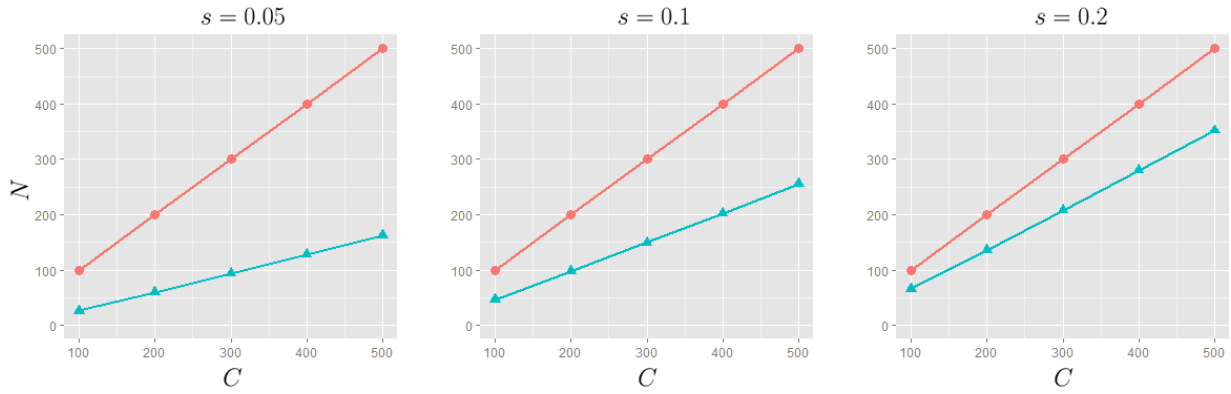

Рис. 2: Размеры выборок, зависящие от стоимости C для разных значений $\gamma = 2p$. Красная линия соответствует iMDR-EFE и sMDR-EFE с $w = 0$, голубая соответствует sMDR-EFE с $w = 0.1$.

2 Оценивание условной энтропии Шеннона

При подготовке данной главы диссертации использован материал публикации [101], выполненной автором в соавторстве с профессором А.В.Булинским. В публикации А.В.Булинскому принадлежит постановка задач и общий подход к их решению, им также доказаны леммы 3.1 и 3.3 (леммы 2.6 и 2.8 в диссертации), все остальные результаты доказаны автором диссертации.

2.1 Постановка задачи

Напомним, что все рассматриваемые в данной главе случайные величины и векторы определены на некотором вероятностном пространстве $(\Omega, \mathcal{F}, \mathsf{P})$. Напомним также, что энтропия Шеннона (см. [85]) дискретной случайной величины Y со значениями в конечном множестве M с распределением $P(y) := \mathsf{P}(Y = y)$, $y \in M$, и (дифференциальная) энтропия случайного вектора X в $\mathbb{R}^d$, имеющего плотность $f(\cdot)$, $x \in \mathbb{R}^d$, относительно меры Лебега μ, определяются следующим образом:

$$H(Y) := -\mathsf{E} \log P(Y) = -\sum_{y \in M} P(y) \log P(y), \tag{33}$$

$$H(X) := -\mathsf{E} \log f(X) = -\int_{\mathbb{R}^d} f(x) \log f(x)\, \mu(dx). \tag{34}$$

Очевидно, что энтропия является функцией, зависящей от вероятностного распределения, поскольку формулы выше основаны соответственно на законах распределения X и Y. Отметим, что дискретизация случайной величины, имеющей плотность (относительно меры Лебега) и вычисление соответствующих энтропий Шеннона не приводит в пределе к значению дифференциальной энтропии Шеннона при стремлении к нулю величины разбиения сетки (см., например, теорему 8.3.1 из [34] и [70]). В общем случае, когда имеется фиксированная мера σ на измеримом пространстве $(S, \mathcal{B})$ можно определить энтропию меры ν, определённой на том же пространстве и абсолютно непрерывной относительно σ, следующим образом:

$$H_\sigma(\nu) := -\int_S \log \left(\frac{d\nu}{d\sigma} \right) d\nu, \tag{35}$$

где $\frac{d\nu}{d\sigma}$ — производная Радона-Никодима. Интеграл должен быть определен, но может принимать бесконечные значения.

Если Y имеет распределение ν на пространстве $(M, 2^M)$, то (33) является частным случаем (35), где $S = M$, $\mathcal{B} = 2^M$ и σ является считающей мерой на M. Если X имеет распределение ν на $(S, \mathcal{B})$, то (35) принимает вид (34), где $S = \mathbb{R}^d$, $\mathcal{B} = \mathcal{B}(\mathbb{R}^d)$ и $\sigma = \mu$. Определение дивергенции (расхождения) Кульбака-Лейблера (см., например, [34], стр. 19, 251) для двух вероятностных мер тесно связано с (35). С различными вариантами f-расхождений можно ознакомиться в работе [83].

Рассмотрим случайный вектор (X, Y) такой, что $X\colon \Omega \to \mathbb{R}^d$ и $Y\colon \Omega \to M$. Здесь M — произвольное конечное множество, $d \in \mathbb{N}$. Без ограничения общности можно считать, что $\mathsf{P}(Y = y) > 0$ для всех $y \in M$. Пусть существует некоторая измеримая функция $f_{X,Y} : \mathbb{R}^d \times M \to \mathbb{R}_+$ такая, что для всех $B \in \mathcal{B}(\mathbb{R}^d)$ и $y \in M$

$$\mathsf{P}(X \in B, Y = y) = \int_B f_{X,Y}(x, y)\, \mu(dx). \tag{36}$$

Иными словами, $f_{X,Y}$ является плотностью случайного вектора (X, Y) относительно меры $\sigma := \mu \otimes \lambda$ на $\mathcal{B}(\mathbb{R}^d) \otimes 2^M$. Для $x \in \mathbb{R}^d$ и $y \in M$ определим следующие функции:

$$f_X(x) := \sum_{y \in M} f_{X,Y}(x, y),$$

$$f_{X|Y}(x|y) := \frac{f_{X,Y}(x, y)}{\mathsf{P}(Y = y)},$$

$$f_{Y|X}(y|x) := \begin{cases} \frac{f_{X,Y}(x,y)}{f_X(x)}, & \text{если } f_X(x) > 0, \\ 0, & \text{если } f_X(x) = 0. \end{cases} \tag{37}$$

Отметим, что f_X является плотностью X, $f_{X|Y}$ — условной плотностью X при условии Y, а $f_{Y|X}$ — условным распределением Y при условии X. Для простоты будем писать dx вместо $\mu(dx)$ и определим $f(x, y) := f_{X,Y}(x, y)$, $f(y|x) := f_{Y|X}(y|x)$.

Согласно (35) (см. также [73]) энтропия вектора (X, Y) в рамках модели (36) определяется формулой

$$H(X, Y) := -\mathsf{E} \log f(X, Y) = -\sum_{y \in M} \int_{\mathbb{R}^d} f(x, y) \log f(x, y)\, dx.$$

Определим условную энтропию Y при условии X:

$$H(Y|X) := -\mathsf{E} \log f(Y|X) = -\sum_{y \in M} \int_{\mathbb{R}^d} f(x, y) \log f(y|x)\, dx. \tag{38}$$

Нетрудно убедиться, что $H(Y|X)$ всегда неотрицательна и конечна. Действительно,

$$0 \le -\mathsf{E}\log f(Y|X) = \sum_{y\in M}\int_{\mathbb{R}^d} f(x,y)\log\frac{1}{f(y|x)}\,dx$$
$$\le \sum_{y\in M}\int_{\mathbb{R}^d}\frac{f(x,y)}{f(y|x)}\,dx = \sum_{y\in M}\int_{\mathbb{R}^d} f_X(x)\,dx = \sharp M < \infty.$$

Здесь мы воспользовались неравенством $\log z \le z$ при $z > 0$ и тем фактом, что $f(y|x) \in [0,1]$ для почти всех $x \in \mathbb{R}^d$.

Взаимная информация X и Y определяется следующим образом:

$$I(X,Y) := D_{KL}(P_{X,Y}||P_X \otimes P_Y),$$

т.е. это расстояние Кульбака-Лейблера D_{KL} между распределением вектора (X,Y) и произведением распределений X и Y. Также имеет место следующее представление:

$$I(X,Y) = H(X) + H(Y) - H(X,Y)$$
$$= H(X) - H(X|Y) = H(Y) - H(Y|X), \quad (39)$$

если все приведённые выражения определены (т.е. мы исключаем случай, когда мы складываем бесконечности с разными знаками). Отметим, что $H(Y)$ и $H(Y|X)$ конечны для рассматриваемой модели, а поэтому $I(X,Y)$ также конечна. Хорошо известно, что $I(X,Y) \ge 0$. Кроме того, $I(X,Y) = 0$ тогда и только тогда, когда X и Y независимы. Таким образом, взаимная информация и ее оценки могут быть использованы для проверки гипотез о независимости случайных величин. Отметим, что существуют и другие подходы, см. например, [33]. Указанное свойство используется также в информационном подходе для выделения значимых факторов, влияющих на значение переменной отклика. Существует обобщение (39) для случая n случайных векторов (см., например, [38]). Оцениванию (безусловной) энтропии посвящено большое число работ. Упомянем недавнюю работу [27], в которой исследуются оценки дифференциальной энтропии Шеннона и в которой можно найти ссылки на другие работы по данной теме.

Рассматриваемая схема (36) включает в себя часто используемую модель логистической регрессии (см., например, [67]). А именно, пусть $M = \{1,2\}$ и

$$\mathsf{P}(Y = 1|X = x) = \frac{1}{1+\exp\{-(w,x)-b\}}, \quad x \in \mathbb{R}^d, \quad w \in \mathbb{R}^d, \quad b \in \mathbb{R}, \quad (40)$$

где $(\cdot,\cdot)$ — скалярное произведение в пространстве $\mathbb{R}^d$ и

$$\mathsf{P}(Y=2|X=x)=1-\mathsf{P}(Y=1|X=x).$$

Обычно модель логистической регрессии формулируется таким образом, что Y принимает значения 0 и 1, однако для удобства в данной работе мы будем использовать значения 1 и 2. Пусть f_X — вероятностная плотность вектора X. Тогда

$$f_{X,Y}(x,1)=\mathsf{P}(Y=1|X=x)f_X(x),$$
$$f_{X,Y}(x,2)=f_X(x)-f_{X,Y}(x,1).$$

Существуют также обобщения логистической регрессии для случая, когда переменная отклика Y принимает более двух значений.

В разделе 2.2 вводится оценка условной энтропии $H(Y|X)$ и формулируются два основных результата. Предлагаемая оценка строится на основе определённых статистик k-ближайших соседей (где $k=k_n$ зависит от числа n независимых одинаково распределённых пар (X^1,Y^1), …, (X^n,Y^n)) и случайного числа наблюдений, входящих в окрестность заданного случайного радиуса. Асимптотическая несмещённость и L_2-состоятельность предлагаемой оценки доказываются при весьма широких условиях в разделах 2.4.1 и 2.4.2, соответственно, а в разделе 2.3 приводятся некоторые вспомогательные результаты. Их доказательства, а также доказательства следствий вынесены в разделы 2.4.4—2.4.7. Исследование оценок взаимной информации, основанных на предлагаемой оценке условной энтропии, приводится в главе 3. Стоит подчеркнуть, что предлагаемая оценка условной энтропии не предполагает существования топологической структуры на множестве значений переменной отклика M (т.е. мы не используем нетривиальные расстояния между Y^i и Y^j, $i,j=1,\dots,n$).

2.2 Основной результат

Пусть $Z^1,Z^2,\dots$ — последовательность таких независимых одинаково распределенных случайных векторов $Z^i=(X^i,Y^i)$, $i\in\mathbb{N}$, что распределение Z^1 совпадает с распределением вектора (X,Y), задаваемого моделью (36). Введем оценку $H(Y|X)$, построенную по выборке $Z^1,\dots,Z^n$:

$$\widehat{H}_{n,k}=\frac{1}{n}\sum_{i=1}^{n}\widehat{H}_{n,k,i}. \tag{41}$$

Здесь $n\in\mathbb{N}$, $n>1$, $k=k(n)\in\{1,\dots,n-1\}$,

$$\widehat{H}_{n,k,i}=-\log(\xi_{n,k,i}(Z^1,\dots,Z^n)+1)+\log k, \tag{42}$$

$$\xi_{n,k,i}(Z^1,\dots,Z^n)$$
$$:= \sharp\{j\in\{1,\dots,n\}\setminus\{i\}\colon Y^j=Y^i, \|X^i-X^j\|\le\rho_{n,k,i}(X^1,\dots,X^n)\}, \quad (43)$$

где $\sharp$ обозначает мощность конечного множества, $\|\cdot\|$ — евклидова норма в пространстве $\mathbb{R}^d$ и

$$\rho_{n,k,i}(X^1,\dots,X^n) := \|X^i - X^{i,(k)}\|, \quad (44)$$

$X^{i,(k)}$ является k-ым ближайшим соседом X^i в выборке $\{X^1,\dots,X^n\}\backslash\{X^i\}$, а $\rho_{n,k,i}(X^1,\dots,X^n)$ — евклидово расстояние от X^i до его k-го ближайшего соседа. Очевидно, что случайная величина $\xi_{n,k,i}(Z^1,\dots,Z^n)$ принимает значения $0,1,\dots,k$. С вероятностью 1 точки $X^1,\dots,X^n$ попарно не совпадают, поскольку вектор X имеет плотность относительно меры Лебега на $\mathbb{R}^d$.

Таким образом, в отличие от известной оценки Козаченко-Леоненко [5] для дифференциальной энтропии случайного вектора важную роль помимо расстояний до k-го ближайшего соседа X^i в выборке $X^1,\dots,X^n$ (исключая точку X^i) играют случайные величины $\xi_{n,k,i}$, $i=1,\dots,n$, т.е. мы строим множество $J\subset\{1,\dots,n\}$, состоящее из всех индексов $j\in\{1,\dots,n\}\backslash\{i\}$ таких, что X^j принадлежит шару $B(X^i,\rho_{n,k,i})$ со случайным центром и случайным радиусом. Затем из множества $\{(X^j,Y^j), j\in J\}$ выбираются такие $\{(X^j,Y^j), j\in I_i\}$, что $I_i := \{j\in J : Y^j = Y^i\}$. Возникает набор случайных величин $\{(X^j,Y^j), j\in I_i\}$, где I_i также случайный набор. Мощность множества I_i, т.е. $\sharp I_i$, равна значению случайной величины $\xi_{n,k,i}$, $i=1,\dots,n$.

Определение 2.1 *Функция $g\colon\mathbb{R}^d\to\mathbb{R}$ называется локально стягиваемой в точке $x\in\mathbb{R}^d$, если существуют строго положительные $R_0(x)$ и $C_0(x)$ такие, что*

$$\left|g(x) - \frac{1}{|B(x,R)|}\int_{B(x,R)} g(v)\,dv\right| \le C_0(x)R \ \text{ для } \ R\in(0,R_0(x)), \quad (45)$$

где $|B(x,R)|$ — объём шара $B(x,R) := \{v\in\mathbb{R}^d : \|v-x\|\le R\}$, т.е. $|B(x,R)| = \mu(B(x,R))$. Функция g называется C_0-стягиваемой, если она локально стягиваема для μ-почти всех точек $x\in\mathbb{R}^d$ и, кроме того, для всех x выполняется неравенство $C_0(x)\le C_0$ и $R_0(x)\ge R_0$, где C_0 и R_0 — строго положительные константы.

Замечание 2.2 *Если функция $g:\mathbb{R}^d\to\mathbb{R}$ удовлетворяет условию Липшица в каждой точке $x\in\mathbb{R}^d$ с параметром $C(x)$, т.е.*

$$|g(v)-g(x)|\le C(x)\|v-x\|$$

для всех $v \in \mathbb{R}^d$, *то* (45) *выполняется для всех* $R_0(x) > 0$. *Нетрудно убедиться, что если* $g(x)$, $x \in \mathbb{R}^d$, *является плотностью невырожденного гауссовского случайного вектора* $N(a, \Sigma)$ *в* $\mathbb{R}^d$, *то эта функция является* C_0*-стягиваемой. с константой*

$$C_0 = \max_{u \in R^d} \|\nabla p_{a,\Sigma}(u)\| < \infty.$$

Для неотрицательных последовательностей $\{a_n\}_{n=1}^{\infty}$ и $\{b_n\}_{n=1}^{\infty}$ пишут $a_n \propto b_n$, если

$$c_1 b_n \le a_n \le c_2 b_n,\ n \in \mathbb{N},$$

где c_1 и c_2 – некоторые положительные константы ($c_1 < c_2$).

Теорема 2.3 *Пусть в рамках модели* (36) *выполняются следующие условия. Для каждого фиксированного* $y \in M$ *и* μ*-почти всех* $x \in \mathbb{R}^d$ *функция* $f(x, y)$, *т.е.* $f(\cdot, y)$, *строго положительна и* C_0*-стягиваема,*

$$k = k_n \propto n^{\alpha} \tag{46}$$

для некоторого $\alpha \in (0, 1)$, *и для некоторого* $\varepsilon > 0$

$$\mathsf{E}|\log f_X(X)|^{1+\varepsilon} < \infty, \tag{47}$$

где $f_X(\cdot)$ — *плотность* X. *Тогда*

$$\mathsf{E}\widehat{H}_{n,k} \to H(Y|X),\ \ n \to \infty, \tag{48}$$

т.е. $\widehat{H}_{n,k}$ *является асимптотически несмещённой оценкой* $H(Y|X)$.

Теорема 2.4 *Пусть условие* (47) *теоремы 2.3 заменяется следующим. Для некоторого* $\varepsilon > 0$

$$\mathsf{E}|\log f_X(X)|^{2+\varepsilon} < \infty. \tag{49}$$

Тогда

$$\mathsf{E}(\widehat{H}_{n,k} - H(Y|X))^2 \to 0,\ \ n \to \infty,$$

т.е. $\widehat{H}_{n,k}$ *является* L_2*-состоятельной оценкой* $H(Y|X)$.

Следствие 2.5 *Пусть в рамках модели* (36) *для всех* $y \in M$ *функция* $f(\cdot, y)$ *является плотностью невырожденного гауссовского случайного вектора со значениями в* $\mathbb{R}^d$ (*со средним и матрицей ковариации, зависящими от* y). *Тогда* $\widehat{H}_{n,k}$, *где* $k(n)$ *удовлетворяет условию* (46), *является асимптотически несмещённой и* L_2*-состоятельной оценкой* $H(Y|X)$ *при* $n \to \infty$.

2.3 Вспомогательные результаты

В этом разделе мы также рассматриваем независимые одинаково распределённые векторы $Z^1, Z^2, \ldots$, имеющие такое же распределение, что и вектор (X, Y) в рамках модели (36). Используя обозначения, введённые в разделе 2.2, для $x \in \mathbb{R}^d$, $y \in M$, $k \in \{1, \ldots, n-1\}$ и $n > 2$, определим

$$\rho_{n,k,1}(x) := \rho_{n,k,1}(x, X^2, \ldots, X^n), \;\; \xi_{n,k,1}(x, y) := \xi_{n,k,1}((x, y), Z^2, \ldots, Z^n).$$

Эти случайные переменные зависят, соответственно, от $x, X^2, \ldots, X^n$ и (x, y), $Z^2, \ldots, Z^n$. Для простоты далее будем опускать случайные аргументы этих функций.

Теперь сформулируем два вспомогательных результата об условных распределениях случайных величин, играющих ключевую роль в изучении асимптотического поведения оценок условной энтропии. Примечательно, что упоминаемые условные распределения являются смесями некоторых биномиальных распределений с явно определёнными весовыми коэффициентами. Доказательства этих результатов вынесены в разделы 2.4.4—2.4.7. Распределение случайной переменной (или вектора) η будем обозначать P_η.

Лемма 2.6 *Для некоторых $y \in M$, $x \in \mathbb{R}^d$, $r = 0, 1, \ldots, k$ и $\mathsf{P}_{\rho_{n,k,1}(x)}$-почти всех $t \in (0, \infty)$, выполняется следующее соотношение:*

$$\begin{aligned}
&\mathsf{P}(\xi_{n,k,1}(x, y) = r | \rho_{n,k,1}(x) = t) \\
&= \binom{k-1}{r} \mathsf{P}(Y = y | \|X - x\| \le t)^r (1 - \mathsf{P}(Y = y | \|X - x\| \le t)^{k-1-r} \alpha(x, y, t) \\
&+ \binom{k-1}{r-1} \mathsf{P}(Y = y | \|X - x\| \le t)^{r-1} (1 - \mathsf{P}(Y = y | \|X - x\| \le t)^{k-r} (1 - \alpha(x, y, t)),
\end{aligned}$$

где $\alpha(x, y, t) = \mathsf{P}(Y \ne y | \|X - x\| = t)$ и $\binom{N}{m} := 0$ для $m < 0$ и $m > N$ ($N \in \mathbb{N}$, $m \in \mathbb{Z}$).

Замечание 2.7 *Как обычно, для случайных векторов $\eta : \Omega \to \mathbb{R}^q$, $\zeta : \Omega \to \mathbb{R}^s$, и для $B \in \mathcal{B}(\mathbb{R}^q)$, $x \in \mathbb{R}^s$, запись $\mathsf{P}(\eta \in B | \zeta = x) = \varphi(x)$ значит, что определена такая борелевская функция $\varphi(x)$, $x \in \mathbb{R}^s$, что $\mathsf{P}(\eta \in B | \zeta) = \varphi(\zeta)$. Функция φ однозначно определена P_ζ-почти наверное, см., например, [7], ч.1, гл.II, раздел 5.*

Положим $z_j = (x_j, y_j)$, где $x_j \in \mathbb{R}^d$, $y_j \in M$, $j \in \{1, 2\}$. Для $n > 2$ определим случайные величины

$$\rho_{n,k,j}(x_1, x_2) := \rho_{n,k,j}(x_1, x_2, X^3, \ldots, X^n), \tag{50}$$

$$\xi_{n,k,j}(z_1, z_2) := \xi_{n,k,j}(z_1, z_2, Z^3, \dots, Z^n). \tag{51}$$

Будем опускать в записи случайные аргументы этих функций.

Лемма 2.8 *Пусть $z_j = (x_j, y_j)$, где $x_j \in \mathbb{R}^d$, $y_j \in M$, $j = 1, 2$, $x_1 \neq x_2$. Зададим случайный вектор $\zeta := (\rho_{n,k,1}(x_1, x_2), \rho_{n,k,2}(x_1, x_2))$. Тогда для всех $n > 2$, $k \in \{0, 1, \dots, [(n-2)/2]\}$, где $[\cdot]$ обозначает целую часть числа, всех $r_1, r_2 \in \{0, \dots, k\}$ и P_ζ-почти всех (t_1, t_2) таких, что $t_1 > 0$, $t_2 > 0$ и $t_1, t_2 < |x_1 - x_2|/2$, выполняется следующее соотношение:*

$$\mathsf{P}(\xi_{n,k,1}(z_1, z_2) = r_1, \xi_{n,k,2}(z_1, z_2) = r_2 | \zeta = (t_1, t_2)) = \prod_{j=1}^{2} \mathsf{P}(\xi_{n,k,j}(z_1, z_2) = r_j | \zeta = (t_1, t_2)). \tag{52}$$

Кроме того,

$$\begin{aligned} &\mathsf{P}\left(\xi_{n,k,j}(z_1, z_2) = r_j | \zeta = (t_1, t_2)\right) \\ &\quad = \binom{k-1}{r_j} p_j^{r_j} (1 - p_j)^{k-1-r_j} \alpha(x_j, y_j, t_j) \\ &\qquad + \binom{k-1}{r_j - 1} p_j^{r_j - 1} (1 - p_j)^{k - r_j} (1 - \alpha(x_j, y_j, t_j)), \end{aligned} \tag{53}$$

где $p_j = \mathsf{P}(Y = y_j | X \in B(x_j, t_j))$, $j = 1, 2$, и $\alpha(x, y, t)$ те же самые, что и в лемме 2.6.

Также для доказательства основных теорем понадобятся следующие простые утверждения.

Лемма 2.9 *Пусть W — случайная переменная с конечным математическим ожиданием $\mathsf{E}W$ и V — случайная величина со значениями в $\mathbb{R}^m$, причем $\mathsf{P}(V \in B) > 0$, где $B \in \mathcal{B}(\mathbb{R}^m)$. Тогда*

$$\mathsf{E}(W | V \in B) = \int_B \mathsf{E}(W | V = y)\, \widetilde{\mathsf{P}}_{V,B}(dy),$$

где $\mathsf{E}(W | V \in B) := \frac{1}{\mathsf{P}(V \in B)} \mathsf{E}(W \mathbb{I}\{V \in B\})$ и $\widetilde{\mathsf{P}}_{V,B}(A) := \mathsf{P}(V \in A | V \in B)$, $A \in \mathcal{B}(\mathbb{R}^m)$.

Лемма 2.10 *Пусть ξ, η — некоторые случайные переменные (величины, векторы) и $\mathsf{E}|\xi| < \infty$. Предположим, что случайная переменная ζ принимает значения в конечном или счётном множестве S. Тогда для P_η-почти всех t*

$$\mathsf{E}(\xi | \eta = t) = \sum_{r \in S} \mathsf{E}(\xi | \zeta = r, \eta = t) \mathsf{P}(\zeta = r | \eta = t).$$

2.4 Доказательства новых результатов

2.4.1 Доказательство теоремы 2.3

Наблюдения $Z^1, Z^2, \dots$ одинаково распределены, поэтому

$$\mathsf{E}\widehat{H}_{n,k} = -\mathsf{E}\log\left(\frac{\xi_{n,k,1}+1}{k}\right),$$

а значит, достаточно показать, что

$$-\mathsf{E}\log\left(\frac{\xi_{n,k,1}+1}{k}\right) \to H(Y|X), \quad n\to\infty. \tag{54}$$

Поскольку Z^1 и $Z^2, \dots, Z^n$ независимы, получаем, что

$$\mathsf{E}\left(\log\left(\frac{\xi_{n,k,1}+1}{k}\right)\middle| X^1 = x, Y^1 = y\right) = \mathsf{E}\log\left(\frac{\xi_{n,k,1}(x,y)+1}{k}\right) := h_{n,k}(x,y).$$

Следовательно,

$$\mathsf{E}\log\left(\frac{\xi_{n,k,1}+1}{k}\right) = \sum_{y\in M}\int_{\mathbb{R}^d} h_{n,k}(x,y)f(x,y)dx. \tag{55}$$

Для того, чтобы доказать (54), рассмотрим правую часть выражения (55). Для каждого $n\in\mathbb{N}$ построим разбиение всего пространства $\mathbb{R}^d$ на подмножества $B_{1,n}$ и $B_{2,n}$, зависящее от вспомогательных параметров $\theta, \nu > 0$. Затем мы рассмотрим интегралы по этим множествам и покажем, что интеграл по $B_{1,n}$ определяет для всех $y\in M$ асимптотическое поведение (при $n\to\infty$) интегралов, фигурирующих в (55). Одновременно производится две процедуры усреднения. Первая задается формулой (55). Вторая основана на телескопическом свойстве математического ожидания и позволяет в формуле (55) вместо $h_{n,k}(x,y)$ написать

$$\mathsf{E}\left(\mathsf{E}\left(\log\left(\frac{\xi_{n,k,1}(x,y)+1}{k}\right)\middle| n^\beta\rho_{n,k,1}(x)\right)\right),$$

где $\beta > 0$ является параметром. Далее мы исследуем комбинации интегралов с помощью леммы 2.6, которая дает явную формулу для условного распределения $\xi_{n,k,1}(x,y)$ при условии $\rho_{n,k,1}(x)$. Здесь для $\delta > 0$ мы также используем разбиение множества значений случайной величины $n^\beta\rho_{n,k,1}(x)$ на множества $(0,\delta]$, (δ,∞) и оцениваем вклад каждого интеграла в сумму.

Подходящий выбор параметров θ, ν, δ и β приводит к желаемому результату. Для удобства изложения мы разбиваем доказательство на три основных шага.

Шаг 1. Введем параметры $\theta, \nu > 0$. В дальнейшем значения параметров будут подобраны особым образом. Покажем, что достаточно рассматривать только такие $x \in \mathbb{R}^d$, что $f_X(x) > n^{-\theta}$ и для всех $y \in M$, $f(y|x) > n^{-\nu}$.

В силу (47) для $n > 2$ и $y \in M$ мы приходим к соотношениям

$$\int\limits_{\{x: f_X(x) \le n^{-\theta}\}} f_X(x)\,dx \le \int\limits_{\{x: f_X(x) \le n^{-\theta}\}} \frac{|\log f_X(x)|^{1+\varepsilon}}{|\log n^{-\theta}|^{1+\varepsilon}} f_X(x)\,dx \le \frac{1}{(\theta \log n)^{1+\varepsilon}} \mathsf{E}|\log f_X(X)|^{1+\varepsilon}, \quad (56)$$

$$\int\limits_{\{x: f(y|x) \le n^{-\nu}\}} f(x,y)\,dx = \int\limits_{\{x: f(y|x) \le n^{-\nu}\}} f(y|x) f_X(x)\,dx \le n^{-\nu} \int\limits_{\mathbb{R}^d} f_X(x)\,dx = n^{-\nu}. \quad (57)$$

Для $n \in \mathbb{N}$ положим $\theta_n := n^{-\theta}$, $\nu_n := n^{-\nu}$ и рассмотрим множества

$$B_{1,n} := \bigcap_{y \in M} \{x \in \mathbb{R}^d \colon f(y|x) > \nu_n\} \cap \{x \in \mathbb{R}^d \colon f_X(x) > \theta_n\},$$
$$B_{2,n} := \mathbb{R}^d \setminus B_{1,n}. \quad (58)$$

Для математического ожидания верно представление

$$\mathsf{E} \log \left(\frac{\xi_{n,k,1} + 1}{k} \right) = I_1(n,k) + I_2(n,k),$$
$$I_j(n,k) := \sum_{y \in M} \int\limits_{B_{j,n}} h_{n,k}(x,y) f(x,y) dx, \quad j = 1, 2.$$

Для $k > 1$, всех $x \in \mathbb{R}^d$ и $y \in M$ $\xi_{n,k,1}(x,y)$ принимает значения $0, 1, \ldots, k$, поэтому выполняется неравенство $|h_{n,k}(x,y)| \le \log k$. Таким образом,

$$|I_2(n,k)| \le \log k \left(\int\limits_{\{x: f_X(x) \le \theta_n\}} f_X(x)\,dx + \sum_{y \in M} \int\limits_{\{x: f(y|x) \le \nu_n\}} f(x,y)\,dx \right). \quad (59)$$

Согласно (56) и (57) мы получаем $I_2(n,k) \to 0$, $n \to \infty$, поскольку $k \propto n^\alpha$.

Зафиксируем параметр $\beta > 0$ и заметим, что

$$\mathsf{E}\left(\log\left(\frac{\xi_{n,k,1}(x,y)+1}{k}\right)\middle| n^\beta \rho_{n,k,1}(x)\right) = \sum_{r=0}^{k} \log\left(\frac{r+1}{k}\right) \mathsf{P}(\xi_{n,k,1}(x,y) = r | n^\beta \rho_{n,k,1}(x)).$$

Следовательно,

$$h_{n,k}(x,y) = \sum_{r=0}^{k} \int_{(0,\infty)} \log\left(\frac{r+1}{k}\right) \mathsf{P}(\xi_{n,k,1}(x,y) = r | n^\beta \rho_{n,k,1}(x) = u) f^{(k)}_{n,x,\beta}(u)du,$$

где $f^{(k)}_{n,x,\beta}(\cdot)$ — плотность положительной случайной величины $n^\beta \rho_{n,k,1}(x)$, а плотность случайной величины $\rho_{n,k,1}(x)$ вводится в доказательстве леммы 2.6 (см. далее формулу (103)), здесь du обозначает интегрирование по мере Лебега на $\mathbb{R}$.

Теперь зафиксируем произвольное $\delta > 0$ и напишем $I_1(n,k) = S_1(n,k) + S_2(n,k)$, где для $V_1 = (\delta, \infty)$, $V_2 = (0, \delta]$ и $j = 1, 2$,

$$S_j(n,k) := \sum_{y \in M} \int_{B_{1,n}} \sum_{r=0}^{k} \int_{V_j} \log\left(\frac{r+1}{k}\right) \mathsf{P}(\xi_{n,k,1}(x,y) = r | \rho_{n,k,1}(x) = u n^{-\beta}) \times f^{(k)}_{n,x,\beta}(u)du f(x,y)dx.$$

Дальнейшее доказательство разбивается на два основных шага.

Шаг 2. Покажем, что $S_1(n,k) \to 0$ при $n \to \infty$, если параметры β и δ выбраны особым образом. Найдем верхнюю границу для $|S_1(n,k)|$. Для всех $n \in \mathbb{N}$, $1 \le k \le n$ и $x, y \in \mathbb{R}^d$ случайная величина $\xi_{n,k,1}(x,y)$ принимает значения $0, \dots, k$. Следовательно, для $k > 1$ верно неравенство

$$\sum_{r=0}^{k} \left|\log\left(\frac{r+1}{k}\right)\right| \mathsf{P}(\xi_{n,k,1}(x,y) = r | \rho_{n,k,1}(x) = u n^{-\beta}) \le \log k.$$

Таким образом,

$$|S_1(n,k)| \le \log k \sum_{y=1}^{m} \int\limits_{B_{1,n}} \int_{(\delta,\infty)} f^{(k)}_{n,x,\beta}(u) f(x,y)\, du\, dx$$
$$= \log k \int\limits_{B_{1,n}} \mathsf{P}\left(\rho_{n,k,1}(x) > \delta n^{-\beta}\right) f_X(x)\, dx$$
$$= \log k \int\limits_{B_{1,n}} \mathsf{P}(\eta_n(\beta,\delta,x) \le k-1) f_X(x)\, dx. \quad (60)$$

Здесь $\eta_n(\beta,u,x) \sim Bin(n-1, p_n(\beta,u,x))$, т.е. $\eta_n(\beta,u,x)$ имеет биномиальное распределение с параметрами $n-1$ и $p_n(\beta,u,x)$, где

$$p_n(\beta,u,x) := \mathsf{P}(X \in U_n(\beta,u,x)) = \int_{U_n(\beta,u,x)} f_X(v)\, dv,$$
$$U_n(\beta,u,x) = \{v \in \mathbb{R}^d \colon \|v-x\| \le u n^{-\beta}\}, \quad u > 0, \quad x \in \mathbb{R}^d.$$

Действительно, $f^{(k)}_{n,x,\beta}(\cdot)$ является плотностью случайной величины $\rho_{n,k}(x) n^{\beta}$, следовательно

$$\int_{(\delta,\infty)} f^{(k)}_{n,x,\beta}(u)\, du = \mathsf{P}(\rho_{n,k,1}(x) > \delta n^{-\beta}). \quad (61)$$

Случайное событие $\{\omega : \rho_{n,k,1}(x) > \delta n^{-\beta}\}$ означает, что в шаре $B(x, \delta n^{-\beta})$ содержится не более $k-1$ точек из $\{X^i\}_{i=2}^n$. Независимость наблюдений обеспечивает выполнение соотношения

$$\mathsf{P}(\rho_{n,k,1}(x) > \delta n^{-\beta}) = \mathsf{P}(\eta_n(\beta,\delta,x) \le k-1).$$

Согласно неравенству для биномиальных сумм [1], для любых $n > 1$, $k = 0, 1, \ldots, n-1$ и всех рассматриваемых значений β, δ и x, имеет место следующая оценка:

$$\mathsf{P}(\eta_n(\beta,\delta,x) \le k-1)$$
$$\le \Phi\left(\operatorname{sgn}\left(\frac{k}{n-1} - p_n(\beta,\delta,x)\right) \sqrt{2(n-1) h\left(\frac{k}{n-1}, p_n(\beta,\delta,x)\right)}\right), \quad (62)$$

где $\Phi(\cdot)$ — функция распределения стандартной нормальной случайной величины,

$$\operatorname{sgn}(t) = \begin{cases} 1, & t > 0, \\ 0, & t = 0, \\ -1, & t < 0, \end{cases}$$

$$h(t,s) = t\log\left(\frac{t}{s}\right) + (1-t)\log\left(\frac{1-t}{1-s}\right), \quad s,t \in (0,1).$$

Для всех $y \in M$, функция $f(\cdot, y)$ является C_0-стягиваемой, следовательно, для μ-почти всех $x \in \mathbb{R}^d$, всех $u \in (0, \delta]$ и всех достаточно больших значений n

$$\left|\frac{\mathsf{P}(X \in U_n(\beta, u, x), Y = y)}{|U_n(\beta, u, x)|} - f(x,y)\right| \le C_0 u n^{-\beta}. \tag{63}$$

Поскольку $f_X(x) = \sum_{y \in M} f(x,y)$, для рассматриваемых x, u и β верно неравенство

$$\left|f_X(x) - \frac{p_n(\beta, u, x)}{|U_n(\beta, u, x)|}\right| \le \sharp M C_0 u n^{-\beta} = C u n^{-\beta}, \tag{64}$$

где $C = \sharp M C_0$. Отсюда следует, что для произвольных $\delta > 0$, $\beta > 0$, μ-почти всех $x \in B_{1,n}$ и любых $n \ge N_0$, где $N_0 = N_0(\delta, \beta)$, выполняется

$$p_n(\beta, \delta, x) \ge V_d \delta^d n^{-d\beta}(f_X(x) - C\delta n^{-\beta})) \ge V_d \delta^d n^{-d\beta}(\theta_n - C\delta n^{-\beta}). \tag{65}$$

Напомним, что $V_d = |B(0,1)|$, $B(0,1) \subset \mathbb{R}^d$. Теперь можно получить верхнюю оценку для аргумента функции sgn в формуле (62). В силу (65) получаем

$$\frac{k}{n-1} - p_n(\beta, \delta, x) \le \frac{k}{n-1} - V_d \delta^d n^{-d\beta}(\theta_n - C\delta n^{-\beta}).$$

Положим $\beta > \theta$. Тогда $n^{-\beta} = o(\theta_n)$ при $n \to \infty$. Согласно (46), верно соотношение $\frac{k}{n-1} \propto n^{\alpha-1}$, $n \to \infty$. Пусть параметры β и θ таковы, что

$$\alpha - 1 < -d\beta - \theta. \tag{66}$$

Для выполнения (66) достаточно того, чтобы было выполнено неравенство

$$(d+1)\theta < 1 - \alpha,$$

поскольку параметр $\beta > \theta$ можно выбрать сколь угодно близким к θ. Тогда $n^{\alpha-1} = o(n^{-d\beta-\theta})$, $n \to \infty$. Таким образом, существует $N \in \mathbb{N}$ (где $N = N(C, d, \alpha, \delta, \beta, \theta)$) такое, что $n > N$. Тогда $\frac{k}{n-1} - p_n(\beta, \delta, x) < 0$ для μ-почти всех $x \in B_{1,n}$, что влечет за собой

$$\operatorname{sgn}\left(\frac{k}{n-1} - p_n(\beta, \delta, x)\right) = -1. \tag{67}$$

Для $s, t \in (0, 1)$ введем функции

$$L_{1,n}(t, s) = 2(n-1)(1-t)\log\left(\frac{1-t}{1-s}\right), \quad L_{2,n}(t, s) = 2(n-1)t\log\left(\frac{t}{s}\right).$$

Тогда $2(n-1)h(t, s) = L_{1,n}(t, s) + L_{2,n}(t, s)$. Теперь для $t = \frac{k}{n-1}$ и $s = p_n(\beta, \delta, x)$ исследуем асимптотическое поведение $L_{1,n}(t, s)$ и $L_{2,n}(t, s)$ (для μ-почти всех $x \in B_{1,n}$) при $n \to \infty$. Применяя неравенство (65) для $n \geq N_0$, мы получаем оценку

$$\begin{aligned} L_{1,n}(t, s) &= 2(n-1-k)\left(\log\left(1 - \frac{k}{n-1}\right) - \log(1 - p_n(\beta, \delta, x))\right) \\ &\geq 2(n-1-k)\left(\log\left(1 - \frac{k}{n-1}\right) - \log(1 - V_d\delta^d n^{-d\beta}(\theta_n - C\delta n^{-\beta}))\right) \\ &:= \mathcal{L}_{1,n}(k). \end{aligned} \tag{68}$$

Очевидно, что $\mathcal{L}_{1,n}(k)$ зависит не только от n и k, но и от набора параметров, фигурирующих в (68). Отметим, что $\log(1 + z) = z + o(z)$ при $z \to 0$. Следовательно, в силу (66) и поскольку $\beta > \theta$, получаем

$$\mathcal{L}_{1,n}(k) \propto n^{1-d\beta-\theta} \text{ при } n \to \infty.$$

Для тех же t, s и $x \in B_{1,n}$, учитывая неравенство $0 < p_n(\beta, \delta, x) \leq 1$, имеем

$$L_{2,n}(t, s) = 2k\left(\log\left(\frac{k}{n-1}\right) - \log p_n(\beta, \delta, x)\right) \geq 2k\log\left(\frac{k}{n-1}\right) := \mathcal{L}_{2,n}(k).$$

Следовательно, $\mathcal{L}_{2,n}(k) \propto n^{\alpha}\log n$, $n \to \infty$. Таким образом, согласно (66) заключаем, что для всех достаточно больших n и μ-почти всех $x \in B_{1,n}$,

$$2(n-1)H\left(\frac{k}{n-1}, p_n(\beta, \delta, x)\right) \geq \mathcal{L}_{1,n}(k) + \mathcal{L}_{2,n}(k) := R_n \propto n^{1-d\beta-\theta}, \tag{69}$$

где $k = k_n \propto n^\alpha$. Здесь R_n зависит не только от n, но еще и от α, β, δ, θ и d. Таким образом, для всех достаточно больших n из (69) следует оценка

$$\Phi\left(\operatorname{sgn}\left(\frac{k}{n-1} - p_n(\beta,\delta,x)\right)\sqrt{2(n-1)h\left(\frac{k}{n-1}, p_n(\beta,\delta,x)\right)}\right) \le \Phi(-\sqrt{R_n}).$$

Поскольку $\Phi(-z) \le \frac{1}{\sqrt{2\pi}z}e^{-\frac{z^2}{2}}$ для $z > 1$, учитывая (60), (62) и (69) получаем для $k = k_n \propto n^\alpha$, $x \in B_{1,n}$ и всех достаточно больших n

$$\int_{(\delta,\infty)} f^{(k)}_{n,x,\beta}(u)\,du \le \frac{1}{\sqrt{2\pi R_n}}e^{-R_n/2}, \tag{70}$$

$$|S_1(n,k)| \le \frac{\log k_n}{\sqrt{2\pi R_n}}e^{-R_n/2} \to 0, \quad n \to \infty.$$

Доказательство шага 2 окончено.

Шаг 3. Покажем, что $S_2(n,k) \to -H(Y|X)$, $n \to \infty$. Для $\mathsf{P}_{n^\beta\rho_{n,k,1}(x)}$-почти всех u в силу леммы 2.6

$$\begin{aligned}\mathsf{E}\left(\log(\xi_{n,k,1}(x,y)+1)|\rho_{n,k,1}(x) = un^{-\beta}\right)&\\ = \mathsf{P}(Y \ne y|\|X-x\| = un^{-\beta})\mathsf{E}\log(\mu_n+1)&\\ + \mathsf{P}(Y = y|\|X-x\| = un^{-\beta})\mathsf{E}\log(\mu_n+2),&\end{aligned} \tag{71}$$

где случайная величина μ_n не зависит от (X,Y) и

$$\mu_n = \mu_n(k,\beta,u,x,y) \sim Bin(k-1, P_n(\beta,u,x,y)), \tag{72}$$

$$P_n(\beta,u,x,y) := \mathsf{P}(Y = y|X \in U_n(\beta,u,x)). \tag{73}$$

Отметим, что $\mathsf{P}(X \in B(x,t)) > 0$ для всех $x \in \mathbb{R}^d$ и любых $t > 0$, поскольку $f_X(\cdot)$ строго положительна μ-почти везде. Сначала для всех $u \in (0,\delta]$, $y \in M$ и μ-почти всех $x \in \mathbb{R}^d$ (таких, что $f_X(x) > 0$) исследуем скорость сходимости $P_n(\beta,u,x,y)$ к $f(y|x)$ при $n \to \infty$:

$$\begin{aligned}|P_n(\beta,u,x,y) - f(y|x)| = \frac{|U_n(\beta,u,x)|}{\mathsf{P}(X \in U_n(\beta,u,x))}&\\ \times\left|\left\{\left(\frac{\mathsf{P}(X \in U_n(\beta,u,x), Y = y)}{|U_n(\beta,u,x)|} - f(x,y)\right)\right.&\\ \left. + f(y|x)\left(f_X(x) - \frac{\mathsf{P}(X \in U_n(\beta,u,x))}{|U_n(\beta,u,x)|}\right)\right\}\right|.&\end{aligned}$$

Для всех $y \in M$ и μ-почти всех $x \in \mathbb{R}^d$, теорема Лебега о дифференцировании мер (см., например, теорему 25.17 [99]) утверждает, что

$$f(y|x) = \lim_{R \to 0+} \frac{\mathsf{P}(Y = y, X \in B(x,R))}{\mathsf{P}(X \in B(x,R))} \le 1.$$

Следовательно, в силу неравенств (63) и (64) для всех достаточно больших n мы получаем неравенство

$$|P_n(\beta,u,x,y) - f(y|x)| \le C_0(\sharp M + 1)\delta n^{-\beta} \frac{|U_n(\beta,u,x)|}{\mathsf{P}(X \in U_n(\beta,u,x))},$$

которое выполняется для $u \in (0,\delta]$, μ-почти всех $x \in \mathbb{R}^d$ и всех $y \in M$. Формула (64) позволяет утверждать, что для μ-почти всех $x \in B_{1,n}$, $u \in (0,\delta]$ и $y \in M$

$$\frac{\mathsf{P}(X \in U_n(\beta,u,x))}{|U_n(\beta,u,x)|} \ge f_X(x) - C_0(\sharp M)un^{-\delta} \ge \theta_n - C_0(\sharp M)\delta n^{-\beta} \ge \frac{1}{2}n^{-\theta},$$

если n достаточно велико ($n \ge N(\sharp M, C_0, \delta, \beta, \theta)$) и $\beta > \theta$. Таким образом, для таких n, u, x и y,

$$|P_n(\beta,u,x,y) - f(y|x)| \le 2C_0\delta(\sharp M + 1)n^{-\beta+\theta}. \tag{74}$$

Отметим, что для $k_n > 1$ и $P_n := P_n(\beta,u,x,y)$

$$\mathsf{E}\log(\mu_n + 1) = \log((k_n - 1)P_n) + \mathsf{E}\log\left(\frac{\mu_n + 1}{(k_n - 1)P_n}\right).$$

Определим $G_n := (0,\delta] \times B_{1,n} \times M$. Покажем, что

$$\sup_{(u,x,y)\in G_n} \left|\mathsf{E}\log\left(\frac{\mu_n + 1}{(k_n - 1)P_n}\right)\right| \to 0, \quad n \to \infty. \tag{75}$$

Согласно неравенству Ляпунова достаточно доказать, что

$$\sup_{(u,x,y)\in G_n} \mathsf{E}\left(\log\left(\frac{\mu_n + 1}{(k_n - 1)P_n}\right)\right)^2 \to 0, \quad n \to \infty. \tag{76}$$

Введем $\eta_n := \frac{\mu_n - (k_n-1)P_n + 1}{(k_n-1)P_n}$. Тогда

$$\begin{aligned}\mathsf{E}\left(\log\left(\frac{\mu_n + 1}{(k_n - 1)P_n}\right)\right)^2 &= \mathsf{E}\left(\log\left(1 + \eta_n\right)\right)^2 \\ &= \mathsf{E}\left((\log(1+\eta_n))^2 \mathbb{I}\left\{|\eta_n| < \frac{1}{2}\right\}\right) + \mathsf{E}\left((\log(1+\eta_n))^2 \mathbb{I}\left\{|\eta_n| \ge \frac{1}{2}\right\}\right) \\ &:= T_1(n) + T_2(n),\end{aligned}$$

где $T_1(n) = T_1(n; k_n, u, x, y, \alpha)$, $T_2(n) = T_2(n; k_n, u, x, y, \alpha)$. Для $k_n > 1$ и $(u, x, y) \in G_n$, $\frac{1}{(k_n-1)P_n} \le \frac{\mu_n+1}{(k_n-1)P_n} \le \frac{2}{P_n}$. Следовательно,

$$\left|\log\left(\frac{\mu_n+1}{(k_n-1)P_n}\right)\right| \le \max\left\{\left|\log\left(\frac{1}{(k_n-1)P_n}\right)\right|, \left|\log\left(\frac{2}{P_n}\right)\right|\right\}.$$

Учитывая соотношение (74) и границу $f(y|x) > n^{-\nu}$ для $(u, x, y) \in G_n$, получаем, что если $0 < \nu < \beta - \theta$, то

$$\frac{1}{2}n^{-\nu} \le P_n \le 1$$

и

$$\frac{1}{4}n^{-\nu}k_n \le (k_n - 1)P_n \le k_n,$$

где n достаточно велико. Следовательно, для всех достаточно больших n

$$\left|\log\left(\frac{\mu_n+1}{(k_n-1)P_n}\right)\right| \le b\log n,$$

где $b = b(\alpha, \nu)$ не зависит от n. Если $0 < \nu < \alpha$, то для всех достаточно больших n

$$T_2(n) \le (b\log n)^2\mathsf{P}\left(|\eta_n| \ge \frac{1}{2}\right) \le (b\log n)^2\mathsf{P}(|\mu_n-(k_n-1)P_n| \ge \frac{1}{4}(k_n-1)P_n).$$

Для случайной переменной $H(m) \sim Bin(m, p)$, $p \in (0, 1)$, $p = p_m$ и $\varepsilon = \varepsilon_m > 0$, неравенство Хёфдинга (см, например, [68], стр. 22) обеспечивает выполнение оценки

$$\mathsf{P}(|H(m) - mp| \ge m\varepsilon) \le 2\exp\{-2m\varepsilon^2\}.$$

Применим это неравенство к $m = (k_n - 1)$, $k_n \propto n^\alpha$, $p = P_n(\beta, u, x, y)$ и $\varepsilon_{k-1} = \frac{P_n}{4}$. Тогда

$$\sup_{(u,x,y)\in G_n} T_2(n) \le 2(b\log n)^2\exp\left\{-\frac{1}{8}(k_n - 1)P_n^2\right\} \to 0, \;\; n \to \infty,$$

когда $\alpha - 2\nu > 0$ (мы можем взять положительное ν достаточно маленьким).

Чтобы получить верхнюю границу для $T_1(n)$ воспользуемся неравенством $|\log(1 + z)| \le 2|z|$ for $|z| < \frac{1}{2}$. Отсюда

$$T_1(n) = \mathsf{E}\left((\log(1+\eta_n))^2\mathbb{I}\left\{|\eta_n| < \frac{1}{2}\right\}\right) \le 4\mathsf{E}\eta_n^2.$$

Таким образом,

$$\mathsf{E}\eta_n^2 = \frac{(k_n-1)P_n(1-P_n)+1}{((k_n-1)P_n)^2} \le \frac{1}{(k_n-1)P_n} + \frac{1}{((k_n-1)P_n)^2}. \qquad (77)$$

Как было отмечено ранее, для $(u,x,y) \in G_n$ и всех достаточно больших n имеем $(k_n-1)P_n \ge \frac{1}{4}n^{-\nu}k_n \to \infty$ при $0 < \nu < \alpha/2$ (мы также предполагаем, что $\nu < \beta - \theta$). Следовательно, правая часть (77) стремится к нулю при $n \to \infty$, а значит, $\sup_{(u,x,y)\in G_n} T_1(n) \to 0$, $n \to \infty$. Таким образом, соотношения (75) и (76) доказаны. Приходим к соотношению

$$\sup_{(u,x,y)\in G_n} |\mathsf{E}\log(\mu_n+1) - \log((k_n-1)P_n(\beta,u,x,y))| \to 0, \quad n \to \infty.$$

Введем обозначение $F_n := \frac{P_n(\beta,u,x,y)}{f(y|x)}$, где $f(y|x) > 0$. Тогда

$$\sup_{(u,x,y)\in G_n} |F_n - 1| = \sup_{(u,x,y)\in G_n} \frac{|P_n(\beta,u,x,y) - f(y|x)|}{f(y|x)}$$
$$\le cn^{-\beta+\theta+\nu} \to 0, \quad n \to \infty,$$

если $0 < \nu < \beta - \theta$, а c определяется отношением (74) и не зависит от n. Нетрудно видеть, что $\sup_{(u,x,y)\in G_n} |F_n - 1| < \frac{1}{2}$ для всех достаточно больших n. Тогда

$$|\log((k_n-1)P_n(\beta,u,x,y)) - \log((k_n-1)f(y|x))| = |\log(1+(F_n-1))| \le 2|F_n-1|.$$

Таким образом, приходим к соотношению

$$\sup_{(u,x,y)\in G_n} |\log F_n| \to 0, \quad n \to \infty.$$

Следовательно,

$$\sup_{(u,x,y)\in G_n} |\mathsf{E}\log(\mu_n+1) - \log((k_n-1)f(y|x)| \to 0, \quad n \to \infty. \qquad (78)$$

Похожим образом можно показать, что

$$\sup_{(u,x,y)\in G_n} |\mathsf{E}\log(\mu_n+2) - \log((k_n-1)f(y|x)| \to 0, \quad n \to \infty.$$

Учитывая формулу (71), мы получаем

$$\sup_{(u,x,y)\in G_n} |\mathsf{E}\left(\log(\xi_{n,k,1}(x,y)+1)|\rho_{n,k,1}(x) = un^{-\beta}\right) - \log((k-1)f(y|x)|$$
$$\le \sup_{(u,x,y)\in G_n} |\mathsf{E}\log(\mu_n+1) - \log((k-1)f(y|x)|$$
$$+ \sup_{(u,x,y)\in G_n} |\mathsf{E}\log(\mu_n+2) - \log((k-1)f(y|x)| \to 0$$

при $n \to \infty$. Итак,

$$\sum_{y\in M}\int_{B_{1,n}}\int_{(0,\delta]}\left(\mathsf{E}\left(\log\left(\frac{\xi_{n,k,1}(x,y)+1}{k}\right)\Big|\rho_{n,k,1}(x)=un^{-\beta}\right)\right.$$
$$\left.-\log\left(\frac{k-1}{k}\right)-\log f(y|x)\right)f^{(k)}_{n,x,\beta}(u)duf(x,y)dx$$
$$=S_2(n,k)-\sum_{y\in M}\int_{B_{1,n}}\int_{(0,\delta]}\left(\log\left(\frac{k-1}{k}\right)+\log f(y|x)\right)$$
$$\times f^{(k)}_{n,x,\beta}(u)duf(x,y)dx\to 0$$

при $n\to\infty$ (напомним, что $k=k_n$). Остается показать, что

$$-\sum_{y\in M}\int_{B_{1,n}}\int_{(0,\delta]}\left(\log\left(\frac{k-1}{k}\right)+\log f(y|x)\right)f^{(k)}_{n,x,\beta}(u)duf(x,y)dx$$
$$\to H(Y|X),\ \ n\to\infty.$$

Вначале заметим, что

$$0\le-\sum_{y\in M}\int_{B_{1,n}}\int_{(0,\delta]}\log\left(\frac{k-1}{k}\right)f^{(k)}_{n,x,\beta}(u)duf(x,y)dx$$
$$=-\log\left(1-\frac{1}{k}\right)\sum_{y\in M}\int_{B_{1,n}}\int_{(0,\delta]}f^{(k)}_{n,x,\beta}(u)duf(x,y)dx$$
$$\le-\log\left(1-\frac{1}{k}\right)\to 0,\ \ n\to\infty.$$

Далее

$$\sum_{y\in M}\int_{B_{1,n}}\int_{(0,\delta]}\log f(y|x)f^{(k)}_{n,x,\beta}(u)duf(x,y)dx$$
$$=\sum_{y\in M}\int_{B_{1,n}}\log f(y|x)\int_{(0,\delta]}f^{(k)}_{n,x,\beta}(u)duf(x,y)dx.$$

Неравенство (70) обеспечивает

$$0\le\Delta_n:=\sup_{x\in B_{1,n}}\left(1-\int_{(0,\delta]}f^{(k)}_{n,x,\beta}(u)du\right)\to 0,\ \ n\to\infty.$$

Таким образом,

$$\left|\sum_{y\in M}\int_{B_{1,n}}\int_{(0,\delta]}\log f(y|x)f^{(k)}_{n,x,\beta}(u)duf(x,y)dx-\sum_{y\in M}\int_{B_{1,n}}\log f(y|x)f(x,y)dx\right|$$
$$\leq\Delta_n\sum_{y\in M}\int_{\mathbb{R}^d}|\log f(y|x)|f(x,y)dx\to 0,\ \ n\to\infty.$$

Теперь заметим, что

$$\sum_{y\in M}\int_{B_{1,n}}\log f(y|x)f(x,y)dx\to\sum_{y\in M}\int_{\mathbb{R}^d}\log f(y|x)f(x,y)dx,\ \ n\to\infty,$$

поскольку $B_{1,n}\nearrow\{(x,y):f_X(x)>0,\cap_{y\in M}\{f(y|x)>0\}\}$ и $\int_B h(x)dx=0$, где $h\in L^1(\mathbb{R})$ и B является борелевским подмножеством $\mathbb{R}^d$ таким, что $\mu(B)=0$. Следовательно,

$$-\sum_{y\in M}\int_{B_{1,n}}\int_0^{\delta}\log f(y|x)f^{(k)}_{n,x,\beta}(u)duf(x,y)dx\to H(Y|X),\ \ n\to\infty.$$

Для завершения доказательства теоремы 2.3 определим параметры $\beta>0$, $\theta>0$, и $\nu>0$ таким образом, чтобы выполнялись следующие условия:

$$\begin{gathered}\beta>\theta,\\ \nu<\beta-\theta,\\ (d+1)\theta<1-\alpha,\\ \alpha-2\nu>0.\end{gathered}$$

Для каждого фиксированного $\alpha\in(0,1)$ можно гарантировать выполнение указанных неравенств, поскольку можно выбрать $\beta\in(0,\frac{1-\alpha}{d+1})$, а затем определить $\theta\in(0,\beta)$. После этого остается зафиксировать $\nu\in(0,\beta-\theta)$ так, чтобы $\nu<\frac{1}{2}\alpha$.

Таким образом, теорема 2.3 полностью доказана. □

Замечание 2.11 *Ведущий член в оценке асимптотики смещения математического ожидания вводимой оценки относительно оцениваемой энтропии оценивается неравенством* (59)*, что позволяет установить скорость сходимости математического ожидания оценки к оцениваемой величине:*

$$\mathsf{E}\widehat{H}_{n,k}-H=O((\log n)^{-\varepsilon}),\ \ n\to\infty,\tag{79}$$

где ε определяется так же, как и в (47). *В недавних работах [52, 57] исследовалось асимптотическое поведение риска определенных оценок дифференциальной энтропии Шеннона при выполнении определенных условия на плотность f (например, f должна принадлежать некоторому шару Гельдера в пространстве функций $f : [0,1]^d \to \mathbb{R}_+$). Мы используем другие гипотезы для изучения оценки условной энтропии Шеннона в смешанной модели. Для оценки, предложенной в работе [44], имеет место степенная скорость сходимости смещения в случае ограниченной плотности f, удовлетворяющей ряду строгих условий, поэтому более низкая скорость сходимости смещения к нулю оценки $\widehat{H}_{n,k}$ в рассматриваемой модели является вполне естественной за счет более широких условий, накладываемых на плотность.*

Замечание 2.12 *Можно рассмотреть модификацию оценки, заменив* $\log k$ *на* $\psi(k)$, *где* ψ — *дигамма функция. Для этого введем* $\widetilde{H}_{n,k}$ *согласно определению* (41) *с*

$$\widetilde{H}_{n,k,i} = -\log(\xi_{n,k,i}(Z^1, \dots, Z^n) + 1) + \psi(k)$$

вместо $\widehat{H}_{n,k,i}$.

Такая поправка используется при произвольном фиксированном k для уменьшения смещения оценки дифференциальной энтропии Козаченко-Леоненко и для оценки взаимной информации Краскова-Штогбауэра-Грассбергера. В нашем случае $k = k(n)$, где $k(n) \propto n^\alpha$, $n \to \infty$, $\alpha \in (0,1)$. Поскольку (см., например, теорему 3.7 в [72])

$$\psi(z) = \log z - \frac{1}{z} + O\left(\log\left(1 + \frac{1}{z}\right)\right), \quad z \to \infty,$$

то доказательство теоремы 2.3 приводит к (79), *если мы заменим* $\widehat{H}_{n,k}$ *на* $\widetilde{H}_{n,k}$.

2.4.2 Доказательство теоремы 2.4

Доказательство теоремы разбивается на несколько шагов. Поскольку $Z^1, Z^2, \dots$ являются независимыми одинаково распределенными случайными векторами, то

$$\mathsf{E}(\widehat{H}_{n,k} - H(Y|X))^2 = \frac{1}{n}\mathsf{E}(\widehat{H}_{n,k,1} - H(Y|X))^2$$
$$+ \left(1 - \frac{1}{n}\right)\mathsf{E}(\widehat{H}_{n,k,1} - H(Y|X))(\widehat{H}_{n,k,2} - H(Y|X)).$$

Далее мы докажем, что математическое ожидание правой части в последней формуле конечно. Более того, будет показано, что при $n \to \infty$

(A) $\mathsf{E}(\widehat{H}_{n,k,1} - H(Y|X))(\widehat{H}_{n,k,2} - H(Y|X)) = o(1)$,

(B) $\mathsf{E}(\widehat{H}_{n,k,1} - H(Y|X))^2 = o(n)$.

Доказательство (A) достаточно трудоемко и включает в себя оценивание множества интегралов, зависящих от нескольких вспомогательных параметров. Эти параметры будут подобраны особым образом. В отличие от доказательства теоремы 2.3, мы будем использовать условные математические ожидания при условии векторов $\zeta = (\rho_{n,k,1}(x_1, x_2), \rho_{n,k,2}(x_1, x_2))$, где $x_1, x_2 \in \mathbb{R}^d$.

Ключевую роль играет лемма 2.8, описывающая условное распределение вектора $(\xi_{n,k,1}(z_1, z_2), \xi_{n,k,2}(z_1, z_2))$ при условии вектора ζ, где $z_i = (x_i, y_i) \in \mathbb{R}^d \times M$, $i = 1, 2$. Она же позволяет доказать (B).

Шаг 1. Доказательство (A) может быть разбито на два основных этапа. Далее понадобятся несколько новых обозначений. Рассмотрим $n > 2$, $k \in \{1, \dots, n\}$, $z_j = (x_j, y_j) \in \mathbb{R}^d \times M$, $j \in \{1, 2\}$. Положим

$$\widehat{H}_{n,k,j}(z_1, z_2) := -\log\left(\frac{\xi_{n,k,j}(z_1, z_2, Z^3, \dots, Z^n) + 1}{k}\right).$$

Независимость наблюдений $Z^1, \dots, Z^n$ влечет за собой равенство

$$\begin{aligned}\mathsf{E}\Big((\widehat{H}_{n,k,1} - H(Y|X))(\widehat{H}_{n,k,2} - H(Y|X))|Z^1 = z_1, Z^2 = z_2\Big)\\ = \mathsf{E}(\widehat{H}_{n,k,1}(z_1, z_2) - H(Y|X))(\widehat{H}_{n,k,2}(z_1, z_2) - H(Y|X)) =: \mathcal{H}_{n,k}(z_1, z_2).\end{aligned}$$

Следовательно,

$$\begin{aligned}\mathsf{E}(\widehat{H}_{n,k,1} - H(Y|X))(\widehat{H}_{n,k,2} - H(Y|X))\\ = \sum_{y_1, y_2 = 1}^{m} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \mathcal{H}_{n,k}(z_1, z_2) f(x_1, y_1) f(x_2, y_2)\, dx_1 dx_2.\end{aligned}$$

По теореме де ла Валле-Пуссена (см., например, [22], стр. 10), для того, чтобы установить соотношение (A) достаточно показать, что верны два следующих утверждения.

1. Если $dQ(x_1, x_2) := f(x_1, y_1) f(x_2, y_2)\ dx_1 dx_2$, то для каждого $y_1, y_2 \in M$ и Q-п.н. $(x_1, x_2) \in \mathbb{R}^d \times \mathbb{R}^d$,

$$\mathcal{H}_{n,k}(z_1, z_2) \to h(x_1, y_1)h(x_2, y_2), \quad n \to \infty, \tag{80}$$

где $h(x, y) := -\log f(y|x) - H(Y|X)$, $x \in \mathbb{R}^d$, $y \in M$, $z_j = (x_j, y_j)$, $j \in \{1, 2\}$.

2. Для некоторого $a > 0$,

$$\sup_n \sum_{y_1, y_2 \in M} \int\limits_{\mathbb{R}^d} \int\limits_{\mathbb{R}^d} |\mathcal{H}_{n,k}(z_1, z_2)|^{1+a} f(x_1, y_1) f(x_2, y_2)\, dx_1 dx_2 < \infty. \tag{81}$$

Действительно, формулы (80) и (81) влекут за собой выполнение

$$\sum_{y_1, y_2 \in M} \int\limits_{\mathbb{R}^d} \int\limits_{\mathbb{R}^d} \mathcal{H}_{n,k}(z_1, z_2) f(x_1, y_1) f(x_2, y_2)\, dx_1 dx_2$$
$$\to \sum_{y_1, y_2 \in M} \int\limits_{\mathbb{R}^d} h(x_1, y_1) f(x_1, y_1)\, dx_1 \int\limits_{\mathbb{R}^d} h(x_2, y_2) f(x_2, y_2)\, dx_2 = 0, \quad n \to \infty.$$

Шаг 2. Докажем (81). Из условного неравенства Йенсена немедленно следует, что (81) выполняется, если

$$\sup_n \mathsf{E} \left| (\widehat{H}_{n,k,1} - H(Y|X))(\widehat{H}_{n,k,2} - H(Y|X)) \right|^{1+a} < \infty.$$

По неравенству Коши-Буняковского-Шварца

$$\mathsf{E} \left| (\widehat{H}_{n,k,1} - H(Y|X))(\widehat{H}_{n,k,2} - H(Y|X)) \right|^{1+a} \le \mathsf{E} |\widehat{H}_{n,k,1} - H(Y|X)|^{2+2a}.$$

Таким образом, поскольку $a > 0$ может быть сколь угодно маленьким, то (81) выполняется, если для некоторого $\varepsilon > 0$

$$\sup_n \mathsf{E} |\widehat{H}_{n,k,1} - H(Y|X)|^{2+\varepsilon} < \infty. \tag{82}$$

Неравенство Ляпунова для моментов случайной величины позволяет утверждать, что (82) обеспечивает выполнение (B).

Рассуждение, приведенное в доказательстве теоремы 2.3, приводит к выражению

$$\mathsf{E} \left| -\log \left(\frac{\xi_{n,k,1}(Z^1, \ldots, Z^n) + 1}{k} \right) - H(Y|X) \right|^{2+\varepsilon}$$
$$= \sum_{y \in M} \int\limits_{\mathbb{R}^d} \mathsf{E} \left| \log \left(\frac{\xi_{n,k,1}(x, y) + 1}{k} \right) + H(Y|X) \right|^{2+\varepsilon} f(x, y)\, dx. \tag{83}$$

Зафиксируем произвольное $\beta > 0$ и для $x \in \mathbb{R}^d$, $y \in M$ и $u > 0$ положим

$$I_{n,k,\beta,\varepsilon}(x,y,u) := \mathsf{E}\left(\left|\log\left(\frac{\xi_{n,k,1}(x,y)+1}{k}\right) + H(Y|X)\right|^{2+\varepsilon}\middle| n^\beta \rho_{n,k,1} = u\right).$$

Тогда имеем

$$\mathsf{E}\left|\log\left(\frac{\xi_{n,k,1}(x,y)+1}{k}\right) + H(Y|X)\right|^{2+\varepsilon}, = \int_0^\infty I_{n,k,\beta,\varepsilon}(x,y,u) f^{(k)}_{n,x,\beta}(u)\, du$$

где $f^{(k)}_{n,x,\beta}(\cdot)$ — плотность случайной величины $n^\beta \rho_{n,k,1}$ (см. формулу (103)). Поскольку множестов M конечно, то достаточно доказать, что для всех $y \in M$ верно неравенство $\sup_n |\mathcal{I}_{n,k,i}(y)| < \infty$, где

$$\mathcal{I}_{n,k,i}(y) := \int\limits_{B_{i,n}} \int_0^\infty I_{n,k,\beta,\varepsilon}(x,y,u) f^{(k)}_{n,x,\beta}(u) f(x,y)\, du dx, \quad i = 1,2,$$

и $B_{i,n}$ определялось в (58).

Шаг 3. Теперь оценим значение интеграла $\mathcal{I}_{n,k,2}(y)$. Очевидно, что при $k > 1$ для всех $x \in \mathbb{R}^d$ и $y \in M$ следующее неравенство выполняется P-п.н.:

$$\left|\log\left(\frac{\xi_{n,k,1}(x,y)+1}{k}\right) + H(Y|X)\right| \le \log k + |H(Y|X)|. \tag{84}$$

Следовательно, используя неравенства (56) и (57) с $2+\varepsilon$ вместо $1+\varepsilon$, мы приходим к соотношению $\sup_n |\mathcal{I}_{n,k,2}| < \infty$.

Шаг 4. Перепишем $\mathcal{I}_{n,k,1}(y)$ в виде суммы двух слагаемых и рассмотрим первое из них. Зафиксируем $\delta > 0$. Имеем

$$\mathcal{I}_{n,k,1}(y) = S_{n,k,1}(y) + S_{n,k,2}(y),$$

где для $V_1 = (\delta, \infty)$ и $V_2 = (0, \delta]$,

$$S_{n,k,j}(y) = \int\limits_{B_{1,n}} \int_{V_j} I_{n,k,\beta,\varepsilon}(x,y,u) f^{(k)}_{n,x,\beta}(u) f(x,y)\, du dx, \quad j = 1,2.$$

По аналогии с (60) для каждого $y \in M$ и всех достаточно больших n, с помощью формулы (84) получим

$$0 \le S_{n,k,1}(y) \le (\log k + |H(Y|X)|)^{2+\varepsilon} \int\limits_{B_{1,n}} \int_\delta^\infty f^{(k)}_{n,x,\beta}(u) f(x,y)\, du\, dx$$

$$\le (\log k + |H(Y|X)|)^{2+\varepsilon} \int\limits_{B_{1,n}} \mathsf{P}(\eta_n(\delta,\beta,x) \le k-1) f_X(x)\, dx.$$

Согласно неравенству (70)

$$0 \le S_{n,k,1}(y) \le \frac{(\log k + |H(Y|X)|)^{2+\varepsilon}}{\sqrt{2\pi R_n}} e^{-R_n/2} \to 0, \quad n \to \infty, \tag{85}$$

где R_n было введено в (69).

Шаг 5. Теперь перейдем к оцениванию значения $S_{n,k,2}(y)$. Лемма 2.6 утверждает, что

$$I_{n,k,\beta,\varepsilon}(x,y,u) = \mathsf{E}\left|\log\left(\frac{\mu_n+1}{k}\right) + H(Y|X)\right|^{2+\varepsilon} \mathsf{P}(Y \ne y | \|X-x\| = un^{-\beta})$$
$$+ \mathsf{E}\left|\log\left(\frac{\mu_n+2}{k}\right) + H(Y|X)\right|^{2+\varepsilon} \mathsf{P}(Y = y | \|X-x\| = un^{-\beta}),$$

где случайная переменная μ_n определялась выражением (72) и не зависит от (X,Y). Мы покажем, что

$$\sup_{(u,x,y)\in G_n} \left| I_{n,k,\beta,\varepsilon}(x,y,u) - |\log f(y|x) + H(Y|X)|^{2+\varepsilon} \right| \to 0, \quad n \to \infty, \tag{86}$$

где $G_n = (0,\delta] \times B_{1,n} \times M$. Достаточно показать, что при $n \to \infty$,

$$\sup_{(u,x,y)\in G_n} \left| \mathsf{E}\left|\log\left(\frac{\mu_n+1}{k}\right) + H(Y|X)\right|^{2+\varepsilon} - |\log f(y|x) + H(Y|X)|^{2+\varepsilon} \right| \to 0, \tag{87}$$

$$\sup_{(u,x,y)\in G_n} \left| \mathsf{E}\left|\log\left(\frac{\mu_n+2}{k}\right) + H(Y|X)\right|^{2+\varepsilon} - |\log f(y|x) + H(Y|X)|^{2+\varepsilon} \right| \to 0. \tag{88}$$

Для $\nu \in (0,1/2)$ и $\gamma \in (0, 1/2-\nu)$ верно представление

$$\mathsf{E}\left|\log(\frac{\mu_n+1}{k}) + H(Y|X)\right|^{2+\varepsilon} = T_{n,1} + T_{n,2},$$

где

$$T_{n,1} := \mathsf{E}\left(\left|\log\left(\frac{\mu_n+1}{k}\right) + H(Y|X)\right|^{2+\varepsilon} \mathbb{I}\{|\mu_n - \mathsf{E}\mu_n| \le (k-1)^{1/2+\gamma}\}\right),$$

$$T_{n,2} := \mathsf{E}\left(\left|\log\left(\frac{\mu_n+1}{k}\right) + H(Y|X)\right|^{2+\varepsilon} \mathbb{I}\{|\mu_n - \mathsf{E}\mu_n| > (k-1)^{1/2+\gamma}\}\right),$$

здесь $T_{n,j} = T_{n,j}(\beta, \varepsilon, \gamma, u, x, y)$, $j = 1, 2$. По свойствам биномиального распределения $\mathsf{E}\mu_n = (k-1)P_n(\beta, u, x, y)$, где $P_n(\beta, u, x, y)$ определялось в (73).

В силу (84) для $(u, x, y) \in G_n$ и всех достаточно больших n верно неравенство

$$0 \le \left| \log\left(\frac{\mu_n + 1}{k}\right) + H(Y|X) \right|^{2+\varepsilon} \le (\log k + |H(Y|X)|)^{2+\varepsilon}.$$

Применяя неравенство Хефдинга с $m = k-1$, $H(m) = \mu_n$, $p = P_n$, $\varepsilon = (k-1)^{-1/2+\gamma}$, получим

$$\begin{aligned} 0 \le T_{n,2} &\le (\log k + |H(Y|X)|)^{2+\varepsilon} \mathsf{P}(|\mu_n - \mathsf{E}\mu_n| > (k-1)^{1/2+\gamma}) \\ &\le (\log k + |H(Y|X)|)^{2+\varepsilon} 2e^{-2(k-1)^{2\gamma}} \to 0, \quad n \to \infty. \end{aligned} \tag{89}$$

Также отметим, что

$$|\log f(y|x) + H(Y|X)|^{2+\varepsilon} \mathsf{P}(|\mu_n - \mathsf{E}\mu_n| > (k-1)^{1/2+\gamma}) \to 0, \quad n \to \infty. \tag{90}$$

Определим $B_i(n, p) = \mathsf{P}(\mu_n = i)$, $i = 0, \ldots, k-1$, и запишем $T_{n,1}$ следующим образом:

$$T_{n,1} = \sum_{i : |i - \mathsf{E}\mu_n| \le (k-1)^{1/2+\gamma}} \left| \log\left(\frac{i+1}{k}\right) + H(Y|X) \right|^{2+\varepsilon} B_i(n, p). \tag{91}$$

Чтобы получить верхнюю границу для $|T_{n,1} - |\log f(y|x) + H(Y|X)|^{2+\varepsilon}|$, применим формулу Лагранжа к функции $g(z) = |\log z + a|^{2+\varepsilon}$, $z > 0$, $a \in \mathbb{R}$. Для $z, z_0 > 0$

$$g(z) - g(z_0) = g'(\xi)(z - z_0), \quad \xi = z_0 + \lambda(z - z_0), \quad \lambda = \lambda(z, z_0) \in (0, 1).$$

Определим $z = \frac{i+1}{k}$, $z_0 = f(y|x)$ и $a = H(Y|X)$. Тогда в силу (74) для всех i, принадлежащих множеству (91), получим

$$\begin{aligned} |z - z_0| &= \left| \frac{i+1}{k} - f(y|x) \right| \\ &\le \left| \frac{i - (k-1)P_n}{k} \right| + \frac{1}{k} + \left| \frac{(k-1)P_n}{k} - P_n \right| + |P_n - f(y|x)| \\ &\le (k-1)^{-1/2+\gamma} + 2k^{-1} + 2C\delta(\sharp M + 1)n^{-\beta+\theta} := \mathcal{Z}_n \propto n^{-\beta+\theta}, \end{aligned}$$

где $\beta - \theta < 1/2 - \gamma$. Заметим, что

$$g'(z) = (2+\varepsilon)z^{-1}(\log z + a)|\log z + a|^{\varepsilon}, \quad z > 0.$$

Следовательно, $|g'(\xi)| \le (2+\varepsilon)|\xi|^{-1}|\log \xi + a|^{1+\varepsilon}$. Для всех достаточно больших n верно неравенство

$$|\log \xi| \le \max\{|\log z_0|, |\log z|\} \le \max\{|\log \nu_n|, \log k\} \le c \log n,$$

где $c = c(\alpha, \nu)$. Понятно, что $|\xi| \ge \min\{z, z_0\}$. В нашем случае $z_0 \ge n^{-\nu}$ и $z \ge P_n(k_n - 1) - (k_n - 1)^{\frac{1}{2}+\gamma}$. Из неравенства (74) следует, что $P_n \ge \frac{1}{2}n^{-\nu}$ для всех достаточно больших n, если $\nu < \beta - \theta$. Следовательно, при всех достаточно больших n неравенство $z \ge c_1 n^{-\nu+\alpha}$ выполняется, когда $\alpha - \nu > \frac{1}{2} + \gamma$ (здесь $c_1 = c_1(\alpha)$). Последнее неравенство выполняется, если $\nu < 2\alpha$ (затем мы выберем положительное достаточно малое γ). Таким образом, $|\xi|^{-1} \le n^{\nu}$ для всех достаточно больших n. Поэтому равномерно по всем i, принадлежащим множеству, по которому ведется суммирование в $T_{n,1}$,

$$\begin{aligned} |g(z) - g(z_0)| &= |g'(\xi)||z - z_0| \\ &\le (2+\varepsilon)(\log n)^{1+\varepsilon} n^{\nu} \mathcal{Z}_n \propto (\log n)^{1+\varepsilon} n^{\nu-\beta+\theta} \to 0, \quad n \to \infty, \end{aligned}$$

где $\nu < \beta - \theta$. Учитывая неравенство $\sum_{i:|i-\mathsf{E}\mu_n| \le (k-1)^{1/2+\gamma}} B_i(n,p) \le 1$, приходим к соотношению

$$\sup_{(u,x,y)\in G_n} |T_{n,1} - |\log f(y|x) + H(Y|X)|^{2+\varepsilon}| \to 0, \quad n \to \infty.$$

Эта формула вместе с (89) и (90) влечет (87). Соотношение (88) доказывается аналогично. Таким образом, мы установили (86). Следовательно,

$$S_{n,k,2}(y) - \int_{B_{1,n}} \int_0^{\delta} |\log f(y|x) + H(Y|X)|^{2+\varepsilon} f^{(k)}_{n,x,\beta}(u) f(x,y)\, du dx \to 0, \quad n \to \infty.$$

Очевидно, что для всех $n \in \mathbb{N}$ и $y \in M$ верно неравенство

$$\begin{aligned} &\int_{B_{1,n}} \int_0^{\delta} |\log f(y|x) + H(Y|X)|^{2+\varepsilon} f^{(k)}_{n,x,\beta}(u) f(x,y)\, du dx \\ &\le \int_{\mathbb{R}^d \cap \{x: f(y|x) > 0\}} |\log f(y|x) + H(Y|X)|^{2+\varepsilon} f(x,y)\, dx < \infty. \end{aligned}$$

Действительно, в силу неравенства Минковского и поскольку $f(y|x) = \mathsf{P}(Y = y|X = x) \le 1$ для всех $y \in M$ и P_X-почти всех x, достаточно показать, что

$$\int_{\mathbb{R}^d \cap \{x: f(y|x) > 0\}} \left(\log \frac{1}{f(y|x)} \right)^{2+\varepsilon} f(x,y)\, dx < \infty, \quad y \in M.$$

Для всех $\varepsilon > 0$ существует такое $\mathcal{T} = \mathcal{T}(\varepsilon) > 1$, что $(\log t)^{2+\varepsilon} \le t$, где $t > \mathcal{T}$. Значит, для всех $y \in M$, последний интеграл может быть представлен в следующем виде:

$$\begin{aligned}
&\int_{\{x: f(y|x) \ge 1/\mathcal{T}\}} \left(\log \frac{1}{f(y|x)} \right)^{2+\varepsilon} f(x,y)\, dx \\
&\qquad + \int_{\{x: 0 < f(y|x) < 1/\mathcal{T}\}} \left(\log \frac{1}{f(y|x)} \right)^{2+\varepsilon} f(y|x) f_X(x)\, dx \\
&\quad \le (\log \mathcal{T})^{2+\varepsilon} \mathsf{P}(Y = y) + \int_{\{x: f(y|x) < 1/T\}} (f(y|x))^{-1} f(y|x) f_X(x)\, dx \\
&\qquad \le (\log \mathcal{T})^{2+\varepsilon} + 1 < \infty.
\end{aligned}$$

Поэтому для $y \in M$ имеем $\sup_n S_{n,k,2}(y) < \infty$. Таким образом, применяя (85), можно убедиться, что для всех $y \in M$

$$\sup_n |\mathcal{I}_{n,k,1}(y)| < \infty.$$

Соотношение (82) доказано.

Шаг 6. Теперь перейдем к доказательству (80). Для $z_1, z_2 \in \mathbb{R}^d \times M$, $n > 2$ и $k = k_n$ введем

$$R_{n,k}(z_1, z_2) := (\widehat{H}_{n,k,1}(z_1, z_2) - H(Y|X))(\widehat{H}_{n,k,2}(z_1, z_2) - H(Y|X)).$$

Для простоты обозначений положим $\rho_{n,k,j} := \rho_{n,k,j}(x_1, x_2)$, $j = 1, 2$. Зафиксируем $\delta > 0$ и определим

$$\begin{aligned}
&\mathcal{U}_{n,k,1}(z_1, z_2) \\
&\quad := \mathsf{E}(R_{n,k}(z_1, z_2) | \rho_{n,k,1} n^\beta \le \delta, \rho_{n,k,2} n^\beta \le \delta) \mathsf{P}(\rho_{n,k,1} n^\beta \le \delta, \rho_{n,k,2} n^\beta \le \delta),
\end{aligned}$$

$$\mathcal{U}_{n,k,2}(z_1, z_2) := \mathcal{H}_{n,k}(z_1, z_2) - \mathcal{U}_{n,k,1}(z_1, z_2),$$

где для интегрируемой случайной величины ξ предполагается, что $\mathsf{E}(\xi|A) := 0$, если $\mathsf{P}(A) = 0$. Для всех $n \ge N_1$, где $N_1 = N_1(H(Y|X))$,

и всех $z_j \in \mathbb{R}^d \times M$, $j = 1, 2$, неравенство $|R_{n,k}(z_1, z_2)| \le 2(\log k)^2$ выполняется с вероятностью 1. Следовательно,

$$|\mathcal{U}_{n,k,2}(z_1, z_2)| \le 2(\log k)^2 \mathsf{P}\left(\left\{\rho_{n,k,1} > \delta n^{-\beta}\right\} \cup \left\{\rho_{n,k,2} > \delta n^{-\beta}\right\}\right) \le 4(\log k)^2 \mathsf{P}\left(\rho_{n,k,1} > \delta n^{-\beta}\right).$$

В силу (61) и (70), для всех $z_j \in \mathbb{R}^d \times M$, $j=1,2$, получаем $|\mathcal{U}_{n,k,2}(z_1,z_2)| \to 0$, $n \to \infty$.

Шаг 7. Остается показать, что для всех $y_1, y_2 \in M$ и Q-почти всех $(x_1, x_2) \in \mathbb{R}^d \times \mathbb{R}^d$,

$$\mathcal{U}_{n,k,1}(z_1, z_2) \to (-\log f(y_1|x_1) - H(Y|X))(-\log f(y_2|x_2) - H(Y|X)), \quad n \to \infty. \tag{92}$$

При доказательстве леммы 2.8 (см. раздел 2.4.5) показывается, что для всех $x_1, x_2 \in \mathbb{R}^d$ $(x_1 \ne x_2)$, конечная мера $G(B) := \mathsf{P}((\rho_{n,k,1}, \rho_{n,k,2}) \in B)$, определенная для $B \in \mathcal{B}(\mathbb{R}^2) \cap D_{x_1,x_2}$, где

$$D_{x_1,x_2} = (0, |x_1 - x_2|/3] \times (0, |x_1 - x_2|/3],$$

может быть представлена как

$$\mathsf{P}((\rho_{n,k,1}, \rho_{n,k,2}) \in B) = \int_B g_{n,k}(x_1, x_2, u_1, u_2)\, du_1 du_2.$$

Здесь $g_{n,k}(x_1, x_2, \cdot, \cdot)\colon D_{x_1,x_2} \to \mathbb{R}_+$ — некоторая интегрируемая функция (относительно ограничения меры Лебега $\Lambda \otimes \Lambda$, заданной на $\mathbb{R}^2$, на D_{x_1,x_2}). Очевидно, что для $0 < k < n - 2$,

$$\mathsf{P}(\rho_{n,k,i}(x_1, x_2) = 0) = \sum_{j=k}^{n-2} \binom{n-2}{j} \mathsf{P}(\|X - x_i\| = 0)^j (1 - \mathsf{P}(\|X - x_i\| = 0))^{n-2-j} = 0, \quad i = 1, 2,$$

поскольку X имеет плотность (и $k = k(n) \propto n^\alpha$, $\alpha \in (0, 1)$).

Теперь положим $g_{n,k}(x_1, x_2, \cdot, \cdot) = 0$ на $\overline{D}_{x_1,x_2} \setminus D_{x_1,x_2}$, где $\overline{D}_{x_1,x_2}$ обозначает замыкание множества D_{x_1,x_2} в $\mathbb{R}^2$. Возьмем $N(x_1, x_2, \beta, \delta)$ такое, чтобы $\delta n^{-\beta} \le |x_1 - x_2|/6$ для всех $n \ge N(x_1, x_2, \beta, \delta)$. Тогда

$$Q_{n,\beta,\delta} := [0, \delta n^{-\beta}] \times [0, \delta n^{-\beta}] \subset \overline{D}_{x_1,x_2},$$

$$B(x_1, u_1 n^{-\beta}) \cap B(x_2, u_1 n^{-\beta}) = \varnothing$$

для $u_i \in [0,\delta]$, $i = 1,2$. Таким образом, для всех $B \in \mathcal{B}(\mathbb{R}^2)$ и $n \geq N(x_1,x_2,\beta,\delta)$,

$$\begin{aligned}\mathsf{P}(\zeta \in B | \zeta \in Q_{n,\delta,\beta}) &= \frac{\mathsf{P}(\zeta \in B \cap Q_{n,\delta,\beta})}{\mathsf{P}(\zeta \in Q_{n,\delta,\beta})} \\ &= \frac{1}{\mathsf{P}(\zeta \in Q_{n,\delta,\beta})} \int_{B \cap Q_{n,\delta,\beta}} g_{n,k}(x_1,x_2,u_1,u_2)\, du_1 du_2,\end{aligned}$$

поскольку $B \cap Q_{n,\delta,\beta} \subset \overline{D}_{x_1,x_2}$. Отметим, что $\mathsf{P}(\zeta \in Q_{n,\delta,\beta}) > 0$ для всех достаточно больших n

$$\begin{aligned}&\mathsf{P}(\zeta \in Q_{n,\delta,\beta}) \\ &\quad = \sum_{\substack{l_1 \geq k, l_2 \geq k, \\ l_1+l_2 \leq n-2}} \binom{n}{l_1} \binom{n-l_1}{l_2} p_{n,x_1,\delta,\beta}^{l_1} p_{n,x_2,\delta,\beta}^{l_2} (1 - p_{n,x_1,\delta,\beta} - p_{n,x_2,\delta,\beta}))^{n-l_1-l_2}.\end{aligned}$$

Здесь $p_{n,x_i,\delta,\beta} = \mathsf{P}(\|X - x_i\| \leq \delta n^{-\beta}) > 0$ для всех $n \in \mathbb{N}$ и $i = 1,2$, поскольку $\mathsf{P}_X(B(x,r)) > 0$ для всех $x \in \mathbb{R}^d$ и $r > 0$ (так как $f_X(z) > 0$ для μ-почти всех $z \in \mathbb{R}^d$). Мы также учитываем, что $p_{n,x_i,\delta,\beta} \to 0$ при $n \to \infty$, $i = 1,2$.

Следовательно, функция $\tilde{g}_{n,k}(x_1,x_2,\cdot,\cdot) : \mathbb{R}^2 \to \mathbb{R}_+$, задаваемая формулой

$$\tilde{g}_{n,k}(x_1,x_2,u_1,u_2) = \begin{cases} \frac{1}{\mathsf{P}(\zeta \in Q_{n,\delta,\beta})} g_{n,k}(x_1,x_2,u_1,u_2), & \text{если } (u_1,u_2) \in Q_{n,\delta,\beta}, \\ 0, & \text{если } (u_1,u_2) \in \mathbb{R}^2 \setminus Q_{n,\delta,\beta}, \end{cases}$$

является вероятностной плотностью меры $\mathsf{P}(\zeta \in \cdot | \zeta \in Q_{n,\delta,\beta})$, которая определена на $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$. Таким образом, мера $\mathsf{P}(n^\beta \zeta \in \cdot | \zeta \in Q_{n,\delta,\beta}))$ на этом пространстве имеет плотность (относительно ограничения $\Lambda \otimes \Lambda$ на $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$) $f_{n,k,\beta,\delta}(x_1,x_2,\cdot,\cdot) \colon \mathbb{R}^2 \to \mathbb{R}_+$.

Теперь применим лемму 2.9 для

$$\begin{gathered}W := R_{n,k}(z_1,z_2), \quad V := (n^\beta \rho_{n,k,1}(x_1,x_2), n^\beta \rho_{n,k,2}(x_1,x_2)), \\ B := \{(u_1,u_2) \in [0,\delta] \times [0,\delta]\}.\end{gathered}$$

Заметим, что $\{V \in B\} = \{\zeta \in Q_{n,\delta,\beta}\}$ и $\mathsf{E}W$ существуют. Следовательно, для рассматриваемых x_1, x_2, β, δ и $n > N(x_1,x_2,\beta,\delta)$ верна формула

$$\begin{aligned}\mathcal{U}_{n,k,1}(z_1,z_2) &= \mathsf{E}(W | V \in B)\, \mathsf{P}(\rho_{n,k,1} n^\beta \leq \delta, \rho_{n,k,2} n^\beta \leq \delta) \\ &= \int_0^\delta \int_0^\delta J_{n,k,\beta}(z_1,z_2,u_1,u_2)\, d\mathsf{P}_{V,B}(u_1,u_2)\, \mathsf{P}(\rho_{n,k,1} n^\beta \leq \delta, \rho_{n,k,2} n^\beta \leq \delta),\end{aligned}$$

где $J_{n,k,\beta}(z_1,z_2,u_1,u_2) := \mathsf{E}(R_{n,k}(z_1,z_2)|V=(u_1,u_2))$. Было показано, что $\mathsf{P}_{V,B}$ имеет плотность $f_{n,k,\beta,\delta}(x_1,x_1,\cdot,\cdot)$, следовательно,

$$\mathcal{U}_{n,k,1}(z_1,z_2) = \int_0^\delta \int_0^\delta J_{n,k,\beta}(z_1,z_2,u_1,u_2) f_{n,k,\beta,\delta}(x_1,x_1,u_1,u_2)\,du_1 du_2 \times \mathsf{P}(\rho_{n,k,1} n^\beta \le \delta, \rho_{n,k,2} n^\beta \le \delta).$$

Шаг 8. Теперь покажем, что равномерно для $\mathsf{P}_{V,B}$-почти всех $(u_1,u_2) \in (0,\delta] \times (0,\delta]$

$$|J_{n,k,\beta}(z_1,z_2,u_1,u_2) - (-\log f(y_1|x_1) - H(Y|X))(-\log f(y_2|x_2) - H(Y|X))| \to 0, \quad n \to \infty. \quad (93)$$

Положим $\xi := (\xi_{n,k,1}, \xi_{n,k,2})$, где $\xi_{n,k,1} = \xi_{n,k,1}(z_1,z_2)$, $i=1,2$. В силу леммы 2.10 для P_V-почти всех (u_1,u_2)

$$J_{n,k,\beta}(z_1,z_2,u_1,u_2) = \sum_{r_1,r_2} \mathsf{E}(R_{n,k}(z_1,z_2)|V=(u_1,u_2),\xi=(r_1,r_2))\mathsf{P}(\xi=(r_1,r_2)|V=(u_1,u_2)).$$

Заметим, что для $\mathsf{P}_{V,\xi}$-почти всех (u_1,u_2,r_1,r_2)

$$\mathsf{E}(R_{n,k}(z_1,z_2)|V=(u_1,u_2),\xi=(r_1,r_2))\mathsf{P}(\xi=(r_1,r_2)|V=(u_1,u_2)) = \left(-\log\left(\frac{r_1+1}{k}\right) - H(Y|X)\right)\left(-\log\left(\frac{r_2+1}{k}\right) - H(Y|X)\right) := h(r_1,r_2), \quad (94)$$

поскольку случайная величина $R_{n,k}(z_1,z_2)$ является измеримой относительно σ-алгебры $\sigma\{V,\xi\}$. Функция $h(r_1,r_2)$ зависит также от n,k,z_1,z_2.

Пусть $O \in \mathcal{B}(\mathbb{R}_+^2 \times M^2)$ является множеством, состоящим их всех таких (u_1,u_2,r_1,r_2), что (94) выполняется. Тогда $\mathsf{P}_{V,\xi}(O)=1$. Поскольку M^2 является конечным множеством, то $O = \bigcup_{(r_1,r_2)\in M^2} O_{r_1,r_2} \times \{(r_1,r_2)\}$, где $O_{r_1,r_2} \in \mathcal{B}(\mathbb{R}_+^2)$. Отметим, что по крайней мере одно из множеств O_{r_1,r_2} непусто, иначе $\mathsf{P}((V,\xi)\in O) \ne 1$. Следовательно, $\bigcup_{(r_1,r_2)\in M^2} O_{r_1,r_2} \ne \varnothing$. Если $\bigcap_{(r_1,r_2)\in M^2} O_{r_1,r_2} \ne \varnothing$, то для всех $(u_1,u_2) \in \bigcap_{(r_1,r_2)\in M^2} O_{r_1,r_2}$,

$$J_{n,k,\beta}(z_1,z_2,u_1,u_2) = \sum_{r_1,r_2} h(r_1,r_2)\mathsf{P}(\xi=(r_1,r_2)|V=(u_1,u_2)). \quad (95)$$

Определим множество $\widetilde{O} = (\bigcup_{(r_1,r_2)\in M} O_{r_1,r_2} \times M^2) \setminus O$. Оно также может быть представлено как

$$\widetilde{O} = \bigcup_{(r_1,r_2)\in M^2} \widetilde{O}_{r_1,r_2} \times \{(r_1,r_2)\}, \;\; \widetilde{O}_{r_1,r_2} = \bigcup_{(s_1,s_2)\in M} O_{s_1,s_2} \setminus O_{r_1,r_2}.$$

Очевидно, что равенство $\mathsf{P}_{V,\xi}(O) = 1$ влечет выполнение соотношения $\mathsf{P}_V(\bigcup_{(r_1,r_2)\in M^2} O_{r_1,r_2}) = 1$. Действительно,

$$\mathsf{P}\left(V \in \bigcup_{(r_1,r_2)\in M} O_{r_1,r_2}\right) \geq \mathsf{P}\left(\bigcup_{(r_1,r_2)\in M} \{V \in O_{r_1,r_2}, \xi = (r_1,r_2)\}\right)$$
$$= \mathsf{P}((V,\xi) \in O) = 1.$$

Если множество $\widetilde{O}$ пусто, то $O_{r_1,r_2} = O_{s_1,s_2}$ для $r_1 \neq s_1, r_2 \neq s_2$ и $\bigcap_{(r_1,r_2)\in M} O_{r_1,r_2} = \bigcup_{(r_1,r_2)\in M} O_{r_1,r_2} \neq \varnothing$, поэтому равенство выполняется для P_V-почти всех (u_1,u_2).

Рассмотрим случай, когда $\widetilde{O} \neq \varnothing$. Введем

$$K := \{(r_1,r_2) \in M^2 : \widetilde{O}_{r_1,r_2} \neq \varnothing\}.$$

Следовательно, $K \neq \varnothing$. Если $L := M^2 \setminus K \neq \varnothing$, то для всех $(r_1,r_2) \in L$ имеем $O_{r_1,r_2} = \bigcup_{(s_1,s_2)\in M} O_{s_1,s_2}$. Таким образом, мы показали, что

$$\mathsf{P}_V(\cup_{(s_1,s_2)\in M^2} O_{s_1,s_2}) = 1.$$

Значит, $\mathsf{P}_V(O_{r_1,r_2}) = 1$ для всех $(r_1,r_2) \in L$. Введем

$$K_1 := \{(r_1,r_2) \in K : \mathsf{P}_V(\widetilde{O}_{r_1,r_2}) = 0\}$$

и $K_2 := K \setminus K_1$. Если $(r_1,r_2) \in K_1$, то

$$\mathsf{P}_V(O_{r_1,r_2}) = \mathsf{P}_V(\bigcup_{(s_1,s_2)\in M} O_{s_1,s_2} \setminus \widetilde{O}_{r_1,r_2}) = \mathsf{P}_V(\bigcup_{(s_1,s_2)\in M} O_{s_1,s_2}) - \mathsf{P}_V(\widetilde{O}_{r_1,r_2}) = 1.$$

Теперь покажем, что если $(r_1,r_2) \in K_2$, то

$$\mathsf{P}(\xi = (r_1,r_2) | V = (u_1,u_2)) = 0$$

для всех $(u_1,u_2) \in S_{r_1,r_2} \subset \widetilde{O}_{r_1,r_2}$, $S_{r_1,r_2} \in \mathcal{B}(\mathbb{R}^2_+)$, где $\mathsf{P}_V(S_{r_1,r_2}) = \mathsf{P}_V(\widetilde{O}_{r_1,r_2}) > 0$. Если последнее утверждение верно, то (96) выполняется, поскольку мы

приходим к тривиальному отношению $0 = 0$ и, следовательно, мы получаем необходимую формулу для

$$(u_1, u_2) \in A := (\cap_{(r_1,r_2)\in L\cup K_1} O_{r_1,r_2}) \cap (\cap_{(r_1,r_2)\in K_2} (O_{r_1,r_2} \cup S_{r_1,r_2})),$$

где $\mathsf{P}_V(A) = 1$, поскольку $\mathsf{P}_V(O_{r_1,r_2}) = 1$ для $(r_1, r_2) \in L \cup K_1$ и

$$\begin{aligned}\mathsf{P}_V(O_{r_1,r_2} \cup S_{r_1,r_2}) &= \mathsf{P}_V(O_{r_1,r_2}) + \mathsf{P}_V(S_{r_1,r_2}) = \mathsf{P}_V(O_{r_1,r_2}) + \mathsf{P}_V(\widetilde{O}_{r_1,r_2}) \\ &= \mathsf{P}_V(O_{r_1,r_2} \cup \widetilde{O}_{r_1,r_2}) = \mathsf{P}_V(\cup_{(s_1,s_2)\in M} O_{s_1,s_2}) = 1.\end{aligned}$$

Здесь мы учитываем, что $S_{r_1,r_2} \subset \widetilde{O}_{r_1,r_2}$ и $O_{r_1,r_2} \cap \widetilde{O}_{r_1,r_2} = \varnothing$.

Для всех $(r_1, r_2) \in K_2$

$$\begin{aligned}&\mathsf{P}(V \in \widetilde{O}_{r_1,r_2}, \xi = (r_1, r_2)) \\ &\qquad = \int_{\widetilde{O}_{r_1,r_2}} \mathsf{P}(\xi = (r_1, r_2)|V = (u_1, u_2))\, d\mathsf{P}_V(u_1, u_2) = 0, \quad (96)\end{aligned}$$

поскольку $(\widetilde{O}_{r_1,r_2} \times \{r_1, r_2\}) \cap O = \varnothing$ и $\mathsf{P}_{V,\xi}(O) = 1$. Принимая в расчет неравенство $\mathsf{P}(\xi = (r_1, r_2)|V = (u_1, u_2)) \geq 0$ для P_V-почти всех (u_1, u_2), из формулы (96) мы получаем $\mathsf{P}(\xi = (r_1, r_2)|V = (u_1, u_2)) = 0$ для некоторого множества $S_{r_1,r_2} \in \mathcal{B}(\mathbb{R}^2_+)$ такого, что $S_{r_1,r_2} \subset \widetilde{O}_{r_1,r_2}$ и $\mathsf{P}_V(S_{r_1,r_2}) = \mathsf{P}_V(\widetilde{O}_{r_1,r_2})$. Также (95) выполняется для P_V-почти всех (u_1, u_2).

Для $n > N(x_1, x_2)$ и $u_i \in [0, \delta]$ имеем $u_i n^{-\beta} < |x_1 - x_2|/2$, $i = 1, 2$, поэтому лемма 2.8 применима к $\mathsf{P}(\xi = (r_1, r_2)|V = (u_1, u_2))$. Тогда

$$\begin{aligned}&J_{n,k,\beta}(z_1, z_2, u_1, u_2) \\ &= \sum_{r_1,r_2} h(r_1, r_2)\, \mathsf{P}(\xi_{n,k,1} = r_1|V = (u_1, u_2))\mathsf{P}(\xi_{n,k,2} = r_2|V = (u_1, u_2)) = \mathcal{J}_{n,1}\mathcal{J}_{n,2},\end{aligned}$$

где

$$\begin{aligned}\mathcal{J}_{n,i} &:= \mathsf{E}\left(-\log\left(\frac{\xi_{n,k,i} + 1}{k}\right) - H(Y|X)\Big|V = (u_1, u_2)\right) \\ &= \mathsf{E}\left(-\log\left(\frac{\mu_{n,i} + 1}{k}\right) - H(Y|X)\right)\mathsf{P}(Y \neq y\big|\|X - x_i\| = u_i n^{-\beta}) \\ &+\mathsf{E}\left(-\log\left(\frac{\mu_{n,i} + 2}{k}\right) - H(Y|X)\right)\mathsf{P}\left(Y = y\big|\|X - x_i\| = u_i n^{-\beta}\right) \\ &= \mathsf{E}\left(-\log\left(\frac{\mu_{n,i} + 1}{k}\right) + \log f(y_i|x_i)\right)\mathsf{P}\left(Y \neq y\big|\|X - x_i\| = u_i n^{-\beta}\right) \\ &+\mathsf{E}\left(-\log\left(\frac{\mu_{n,i} + 2}{k}\right) + \log f(y_i|x_i)\right)\mathsf{P}\left(Y = y\big|\|X - x_i\| = u_i n^{-\beta}\right) \\ &+(-\log f(y_i|x_i) - H(Y|X)),\end{aligned}$$

$\mu_{n,i} = \mu_n(k, \beta, u_i, x_i, y_i)$, $J_{n,i} = J_{n,i}(k, \beta, u_i, x_i, y_i)$ и $i = 1, 2$.

Согласно (78) и поскольку $\log k - \log(k-1) \to 0$ при $n \to \infty$, мы можем утверждать, что

$$\sup_{(u,x,y)\in G_n} \left| \mathsf{E}\left(\log\left(\frac{\mu_n(k,\beta,u,x,y)+1}{k} \right) \right) - \log f(y|x) \right| \to 0.$$

Заметим, что $G_n \nearrow (0, \delta] \times \{x \in \mathbb{R}^d \colon f_X(x) > 0\} \times M$ при $n \to \infty$. Для рассматриваемой версии плотности f_X и произвольных $x \in \mathbb{R}^d$, для которых $f_X(x) > 0$, можно найти такое $N(x) \in \mathbb{N}$, что $x \in B_{1,n}$ при $n > N(x)$. Рассмотрим $x_i \in \mathbb{R}^d$ такое, что $f_X(x_i) > 0$, $i = 1, 2$. Тогда для $n \geq \max\{N(x_1), N(x_2)\}$ и $i = 1, 2$

$$\sup_{0<u_i\leq\delta} \left| \log\left(\frac{\mu_{n,i}+1}{k} \right) - \log f(y_i|x_i) \right| \leq \sup_{(u,x,y)\in G_n} \left| \log\left(\frac{\mu_n(k,\beta,u,x,y)+1}{k} \right) - \log f(y|x) \right| \to 0,$$

при $n \to \infty$. Аналогичным образом

$$\sup_{0<u_i\leq\delta} \left| \log\left(\frac{\mu_{n,i}+2}{k} \right) - \log f(y_i|x_i) \right| \to 0, \quad n \to \infty, \quad i = 1, 2.$$

Следовательно, для P_X-почти всех $x_i \in \mathbb{R}^d$ и любых $y_i \in M$ $(i = 1, 2)$,

$$\sup_{0<u_i\leq\delta} |\mathcal{J}_{n,i} - (-\log f(y_i|x_i) - H(Y|X))| \to 0, \quad n \to \infty. \tag{97}$$

Положим $F_i := -\log f(y_i|x_i) - H(Y|X)$. Тогда получим

$$\sup_{0<u_1,u_2\leq\delta} |\mathcal{J}_{n,1}\mathcal{J}_{n,2} - F_1F_2| \leq \sup_{0<u_1\leq\delta} |\mathcal{J}_{n,1}| \sup_{0<u_2\leq\delta} |\mathcal{J}_{n,2} - F_2| + |F_2| \sup_{0<u_1\leq\delta} |\mathcal{J}_{n,1} - F_1|.$$

В силу (97) для $n \geq N(x_i, y_i, \delta)$ выполняется неравенство

$$\sup_{0<u_1\leq\delta} |\mathcal{J}_{n,1}| \leq 2|F_1|.$$

Также для всех $y_1, y_2 \in M$ и $\mathsf{P}_X \otimes \mathsf{P}_X$-почти всех $(x_1, x_2) \in \mathbb{R}^d \times \mathbb{R}^d$,

$$\sup_{u_1,u_2<\delta} |\mathcal{J}_{n,1}\mathcal{J}_{n,2} - F_1F_2| \to 0, \quad n \to \infty.$$

Таким образом, (93) доказано.

Шаг 9. Теперь перейдем к завершающей части доказательства. В силу предыдущих шагов получим

$$
\begin{aligned}
&|\mathcal{U}_{n,k,1}(z_1,z_2)-F_1F_2| \\
&\quad\le \Big|\int_0^\delta\int_0^\delta J_{n,k,\beta}(z_1,z_2,u_1,u_2)f_{n,k,\beta}(x_1,x_1,u_1,u_2)\,du_1du_2-F_1F_2\Big| \\
&\qquad+\Big|\int_0^\delta\int_0^\delta J_{n,k,\beta}(z_1,z_2,u_1,u_2)f_{n,k,\beta}(x_1,x_1,u_1,u_2)\,du_1du_2\Big| \\
&\qquad\qquad\times(1-\mathsf{P}(\rho_{n,k,1}n^\beta\le\delta,\rho_{n,k,2}n^\beta\le\delta)) \\
&\quad\le \sup_{0<u_1,u_2\le\delta}|\mathcal{J}_{n,1}\mathcal{J}_{n,2}-F_1F_2|\int_0^\delta\int_0^\delta f_{n,k,\beta}(x_1,x_2,u_1,u_2)\,du_1du_2 \\
&\qquad\qquad+|\log k+H(Y|X)|(1-\mathsf{P}(\rho_{n,k,1}n^\beta\le\delta,\rho_{n,k,2}n^\beta\le\delta)).
\end{aligned}
$$

Более того,

$$
\begin{aligned}
&1-\mathsf{P}(\rho_{n,k,1}n^\beta\le\delta,\rho_{n,k,2}n^\beta\le\delta) \\
&\qquad\le \mathsf{P}\left(\left\{\rho_{n,k,1}>\delta n^{-\beta}\right\}\cup\left\{\rho_{n,k,2}>\delta n^{-\beta}\right\}\right)\le 2\mathsf{P}\left(\rho_{n,k,1}>\delta n^{-\beta}\right).
\end{aligned}
$$

Учитывая (61) и (70) мы видим, что для всех $z_j\in\mathbb{R}^d\times M$, $j=1,2$,

$$
|\log k+H(Y|X)|(1-\mathsf{P}(\rho_{n,k,1}n^\beta\le\delta,\rho_{n,k,2}n^\beta\le\delta))\to 0,\quad n\to\infty.
$$

Следовательно, $\mathcal{U}_{n,k,1}(z_1,z_2)-F_1F_2\to 0$ при $n\to\infty$. Значит, (92) доказано и доказательство теоремы 2.4 завершено. $\square$

2.4.3 Доказательство следствия 2.5

Пусть $X\sim N(a,\Sigma)$, где $a\in\mathbb{R}^d$ и $\Sigma>0$. Нетрудно убедиться, что для $\varepsilon>0$ выполняется неравенство $\mathsf{E}|\log f_X(X)|^\varepsilon<\infty$ (см., например, [27]). Поскольку для всех $x\in\mathbb{R}^d$ и $y\in M$,

$$
\begin{aligned}
f(x,y)&=\mathsf{P}(Y=y|X=x)f_X(x),\\
\mathsf{P}(Y=1|X=x)&=\frac{1}{1+\exp\{-(w,x)-b\}}>0,\\
f_X(x)&=\frac{1}{(2\pi)^{d/2}|\det\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(x-a)^T\Sigma^{-1}(x-a)\right\}>0,
\end{aligned}
$$

то $f(\cdot,y)$ — μ-почти всюду положительна. Мы покажем, что $f(\cdot,y)$ является C_0-стягиваемой для всех $y\in\{0,1\}$. Согласно замечанию 2.2 достаточно

проверить, что эта функция является липшицевой. Имеем

$$\begin{aligned}|f(u,y)-f(v,y)| &= |\mathsf{P}(Y=y|X=u)f_X(u)-\mathsf{P}(Y=y|X=v)f_X(v)|\\ &\le \mathsf{P}(Y=y|X=u)|f_X(u)-f_X(v)|\\ &\quad + f_X(v)|\mathsf{P}(Y=y|X=u)-\mathsf{P}(Y=y|X=v)|\\ &\le |f_X(u)-f_X(v)| + \max_x |f_X(x)|\,|\mathsf{P}(Y=y|X=u)-\mathsf{P}(Y=y|X=v)|.\end{aligned}$$

Заметим, что $\max_x |f_X(x)| < \infty$ и $f_X(\cdot)$ удовлетворяет условию Липшица, поэтому достаточно показать, что функция $\mathsf{P}(Y=y|X=x)$ (как функция переменной x) удовлетворяет условию Липшица с константой C. Для любого $x \in \mathbb{R}^d$ и $j \in \{1,\dots,d\}$,

$$\left|\frac{\partial}{\partial x_j}\mathsf{P}(Y=1|X=x)\right| = \frac{\exp\{-(w,x)-b\}}{(1+\exp\{-(w,x)-b\})^2}|w_j| \le \frac{1}{4}\|w\| < \infty.$$

Таким образом, для $\mathsf{P}(Y=1|X=x)$ желаемое свойство выполняется. Очевидно, что $\mathsf{P}(Y=0|X=x) = 1-\mathsf{P}(Y=1|X=x)$ и, следовательно, $\mathsf{P}(Y=0|X=x)$ также удовлетворяет условию Липшица. □

2.4.4 Доказательство леммы 2.6

Зафиксируем $n>1$ и $k \in \{1,\dots,n-1\}$. Для $t>0$, $x \in \mathbb{R}^d$ и $\delta \in (0,t)$ определим события

$$A_{n,k}(x,t,\delta) := \{t-\delta < \rho_{n,k,1}(x) \le t+\delta\}.$$

Событие $\{\rho_{n,k,1}(x) > t-\delta\}$ означает, что менее чем k точек (т.е. $0,1,\dots,k$–1) из $X^2,\dots,X^n$ попали в шар $B(x,t-\delta)$. Событие же $\{\rho_{n,k}(x) \le t+\delta\}$ означает, что как минимум k точек (т.е. $k,k+1,\dots,n-1$) из $X^2,\dots,X^n$ содержатся в шаре $B(x,t+\delta)$. Для $x \in \mathbb{R}^d$ и $t>0$, рассмотрим множество

$$S_x(t,\delta) := \{z \in \mathbb{R}^d : t-\delta < \|z-x\| \le t+\delta\}.$$

Пусть $P_x(t,\delta) := \mathsf{P}(X \in S_x(t,\delta))$. Заметим, что $P_x(t,\delta) \to 0$ при $\delta \to 0+$ поскольку $\mu(S_x(t,\delta)) \to 0$ при $\delta \to 0+$ и существует плотность P_X относительно меры Лебега μ. Для $2 \le i_1 < \ldots < i_q \le n$, учитывая независимость $X^1,\dots,X^n$, получаем

$$\mathsf{P}(X^{i_1} \in S_x(t,\delta),\dots,X^{i_q} \in S_x(t,\delta)) = P_x(t,\delta)^q.$$

Отметим, что

$$A_{n,k}(x,t,\delta) = \bigcup_{(s,m)\in J} B_s D_m G_{n-1-(s+m)},$$

здесь $J := \{(s,m) : s \in \{0,\dots,k-1\}, m \in \{1,\dots,n-1\}, k \le s+m \le n-1\}$, $J = J(n,k)$,

$$B_s := \{s \text{ наблюдений из } X^2,\dots,X^n \text{ содержатся в } B(x,t-\delta)\},$$
$$D_m := \{m \text{ наблюдений из } X^2,\dots,X^n \text{ содержатся в } S(t,\delta)\},$$
$$G_{n-1-(s+m)} := \{n-1-(s+m) \text{ наблюдений из } X^2,\dots,X^n \text{ содержатся в } \mathbb{R}^d \setminus B(x,t+\delta)\}.$$

Очевидно, что события B_s, D_m и $G_{n-1-(s+m)}$ зависят от x,t,δ,n. Получим

$$\mathsf{P}(A_{n,k}(x,t,\delta)) = \mathsf{P}(B_{k-1}D_1G_{n-1-k}) + O(P_x(t,\delta)^2), \quad \delta \to 0+, \tag{98}$$

при $\mathsf{P}(D_m) = \binom{n-1}{m} P_x(t,\delta)^m (1-P_x(t,\delta))^{n-1-m}$ и $\mathsf{P}(B_sD_mG_{n-1-(s+m)}) \le \mathsf{P}(D_m)$ для $(m,s) \in J$. Определим $p_x(u) = \mathsf{P}(X \in B(x,u))$, $x \in \mathbb{R}^d$, $u > 0$. Тогда

$$\mathsf{P}(B_{k-1}D_1G_{n-1-k}) = \binom{n-1}{k-1}(p_x(t-\delta))^{k-1}\binom{n-k}{1}P_x(t,\delta)(1-p_x(t+\delta))^{n-1-k}.$$

Действительно, существует $\binom{n-1}{k-1}$ способов выбрать $k-1$ точку из $X^2,\dots,X^n$, содержащуюся в $B(x,t-\delta)$, а затем $n-1-(k-1)$ способов выбрать одну точку, попадающую в $S_x(t,\delta)$. Другие точки (их всего $n-1-k$) будут лежать вне шара $B(x,t+\delta)$.

Теперь заметим, что $r = 0,1,\dots,k-1$,

$$\begin{aligned}\mathsf{P}(\xi_{n,k,1}(x,y) = r, A_{n,k}(x,t,\delta) \\ = \mathsf{P}(\xi_{n,k,1}(x,y) = r, B_{k-1}D_1G_{n-1-k}) + O(P_x(t,\delta)^2), \quad \delta \to 0+,\end{aligned}$$

и

$$\begin{aligned}&\mathsf{P}(\xi_{n,k,1}(x,y) = r, B_{k-1}D_1G_{n-1-k}) \\ &= \binom{n-1}{k-1}\binom{k-1}{r}\mathsf{P}(Y=y, X \in B(x,t-\delta))^r \mathsf{P}(Y \ne y, X \in B(x,t-\delta))^{k-1-r} \\ &\qquad \times \binom{n-k}{1} P(Y \ne y, X \in S_x(t,\delta))(1-p_x(t+\delta))^{n-1-k} \\ &\qquad + \binom{n-1}{k-1}\binom{k-1}{r-1}\mathsf{P}(Y=y, X \in B(x,t-\delta))^{r-1} \\ &\qquad\qquad \times \mathsf{P}(Y \ne y, X \in B(x,t-\delta))^{k-1-(r-1)} \\ &\qquad \times \binom{n-k}{1}\mathsf{P}(Y=y, X \in S_x(t,\delta))(1-p_x(t+\delta))^{n-1-k}. \end{aligned} \tag{99}$$

Имеется $\binom{n-1}{k-1}$ способов выбрать $k-1$ точку из $X^2,\dots,X^n$, которая будет лежать в $B(x,t-\delta)$, и существует $\binom{n-k}{1}$ способов выбрать из оставшихся наблюдений точку X^q, принадлежащую $S_x(t,\delta)$. Далее рассмотрим две возможности.

1. Если $Y^q \neq y$, то существует $\binom{k-1}{r}$ способов выбрать из точек, содержащихся в $B(x,t-\delta)$, r точек $X^{i_1},\dots,X^{i_r}$ таких, что $Y^{i_m}=y$, $m=1,\dots,r$. Для остальных $k-1-r$ точек $X^{j_1},\dots,X^{j_{k-1-r}}$, принадлежащих $B(x,t-\delta)$, будет выполняться $Y^{j_s}\neq y$, $s=1,\dots,k-1-r$.

2. Если $Y^q = y$, то существует $\binom{k-1}{r-1}$ способов выбрать из точек, содержащихся в $B(x,t-\delta)$, $r-1$ точек $X^{i_1},\dots,X^{i_{r-1}}$ таких, что $Y^{i_m}=y$, $m=1,\dots,r-1$. Для остальных $k-1-(r-1)$ точек $X^{j_1},\dots,X^{j_{k-1-(r-1)}}$, принадлежащих $B(x,t-\delta)$, будет выполняться $Y^{j_s}\neq y$, $s=1,\dots,k-1-(r-1)$.

Оставшиеся $n-1-k$ точек будут принадлежать дополнению шара $B(x,t+\delta)$. Вероятность для каждого наблюдения X^m попасть в дополнение будет равна $1-p_x(t+\delta)$.

Для $r=k$ получаем, что

$$\begin{aligned}\mathsf{P}(\xi_{n,k,1}(x,y)=k, B_{k-1}D_1G_{n-1-k}) \\ = \binom{n-1}{k-1}\binom{n-k}{1}\mathsf{P}(Y=y, X\in B(x,t-\delta))^{k-1} \\ \times \mathsf{P}(Y=y, X\in S_x(t,\delta))(1-p_x(t+\delta))^{n-1-k}.\end{aligned}$$

В этом случае рассуждение проводится аналогично, но разница заключается в следующем. Не только для всех $(k-1)$ точек $X^{i_1},\dots,X^{i_{k-1}}$ (из $X^2,\dots,X^n$, принадлежащих $B(x,t-\delta)$) верно равенство $Y^{i_1}=y,\dots,Y^{i_{k-1}}=y$, но также и для X^q, содержащихся в $S_x(t,\delta)$, выполняется $Y^q=y$. Случай $r=k$ вытекает из формулы (99), поскольку $\binom{k-1}{k}=0$.

Если случайная переменная τ такова, что $\tau\geq 0$ п.н., $\mathsf{E}\tau<\infty$ и случайный вектор ζ принимает значения в $\mathbb{R}^s$, то (см., например, [7], том II, глава 7.5) функция $\mathsf{E}(\tau|\zeta=t)$ может быть определена следующим образом. Положим $G(B):=\mathsf{E}(\tau\mathbb{I}\{\zeta\in B\})$, где $B\in\mathcal{B}(\mathbb{R}^s)$. Очевидно, что G является конечной мерой, абсолютно непрерывной относительно P_ζ. Следовательно, существует такая борелевская функция $\varphi:\mathbb{R}^s\to\mathbb{R}$, что для всех $B\in\mathcal{B}(\mathbb{R}^s)$,

$$\mathsf{E}(\tau\mathbb{I}\{\zeta\in B\})=\int_B\varphi(x)\mathsf{P}_\zeta(dx).$$

Иными словами, $\varphi(t)$ — производная Радона-Никодима $\frac{dG}{d\mathsf{P}_\zeta}(t)$, $t \in \mathbb{R}^s$. Значит, $\mathsf{E}(\tau|\zeta) = \varphi(\zeta)$. Согласно теореме 5.8.8 [2] (с учетом того, что $G \ll \mathsf{P}_\zeta$) существует

$$\lim_{\delta \to 0+} \frac{G(B(t,\delta))}{\mathsf{P}_\zeta(B(t,\delta))} = \frac{dG}{d\mathsf{P}_\zeta}(t), \quad t \in \mathbb{R}^s. \tag{100}$$

Точнее, этот предел существует для P_ζ-почти всех $t \in \mathbb{R}^s$ и является производной Радона-Никодима G относительно меры P_ζ, т.е. версией $\mathsf{E}(\tau|\zeta = t)$. Мы применим этот результат к $\tau = \mathbb{I}\{\xi_{n,k,1}(x,y) \in D\}$, где $D \in \mathcal{B}(\mathbb{R}_+)$, $\zeta = \rho_{n,k,1}(x)$, $x \in \mathbb{R}^d$, $y \in M$. Очевидно, что τ является интегрируемой случайной величиной относительно любой конечной меры. Формула (100) верна для $\mathsf{P}_{\rho_{n,k,1}(x)}$-почти всех $t \in (0,\infty)$ и

$$\mathsf{P}(\xi_{n,k,1}(x,y) \in D | \rho_{n,k,1}(x) = t) = \lim_{\delta \to 0+} \frac{\mathsf{P}(\xi_{n,k,1}(x,y) \in D, \rho_{n,k,1}(x) \in B(t,\delta))}{\mathsf{P}(\rho_{n,k,1}(x) \in B(t,\delta))}. \tag{101}$$

Отметим, что вместо $B(t,\delta) = [t-\delta, t+\delta]$, где $0 < \delta < t$, мы можем взять $(t-\delta, t+\delta]$, поскольку для всех $n \in \mathbb{N}$, $n > 1$, $k \in \{0,\ldots,n-1\}$ и $x \in \mathbb{R}^d$ случайная величина $\rho_{n,k,1}(x)$ имеет плотность. Действительно, $\mathsf{P}(\rho_{n,k,1}(x) \le 0) = 0$, поскольку существует плотность $f_X(\cdot)$, а для $t > 0$

$$\mathsf{P}(\rho_{n,k,1}(x) \le t) = \sum_{j=k}^{n-1} \binom{n-1}{j} p_x(t)^j (1-p_x(t))^{n-1-j}, \tag{102}$$

где $p_x(t) = \mathsf{P}(X \in B(x,t))$. Понятно, что $p_x(t) = \mathsf{P}(\|X-x\| \le t)$ является функцией распределения неотрицательной случайной величины $\|X-x\|$. Вначале покажем, что $p_x(\cdot)$ имеет плотность $f_x(\cdot)$ относительно меры Лебега на $\mathcal{B}(\mathbb{R}_+)$. Затем докажем, что существует плотность случайной величины $\rho_{n,k,1}(x)$.

Известно, что $X = (X^1,\ldots,X^d)$ имеет плотность $f_X(\cdot)$ относительно меры Лебега μ на $\mathbb{R}^d$ (i.e. $\mathsf{P}_X \ll \mu$). С другой стороны, поскольку $f_X(x)$ строго положительна для μ-почти всех $x \in \mathbb{R}^d$, можно легко показать, что $\mu \ll \mathsf{P}_X$. Следовательно, $\mathsf{P}_X \sim \mu$.

Пусть μ_1 и μ_2 — некоторые меры на пространстве $(S,\mathcal{B})$ и $h : S \to T$ — некоторая $\mathcal{B}|\mathcal{D}$-измеримая функция, где T снабжено σ-алгеброй $\mathcal{D}$. Введем меры $\nu_i := \mu_i h^{-1}$, $i = 1,2$. Тогда, очевидно, $\mu_1 \ll \mu_2$ влечет $\nu_1 \ll \nu_2$. Если Q является гауссовской мерой на $\mathcal{B}(\mathbb{R}^d)$, имеющей плотность относительно меры μ, то $Q \sim \mu$, поскольку существует строго положительная плотность $dQ/d\mu$ на $\mathbb{R}^d$. Рассмотрим $(S,\mathcal{B}) := (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, $(T,\mathcal{D}) := (\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+))$, $h : \mathbb{R}^d \to \mathbb{R}_+$, где $h(x) = x_1^2 + \ldots + x_d^2$ для $x = (x_1,\ldots,x_d) \in \mathbb{R}^d$. Пусть $\mu_1 = \mathsf{P}_X$ и μ_2 является гауссовской мерой $N(0,I)$ на $\mathbb{R}^d$ с нулевым средним

и единичной матрицей ковариаций I. Тогда $\mu_1 \sim \mu_2$, поскольку $\mu_1 \sim \mu$ и $\mu_2 \sim \mu$. Следовательно, $\nu_1 \sim \nu_2$. Мера $\nu_2 = \mu_2 h^{-1}$ имеет распределение χ^2_d и ее плотность относительно меры Лебега $\Lambda_{\mathbb{R}_+}$ на $(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+))$ выражается как

$$\mathsf{P}_{\chi^2_d}(u) = \frac{u^{\frac{d}{2}-1}e^{-\frac{u}{2}}}{2^{\frac{d}{2}}\Gamma(\frac{d}{2})}, \quad u \geq 0.$$

Эта плотность строго положительна на $(0, \infty)$, а значит, $\nu_2 \sim \Lambda_{\mathbb{R}_+}$. Таким образом, $\mathsf{P}_X h^{-1} \sim \nu_2$, следовательно $\mathsf{P}_X h^{-1} \sim \Lambda_{\mathbb{R}_+}$. Мы доказали, что существует плотность g случайной величины $(X_1)^2 + \ldots + (X_d)^2$ относительно меры Лебега на $\mathbb{R}_+$. Можем написать, что $d(\mathsf{P}_X h^{-1})/d\,\Lambda_{\mathbb{R}_+} = g$. Для всех $B \in \mathcal{B}(\mathbb{R}_+)$ верно

$$\int_B g(t)dt = \mathsf{P}_X h^{-1}(B).$$

Если $\int_B g(t)dt = 0$, то $\mathsf{P}_X h^{-1}(B) = 0$ и следовательно $\Lambda_{\mathbb{R}_+}(B) = 0$. Положим $B := \{t \in \mathbb{R}_+ : g(t) = 0\}$. Тогда $\int_B g(t)dt = 0$ и, следовательно, $\Lambda_{\mathbb{R}_+}\{t : g(t) = 0\} = 0$. Другими словами, g строго положительна $\Lambda_{\mathbb{R}_+}$-почти всюду.

Если случайный вектор V имеет плотность (относительно меры μ) $q(z)$, $z \in \mathbb{R}^d$, тогда для $x \in \mathbb{R}^d$ вектор $V - x$ имеет плотность

$$q_x(z) = q(z + x), \quad z \in \mathbb{R}^d.$$

Следовательно, мы можем потребовать, чтобы для всех $x \in \mathbb{R}^d$ существовала плотность случайной величины $\|X - x\|^2$ относительно меры Лебега $\Lambda_{\mathbb{R}_+}$ на $\mathbb{R}_+$. Эта плотность строго положительна относительно меры $\Lambda_{\mathbb{R}_+}$, где $f_X(\cdot)$ строго положительна относительна μ. Если случайная величина $\xi_x \geq 0$ имеет плотность $\gamma_x(u)$, $u \geq 0$ ($x \in \mathbb{R}^d$), то случайная величина $\sqrt{\xi_x}$ имеет плотность $p_x(u) = 2u\gamma_x(u^2)$, $u \geq 0$. Таким образом, существует плотность $f_x(u)$, $u \geq 0$, случайной величины $\|X - x\|$ и эта плотность строго положительна для $\Lambda_{\mathbb{R}_+}$-почти всех $u \geq 0$ и $\mathsf{P}_{\|X-x\|} \sim \Lambda_{\mathbb{R}_+}$.

Теперь мы можем доказать, что плотность (относительно $\Lambda_{\mathbb{R}_+}$) случайной величины $\rho_{n,k,1}(x)$ имеет вид

$$\begin{aligned} &h_{n-1,k,x}(u) \\ =&\sum_{j=k}^{n-1} \binom{n-1}{j} (jp_x(u)^{j-1}(1-p_x(u))^{n-1-j} - p_x(u)^j(n-j-1)(1-p_x(u))^{n-j-2})f_x(u) \end{aligned} \tag{103}$$

где $f_x(\cdot)$ — плотность, соответствующая функции распределения функции $p_x(\cdot)$, $x \in \mathbb{R}^d$.

Стоит подчеркнуть, что в общем случае мы не можем продифференцировать функцию распределения, чтобы получить плотность (примером может служить функция Кантора), поэтому надо использовать интегральные соотношения. Пусть F — функция распределения положительной случайной величины с плотностью f (таким образом, $F(0) = 0$ и $f(u) = 0$, $u < 0$). Тогда используя интегрирование по частям (см., например, [7], том II, глава 6.12) и метод математической индукции, мы можем доказать, что для всех $n \in \mathbb{N}$ функция распределения $F^n(u)$ имеет плотность $nF^{n-1}(u)f(u)$, $u \in \mathbb{R}$. Для $m \in \mathbb{N}$ и $j = 0, \ldots, m$ мы можем написать, что

$$F(t)^j(1-F(t))^{m-j} = \sum_{r=0}^{m-j} \binom{m-j}{r} (-1)^{m-j-r} F(t)^{j+m-j-r}$$
$$= \sum_{r=0}^{m-j} \binom{m-j}{r} (-1)^{m-j-r} \int_{(0,t]} (m-r) F^{m-r-1}(u) f(u) du.$$

Последняя формула и (102) для $t \geq 0$ приводит к соотношению

$$\mathsf{P}(\rho_{n,k,1}(x) \leq t) = \int_0^t h_{n-1,k,x}(u) du, \tag{104}$$

где $h_{n-1,k,1}$ задается формулой (103). Положим $h_{n-1,k,1}(u) = 0$ для $u < 0$. Отметим, что мы приходим к (103), используя полином степени $n-1$ от функции распределения $p_x(\cdot)$. Следовательно, $h_{n-1,k,x}$ является интегрируемой функцией (относительно меры Лебега Λ на $\mathbb{R}$). Однако упомянутый полином имеет положительные и отрицательные коэффициенты. Следовательно, необходимо пояснить, почему $h_{n-1,k,x}$ является вероятностной плотностью. Покажем, что если для функции распределения F верно представление

$$F(t) = \int_{(-\infty,t]} f(u) du, \quad t \in \mathbb{R},$$

где $f \in L^1(\mathbb{R}, \mathcal{B}(\mathbb{R}), \Lambda)$, то f является вероятностной плотностью. Теорема Лебега о мажорируемой сходимости позволяет получить равенство $\int_{\mathbb{R}} f(u) du = 1$, поскольку $\lim_{t\to\infty} F(t) = 1$. Остается показать, что $f(u) \geq 0$ для Λ-почти всех u. Введем функцию

$$Q(B) := \int_B f(u) du, \quad B \in \mathcal{B}(\mathbb{R}).$$

Очевидно, $Q((a,b]) = F(b) - F(a) \geq 0$ для $-\infty \leq a \leq b \leq \infty$ (положим $F(-\infty) := 0$, $F(\infty) := 1$ и $(a, \infty] := (a, \infty)$). Пусть G является вероятностной мерой на $\mathcal{B}(\mathbb{R})$, порожденной функцией распределения F, тогда

$G((a,b]) := F(b) - F(a)$. Меры G и Q совпадают на алгебре $\mathcal{A}$, состоящей из конечных объединений попарно непересекающихся интервалов вида $(a,b]$, $-\infty \le a \le b \le \infty$. Следовательно, Q является конечной неотрицательной функцией на $\mathcal{A}$. Также Q является счетно-аддитивной функцией на $\mathcal{B}(\mathbb{R})$ и $Q(\mathbb{R})$ конечно. Остается заметить, что для любых $B \in \mathcal{B}(\mathbb{R})$ и всех $\varepsilon > 0$ существует $A \in \mathcal{A}$ такое, что $|Q(B) - Q(A)| < \varepsilon$. Действительно,

$$|Q(B) - Q(A)| = \left| \int_{\mathbb{R}} (\mathbb{I}\{B\} - \mathbb{I}\{A\}) f(u) du \right| \le \int_{\mathbb{R}} \mathbb{I}\{B \triangle A\} |f(u)| du.$$

Следовательно, для всех $B \in \mathcal{B}(\mathbb{R})$ можно найти такие $A_n \in \mathcal{A}$ ($n \in \mathbb{N}$), что $Q(A_n) \to Q(B)$ при $n \to \infty$. Учитывая тот факт, что $Q(A_n) \ge 0$, мы получим $Q(B) \ge 0$. Теперь предположим, что $\mu(E) > 0$, где $E = \{x : f(x) < 0\}$. Отметим, что $E = \cup_{m=1}^{\infty} \{-\infty < f(x) \le -\frac{1}{m}\}$. Тогда стандартными рассуждениями мы придем к противоречию. Значит, $\mu(E) = 0$. Таким образом, формула (103) дает вероятностную плотность $h_{n-1,k,x}(\cdot)$ случайной величины $\rho_{n,k,1}(x)$, где $x \in \mathbb{R}^d$.

Следовательно, для всех $x \in \mathbb{R}^d$, $y \in M$, $r \in \{0, 1, \ldots, k\}$ и $\mathsf{P}_{\rho_{n,k,1}(x)}$-почти всех $t \in (0, \infty)$ в силу (101) получим

$$\mathsf{P}(\xi_{n,k,1}(x,y) = r | \rho_{n,k,1}(x) = t) = \lim_{\delta \to 0+} \frac{\mathsf{P}(\xi_{n,k,1}(x,y) = r, A_{n,k}(x,t,\delta))}{\mathsf{P}(A_{n,k}(x,t,\delta))}. \quad (105)$$

Применяя формулы, полученные для числителя и знаменателя последней дроби в (105), и учитывая тот факт, что функция $\mathsf{P}(X \in B(x,t))$ непрерывна по $(x,t) \in \mathbb{R}^d \times \mathbb{R}_+$ (см, например, лемму 2.1 в [27]), мы получим для всех $n \in \mathbb{N}$ ($n > 1$), $k \in \{1, \ldots, n-1\}$, $r = 0, \ldots, k$ и $\mathsf{P}_{\rho_{n,k,1}(x)}$-почти всех $t \in (0, \infty)$,

$$\begin{aligned} &\mathsf{P}(\xi_{n,k,1}(x,y) = r | \rho_{n,k,1}(x) = t) \\ &\quad = \binom{k-1}{r} p^r (1-p)^{k-1-r} \lim_{\delta \to 0+} \frac{\mathsf{P}(Y \ne y, X \in S_x(t,\delta))}{\mathsf{P}(X \in S_x(t,\delta))} \\ &\qquad + \binom{k-1}{r-1} p^{r-1} (1-p)^{k-r} \lim_{\delta \to 0+} \frac{\mathsf{P}(Y = y, X \in S_x(t,\delta))}{\mathsf{P}(X \in S_x(t,\delta))}, \quad (106) \end{aligned}$$

где $p := \mathsf{P}(Y = y | X \in B(x,t))$, $p = p(x,y,t)$. Ранее мы показали, что $\mathsf{P}(X \in B(x,t)) > 0$ для μ-почти всех $x \in \mathbb{R}^d$ и $t > 0$, поскольку $f_X(z)$ является строго положительной для μ-почти всех $z \in \mathbb{R}^d$. Однако необходимо объяснить существование предела в формуле (106). Применим формулу (100) для $\zeta := \|X - x\|$, $G(D) := \mathsf{E}(\mathbb{I}\{\zeta \in D\} \mathbb{I}\{Y = y\})$, $x \in \mathbb{R}^d$, $y \in M$ и $D \in \mathcal{B}(\mathbb{R})$. Мы можем потребовать, чтобы для всех $x \in \mathbb{R}^d$ и $y \in M$

пределы, фигурирующие в формуле (106) существовали для $\mathsf{P}_{\|X-x\|}$-почти всех $t > 0$. Действительно,

$$\lim_{\delta\to 0+} \frac{\mathsf{P}(Y \neq y, X \in S_x(t,\delta))}{\mathsf{P}(X \in S_x(t,\delta))} = \lim_{\delta\to 0+} \frac{\mathsf{P}(Y \neq y, t-\delta < \|X-x\| \leq t+\delta)}{\mathsf{P}(t-\delta < \|X-x\| \leq t+\delta)}$$
$$= \mathsf{P}(Y \neq y | \|X-x\| = t) := \alpha(x,y,t). \quad (107)$$

Мы показали, что $\mathsf{P}_{\|X-x\|} \sim \Lambda_{\mathbb{R}_+}$ для всех $x \in \mathbb{R}^d$. Следовательно, $\mathsf{P}(X \in S_x(t,\delta)) > 0$ для всех $x \in \mathbb{R}^d$ и $\delta > 0$. Кроме того, для всех y, принадлежащих конечному множеству M, и всех $x \in \mathbb{R}^d$, пределы в (106) существуют для $\Lambda_{\mathbb{R}_+}$-почти всех $t \in (0,\infty)$, поскольку $\mathsf{P}_{\|X-x\|} \sim \Lambda_{\mathbb{R}_+}$. Следовательно, мера $\mathsf{P}_{\rho_{n,k,1}(x)}$ множества точек $t \in \mathbb{R}_+$ таких, что пределы в (106) не существуют, равна нулю, поскольку $\mathsf{P}_{\rho_{n,k,1}(x)} \ll \Lambda_{\mathbb{R}_+}$. Доказательство окончено. □

2.4.5 Доказательство леммы 2.8

Зафиксируем произвольное $z_j = (x_j, y_j) \in \mathbb{R}^d \times M$, $j \in \{1,2\}$ такое, что $x_1 \neq x_2$. Предположим, что $n \in \mathbb{N}$, $n > 2$, и $k \in \{1,\dots,n-1\}$. Заметим, что

$$\xi_{n,k,1}(z_1,z_2) = \sharp\{i : i \in \{3,\dots,n\}, Y^i = y_1, X^i \in B(x_1,R_1)\}$$
$$+ \mathbb{I}\{y_2 = y_1, x_2 \in B(x_1,R_1)\},$$

$$\xi_{n,k,2}(z_1,z_2) = \sharp\{i : i \in \{3,\dots,n\}, Y^i = y_2, X^i \in B(x_2,R_2)\}$$
$$+ \mathbb{I}\{y_1 = y_2, x_1 \in B(x_2,R_2)\},$$

где $R_j := \rho_{n,k,j}(x_1,x_2)$, $j = 1,2$. Напомним, что $\rho_{n,k,j}(x_1,x_2)$ и $\xi_{n,k,j}(z_1,z_2)$ определяются формулами (50) и (51), соответственно. Возьмем произвольное $\varepsilon \in (0, |x_1-x_2|/2)$ и $t_j \in (0, |x_1-x_2|/2-\varepsilon)$, $j = 1,2$. Тогда существует такое $\delta > 0$, что $\delta < t_j$, $j = 1,2$, и $t_1 + t_2 + 2\delta < |x_1-x_2|$. Введем события

$$A_{n,k} := \bigcap_{j=1}^{2} \{t_j - \delta < \rho_{n,k,j}(x_1,x_2) \leq t_j + \delta\}.$$

Очевидно, что $A_{n,k} = A_{n,k}(x_1,x_2,t_1,t_2,\delta)$. Далее в доказательстве мы будем рассматривать $j \in \{1,2\}$ без отдельного упоминания. Для простоты изложения мы будем использовать те же обозначения, что и при доказательстве леммы 2.6. Однако необходимо подчеркнуть тот факт, что теперь

мы используем векторы, состоящие из двух компонент, вместо случайных величин, фигурирующих в лемме 2.6. Например, событие $A_{n,k}$ не то же самое, что раньше. Событие $\{\rho_{n,k,j}(x_1,x_2) > t_j - \delta\}$ означает, что замкнутый шар $B(x_j, t_j - \delta)$ содержит менее k (т.е. $0, 1, \ldots, k-1$) наблюдений из $X^3, \ldots, X^n$, поскольку x_i, $i \in \{1,2\} \setminus \{j\}$, не принадлежат этому шару, так как $|x_1 - x_2| > t_j - \delta$. Событие $\{\rho_{n,k,j}(x_1,x_2) \le t_i + \delta\}$ означает, что в шаре $B(x_j, t_j + \delta)$ содержится k (т.е. $k, k+1, \ldots, n-2$) точек из $X^3, \ldots, X^n$, поскольку x_i, $i \in \{1,2\} \setminus \{j\}$, не принадлежат этому шару, т.к. $|x_1 - x_2| > t_j + \delta$. Имеем

$$A_{n,k} = \bigcup_{(s_1,s_2,m_1,m_2)\in J_{n,k}} B_{1,s_1} D_{1,m_1} B_{2,s_2} D_{2,m_2} G_{n-2-(s_1+s_2+m_1+m_2)},$$

где $J_{n,k}$ состоит из таких точек (s_1, s_2, m_1, m_2), что

$$s_1, s_2 \in \{0, \ldots, k-1\},$$

$$m_1, m_2 \in \{1, \ldots, n-1\},$$

$$s_1 + m_1 \ge k, s_2 + m_2 \ge k, s_1 + s_2 + m_1 + m_2 \le n-2$$

и

$$B_{j,s} := \{s \text{ величин из } X^3, \ldots, X^n \text{ принадлежат } B(x_j, t_j - \delta)\},$$

$$D_{j,m} := \{m \text{ величин из } X^3, \ldots, X^n \text{ принадлежат } S_{x_j}(t_j, \delta)\},$$

$$G_l := \left\{ l \text{ величин из } X^3, \ldots, X^n \text{ принадлежат } \mathbb{R}^d \setminus \bigcup_{j=1}^{2} B(x_j, t_j + \delta) \right\}.$$

Поскольку $k \le [(n-2)/2]$, то множество $J_{n,k}$ непусто, так как $(k-1, k-1, 1, 1) \in J_{n,k}$. Здесь и далее мы упрощаем запись, поскольку, на самом деле, $B_{j,s} = B_{j,s}(x_j, t_j, \delta, n)$, $D_{j,m} = D_{j,m}(x_j, t_j, \delta, n)$ и $G_l = G_l(x_1, x_2, t_1, t_2, \delta)$. Аналогично (109) получаем

$$\mathsf{P}(A_{n,k}) = \mathsf{P}(B_{1,k-1} D_{1,1} B_{2,k-1} D_{2,1} G_{n-2-2k}) + O(P_{x_1}(t_1,\delta)^2 P_{x_2}(t_2,\delta)) + O(P_{x_1}(t_1,\delta) P_{x_2}(t_2,\delta)^2)$$

при $\delta \to 0+$, поскольку для $(s_1, s_2, m_1, m_2) \in J_{n,k}$

$$\mathsf{P}(D_{1,m_1} D_{2,m_2}) = \binom{n-2}{m_1} P_{x_1}(t_1,\delta)^{m_1} \binom{n-2-m_1}{m_2} P_{x_2}(t_2,\delta)^{m_2} \times (1 - P_{x_1}(t_1,\delta) - P_{x_2}(t_2,\delta))^{n-2-m_1-m_2},$$

$$\mathsf{P}(B_{1,s_1}D_{1,m_1}B_{2,s_2}D_{2,m_2}G_{n-2-(s_1+s_2+m_1+m_2)}) \le \mathsf{P}(D_{1,m_1}D_{2,m_2}),$$

$P_x(t,\delta) = \mathsf{P}(X \in S_x(t,\delta))$, $x \in \mathbb{R}^d$, $t > 0$ и $\delta > 0$. Нетрудно убедиться, что

$$\begin{aligned}\mathsf{P}(B_{1,k-1}D_{1,1}B_{2,k-1}D_{2,1}G_{n-2-2k}) &= Poly(k-1,1,k-1,1,n-2-2k)\\ &\times (p_{x_1}(t_1-\delta))^{k-1}P_{x_1}(t_1,\delta)(p_{x_2}(t_2-\delta))^{k-1}P_{x_2}(t_2,\delta)\\ &\times (1-p_{x_1}(t_1+\delta)-p_{x_2}(t_2+\delta))^{n-2-2k},\end{aligned}$$

где $p_x(t) = \mathsf{P}(X \in B(x,t))$, $x \in \mathbb{R}^d$, $t > 0$, и

$$Poly(k_1,\dots,k_q) := \frac{(k_1+\ldots+k_q)!}{k_1!\ldots k_q!}, \quad k_i \in \mathbb{Z}_+,\ i = 1,\dots,q.$$

Таким образом, для $r_1, r_2 \in \{0,\dots,k\}$,

$$\begin{aligned}&\mathsf{P}(\xi_{n,k,1}(z_1,z_2) = r_1, \xi_{n,k,2}(z_1,z_2) = r_2, A_{n,k})\\ &\quad = \mathsf{P}(\xi_{n,k,1}(z_1,z_2) = r_1, \xi_{n,k,2}(z_1,z_2) = r_2, B_{1,k-1}D_{1,1}B_{2,k-1}D_{2,1}G_{n-2-2k})\\ &\qquad + O(P_{x_1}(t_1,\delta)^2P_{x_2}(t_2,\delta)) + O(P_{x_1}(t_1,\delta)P_{x_2}(t_2,\delta)^2), \quad \delta \to 0+.\end{aligned}$$

Введем вспомогательные события. Пусть $B^l_{j,s}$ означает, что только s наблюдений из $X^3,\dots,X^n$ содержатся в $B(x_j, t_j-\delta)$, более того, l точек из X^i, содержащихся в шаре, т.е. $X^{i_1},\dots,X^{i_l}$ таковы, что $Y^{i_m} = y_j$, $m = 1,\dots,l$. Ясно, что $B^l_{j,s} = B^l_{j,s}(x_j,y_j,t_j,\delta,n)$. Аналогично, можно определить событие $D^l_{j,s}$ (s точек из $X^3,\dots,X^n$ содержатся в $S_{x_j}(t_j,\delta)$, а другие не принадлжат этому множеству, более того, l точек из X^i, содержащихся в шаре $S_{x_j}(t_j,\delta)$ таковы, что $Y^i = y_j$). Отметим, что $D^l_{j,m} = D^l_{j,m}(x_j,y_j,t_j,\delta,n)$. Тогда для $r_1, r_2 \in \{0,1,\dots,k\}$,

$$\begin{aligned}&\{\xi_{n,k,1}(z_1,z_2) = r_1, \xi_{n,k,2}(z_1,z_2) = r_2\} \cap B_{1,k-1}D_{1,1}B_{2,k-1}D_{2,1}G_{n-2-2k}\\ &\quad = B^{r_1}_{1,k-1}D^0_{1,1}B^{r_2}_{2,k-1}D^0_{2,1}G_{n-2-2k} \cup B^{r_1-1}_{1,k-1}D^1_{1,1}B^{r_2}_{2,k-1}D^0_{2,1}G_{n-2-2k}\\ &\quad \cup B^{r_1}_{1,k-1}D^0_{1,1}B^{r_2-1}_{2,k-1}D^1_{2,1}G_{n-2-2k} \cup B^{r_1-1}_{1,k-1}D^1_{1,1}B^{r_2-1}_{2,k-1}D^1_{2,1}G_{n-2-2k}. \quad (108)\end{aligned}$$

Если $r = 0$, то $B^{r-1}_{j,k-1} := \varnothing$ (для $j \in \{1,2\}$). Понятно, что четыре события, фигурирующих в объединении в формуле (108), попарно несовместны.

Оценим их вероятности. Имеем

$$\begin{aligned}
&\mathsf{P}(B^{r_1}_{1,k-1}D^0_{1,1}B^{r_2}_{2,k-1}D^0_{2,1}G_{n-2-2k})\\
&\qquad = Poly(k-1,1,k-1,1,n-2-2k)\binom{k-1}{r_1}\binom{k-1}{r_2}\\
&\qquad\qquad \times (1-p_{x_1}(t_1+\delta)-p_{x_2}(t_2+\delta))^{n-2-2k}\\
&\quad \times \mathsf{P}(Y=y_1, X\in B(x_1,t_1-\delta))^{r_1}\mathsf{P}(Y\neq y_1, X\in B(x_1,t_1-\delta))^{k-1-r_1}\\
&\qquad \times P(Y\neq y_1, X\in S_{x_1}(t_1,\delta))\mathsf{P}(Y=y_2, X\in B(x_2,t_2-\delta))^{r_2}\\
&\qquad\quad \times \mathsf{P}(Y\neq y_2, X\in B(x_2,t_2-\delta))^{k-1-r_2}P(Y\neq y_2, X\in S_{x_2}(t_2,\delta)).
\end{aligned}$$

Действительно, существует $Poly(k-1,1,k-1,1,n-2-2k)$ способов разбить $X^3,\ldots,X^n$ на группы, принадлежащие, соответственно, попарно непересекающимся (при условиях, наложенных на $t_1,t_2,|x_1-x_2|$ и δ) множествам $B(x_1,t_1-\delta)$, $S_{x_1}(t_1,\delta)$, $B(x_2,t_2-\delta)$, $S_{x_2}(t_2,\delta)$ и $\mathbb{R}^d\setminus\cup_{j=1}^2 B(x_j,t_j+\delta)$. Заметим, что существует $\binom{k-1}{r_1}$ способов выбрать r_1 точку X^i, $i\in I$, из $X^{q_1},\ldots,X^{q_{k-1}}$ $(3\le q_1<\ldots<q_{k-1}\le n)$, принадлежащих такому множеству $B(x_1,t_1-\delta)$, что $Y^i=y_1$ for $i\in I$ и $Y^q\neq y_1$ для $q\in\{q_1,\ldots,q_{k-1}\}\setminus I$, $\sharp I=r_1$. Подобным образом можно объяснить появление множителя $\binom{k-1}{r_2}$. Для остальных трех событий вероятности могут быть вычислены аналогично. В результате мы получим

$$\begin{aligned}
&\mathsf{P}(\xi_{n,1}=r_1,\xi_{n,2}=r_2,B_{1,k-1}D_{1,1}B_{2,k-1}D_{2,1}G_{n-2-2k})\\
&= Poly(k-1,1,k-1,1,n-2-2k)(1-p_{x_1}(t_1+\delta)-p_{x_2}(t_2+\delta))^{n-2-2k}\\
&\quad\times\prod_{j=1}^{2}\Bigg(\binom{k-1}{r_j}\mathsf{P}(Y=y_j,F(x_j,t_j,\delta))^{r_j}\mathsf{P}(Y\neq y_j,F(x_j,t_j,\delta))^{k-1-r_j}\\
&\qquad\qquad\times P(Y\neq y_j, X\in S_{x_j}(t_j,\delta))\\
&\qquad+\binom{k-1}{r_j-1}\mathsf{P}(Y=y_j,F(x_j,t_j,\delta))^{r_j-1}\mathsf{P}(Y\neq y_j,F(x_j,t_j,\delta))^{k-r_j}\\
&\qquad\qquad\times P(Y=y_j, X\in S_{x_j}(t_j,\delta))\Bigg),
\end{aligned}$$

где $F(x_j,t_j,\delta):=\{X\in B(x_j,t_j-\delta)\}$, $j=1,2$. Следовательно, в силу (107) и поскольку для всех $x\in\mathbb{R}^d$ распределение $\|X-x\|$ на $\mathbb{R}_+$ эквивалентно мере Лебега $\Lambda_{\mathbb{R}_+}$, можно утверждать следующее. Для всех $x_1,x_2\in\mathbb{R}^d$ $(x_1\neq x_2)$ и $\Lambda\otimes\Lambda$-почти всех $t=(t_1,t_2)\in(0,\infty)\times(0,\infty)$ таких, что $t_1+t_2<|x_1-x_2|$,

имеем

$$\lim_{\delta\to 0+}\frac{\mathsf{P}(\xi_{n,k,1}(z_1,z_2)=r_1,\xi_{n,k,2}(z_1,z_2)=r_2,A_{n,k})}{\mathsf{P}(A_{n,k})}$$

$$=\lim_{\delta\to 0+}\prod_{j=1}^{2}\Bigg\{\binom{k-1}{r_j}\left(\frac{\mathsf{P}(Y=y_j,X\in B(x_j,t_j-\delta))}{p_{x_1}(t_1-\delta)}\right)^{r_j}$$

$$\times\left(\frac{\mathsf{P}(Y\neq y_j,X\in B(x_j,t_j-\delta))}{p_{x_1}(t_1-\delta)}\right)^{k-1-r_j}\left(\frac{P(Y\neq y_j,X\in S_{x_j}(t_j,\delta))}{P(X\in S_{x_j}(t_j,\delta))}\right)$$

$$+\binom{k-1}{r_j-1}\left(\frac{\mathsf{P}(Y=y_j,X\in B(x_j,t_j-\delta))}{p_{x_1}(t_1-\delta)}\right)^{r_j-1}$$

$$\times\left(\frac{\mathsf{P}(Y\neq y_j,X\in B(x_j,t_j-\delta))}{p_{x_1}(t_1-\delta)}\right)^{k-r_j}\left(\frac{P(Y=y_j,X\in S_{x_j}(t_j,\delta))}{P(X\in S_{x_j}(t_j,\delta))}\right)\Bigg\}$$

$$=\prod_{j=1}^{2}\Bigg\{\binom{k-1}{r_j}p_j^{r_j}(1-p_j)^{k-1-r_j}\lim_{\delta\to 0+}\left(\frac{P(Y\neq y_j,X\in S_{x_j}(t_j,\delta))}{P(X\in S_{x_j}(t_j,\delta))}\right)$$

$$+\binom{k-1}{r_j-1}p_j^{r_j-1}(1-p_j)^{k-r_j}\lim_{\delta\to 0+}\left(\frac{P(Y=y_j,X\in S_{x_j}(t_j,\delta))}{P(X\in S_{x_j}(t_j,\delta))}\right)\Bigg\}$$

$$=\prod_{j=1}^{2}\Bigg\{\binom{k-1}{r_j}p_j^{r_j}(1-p_j)^{k-1-r_j}\alpha(x_j,y_j,t_j)$$

$$+\binom{k-1}{r_j-1}p_j^{r_j-1}(1-p_j)^{k-r_j}(1-\alpha(x_j,y_j,t_j))\Bigg\},\quad(109)$$

где $p_j:=p_j(x_j,y_j,t_j)=\mathsf{P}(Y=y_j|X\in B(x_j,t_j))$ и мы используем тот факт, что функция $\mathsf{P}(X\in B(x,t))$ непрерывна по $(x,t)\in\mathbb{R}^d\times\mathbb{R}_+$. Также применим неравенство $\mathsf{P}(X\in B(x,t))>0$ для всех $x\in\mathbb{R}^d$ и $t>0$, поскольку $f_X(z)$ строго положительна для μ-почти всех $z\in\mathbb{R}^d$.

Заметим, что доказательство теоремы 5.8.8 в [2] обеспечивает выполнение (100) и в случае, если мы заменим шары $B(t,\delta)$ кубами $\widetilde{B}(t,\delta):=\prod_{i=1}^{s}[t_i-\delta,t_i+\delta]$, где $t=(t_1,\ldots,t_s)\in\mathbb{R}^s$. Таким образом, для P_ζ-почти всех $t\in\mathbb{R}^s$ вместо (100) можно написать, что

$$\lim_{\delta\to 0+}\frac{G(\tilde{B}(t,\delta))}{\mathsf{P}_\zeta(\tilde{B}(t,\delta))}=\mathsf{E}(\tau|\zeta=t).$$

Теперь возьмем $s = 2$, $\tau := \mathbb{I}\{\xi_{n,k,1}(z_1, z_2) = r_1, \xi_{n,k,2}(z_1, z_2) = r_2\}$, $\zeta := ((\rho_{n,k,1}(x_1, x_2), \rho_{n,k,2}(x_1, x_2))$, где $n \in \mathbb{N}$, $n > 2$, $k = 0, \ldots, n-1$, $z_1 = (x_1, y_1) \in \mathbb{R}^d \times M$, $z_2 = (x_2, y_2) \in \mathbb{R}^d \times M$ и $r_1, r_2 \in \{0, \ldots, k\}$. Тогда для P_ζ-почти всех $t = (t_1, t_2) \in (0, \infty) \times (0, \infty)$ мы получим

$$\mathsf{P}(\xi_{n,k,1}(z_1, z_2) = r_1, \xi_{n,k,2}(z_1, z_2) = r_2 | \rho_{n,k,1}(x_1, x_2) = t_1, \rho_{n,k,2}(x_1, x_2) = t_2)$$

$$= \lim_{\delta \to 0+} \frac{\mathsf{E}(\mathbb{I}\{\zeta \in \widetilde{B}(t, \delta)\}\tau)}{\mathsf{P}(\zeta \in \widetilde{B}(t, \delta))}. \tag{110}$$

Теперь покажем, что для всех $x_1, x_2 \in \mathbb{R}^d$ $(x_1 \neq x_2)$ и P_ζ-почти всех $t = (t_1, t_2)$, принадлежащих множеству

$$B_{x_1,x_2}(\varepsilon) = \{t \in \mathbb{R}^2 : 0 < t_j \leq |x_1 - x_2|/2 - \varepsilon,\ j = 1, 2\},$$

последний предел совпадает со значением, полученным в (109) для

$$\lim_{\delta \to 0+} \mathsf{P}(\xi_{n,k,1}(z_1, z_2) = r_1, \xi_{n,k,2}(z_1, z_2) = r_2 | A_{n,k}),$$

где $A_{n,k}$ зависит от x_1, x_2 и δ.

Для этого убедимся, что для всех $x_1, x_2 \in \mathbb{R}^d$ $(x_1 \neq x_2)$, существует такая положительная измеримая функция $f_{n,k}(x_1, x_2, \cdot, \cdot)$, что $B \in \mathcal{B}(\mathbb{R}^2)$ и $B \subset B_{x_1,x_2}(\varepsilon)$, тогда

$$\mathsf{P}(\zeta \in B) = \int_B f_{n,k}(x_1, x_2, u_1, u_2) du_1 du_2. \tag{111}$$

Для $(u_1, u_2) \in B_{x_1,x_2}(\varepsilon)$ можно написать

$$\mathsf{P}(\rho_{n,k,1}(x_1, x_2) \leq u_1, \rho_{n,k,2}(x_1, x_2) \leq u_2)$$
$$= \sum_{(l_1,l_2) \in J(n,k)} \binom{n}{l_1} \binom{n - l_1}{l_2} p_{x_1}^{l_1}(u_1) p_{x_2}^{l_2}(u_2)(1 - p_{x_1}(u_1) - p_{x_2}(u_2)))^{n-l_1-l_2},$$

где $J(n, k) := \{(l_1, l_2) : l_1 \geq k, l_2 \geq k, l_1 + l_2 \leq n - 2\}$. Это множество непусто, поскольку $(k, k) \in J(n, k)$.

Следовательно, мы получили полином $p_{x_1}(u_1)$ и $p_{x_2}(u_2)$. При доказательстве леммы 2.6 мы убедились, что для $x \in \mathbb{R}^d$, $u > 0$ и $l \in \mathbb{N}$, функция распределения $p_x(u)^l$ имеет плотность $l p_x^{l-1}(u) f_x(u)$, где $f_x(\cdot)$ является плотностью $p_x(\cdot)$ относительно $\Lambda_{\mathbb{R}_+}$. Таким образом, $p_{x_1}^{l_1}(u_1) p_{x_2}^{l_2}(u_2)$ абсолютно непрерывна относительно $\Lambda_{\mathbb{R}_+} \otimes \Lambda_{\mathbb{R}_+}$. Следовательно, существует интегрируемая (относительно ограничения меры $\Lambda \otimes \Lambda$ на $B_{x_1,x_2}(\varepsilon)$)

функция $f_{n,k}(x_1, x_2, \cdot, \cdot)\colon B_{x_1,x_2}(\varepsilon) \to \mathbb{R}$ такая, что для параллелепипеда $(0, u_1] \times (0, u_2]$, $(u_1, u_2) \in B_{x_1,x_2}(\varepsilon)$, формула (111) имеет место. Аддитивность интеграла влечет выполнение указанной формулы и для параллелепипедов $(a_1, b_1] \times (a_2, b_2]$, где $a_i \le b_i$ $(i = 1, 2)$, $(a_1, a_2), (b_1, b_2) \in B_{x_1,x_2}(\varepsilon)$. Более того, формула (111) верна для алгебры $\mathcal{E}$ таких подмножеств $B_{x_1,x_2}(\varepsilon)$, которые могут быть представлены в виде конечного объединения таких параллелепипедов. Таким образом, мы убеждаемся, что

$$Q(B) := \int_B f_{n,k}(x_1, x_2, u_1, u_2)\, du_1 du_2, \quad B \in \mathcal{B}(\mathbb{R}^2) \cap B_{x_1,x_2}(\varepsilon),$$

и $G(B) := \mathsf{P}(\zeta \in B)$ совпадают на $\mathcal{E}$. Подобно доказательству леммы 2.6 мы получаем $f_{n,k}(x_1, x_2, u_1, u_2) \ge 0$ для $(\Lambda \otimes \Lambda)$-почти всех $u = (u_1, u_2)$ из $B_{x_1,x_2}(\varepsilon)$. Следовательно, формула (111) доказана.

Сравним (109) и (110). Мы показали, что для всех $(t_1, t_2) \in B_{x_1,x_2}(\varepsilon)$ и всех достаточно малых $\delta > 0$ (т.е. таких, что $\delta < \Delta(x_1, x_2, t_1, t_2)$), мы имеем

$$\mathsf{E}(\mathbb{I}\{\zeta \in C(t, \delta)\}\tau) = \mathsf{E}(\mathbb{I}\{\zeta \in \widetilde{B}(t, \delta)\tau)$$

и

$$\mathsf{P}(\zeta \in C(t, \delta)) = \mathsf{P}(\zeta \in \widetilde{B}(t, \delta)),$$

где $C(t, \delta) := (t_1 - \delta, t_1 + \delta] \times (t_2 - \delta, t_2 + \delta]$. Используя соотношение $C(t, \delta) \subset \widetilde{B}(t, \delta) \subset B_{x_1,x_2}(\varepsilon)$ для всех $\delta \in (0, \Delta(x_1, x_2, t_1, t_2))$ и в силу (111), мы получаем $\mathsf{P}(\zeta \in \widetilde{B}(t, \delta) \setminus C(t, \delta)) = 0$, поскольку

$$(\Lambda \otimes \Lambda)(\widetilde{B}(t, \delta) \setminus C(t, \delta)) = 0.$$

Учитывая, что ограничение P_ζ на $B_{x_1,x_2}(\varepsilon)$ абсолютно непрерывно относительно соответствующего ограничения $\Lambda_{\mathbb{R}_+} \otimes \Lambda_{\mathbb{R}_+}$, мы заключаем, что для всех $x_1, x_2 \in \mathbb{R}^d$ $(x_1 \ne x_2)$ и P_ζ-почти всех $(t_1, t_2) \in B_{x_1,x_2}(\varepsilon)$, формулы (52) и (53) установлены.

Заметим, что $B_{x_1,x_2}(\varepsilon) \uparrow (0, |x_1 - x_2|/2) \times (0, |x_1 - x_2|/2)$ при $\varepsilon \to 0$. Следовательно, формулы (52) и (53) выполняются P_ζ-п.н. для точек множества $(0, |x_1 - x_2|/2) \times (0, |x_1 - x_2|/2)$. $\square$

2.4.6 Доказательство леммы 2.9

Имеем

$$\mathsf{E}(W\mathbb{I}\{V \in B\}) = \mathsf{E}(\mathsf{E}(W\mathbb{I}\{V \in B\})|V) = \mathsf{E}(\mathbb{I}\{V \in B\}\mathsf{E}(W|V))$$

$$= \int_{\mathbb{R}^m} \mathbb{I}\{t \in B\}\mathsf{E}(W|V = t)\mathsf{P}_V(dt) = \int_B \mathsf{E}(W|V = t)\mathsf{P}_V(dt).$$

Следовательно,

$$\mathsf{E}(W|V\in B)=\frac{\mathsf{E}(W\mathbb{I}\{V\in B\})}{\mathsf{P}(V\in B)}$$
$$=\frac{\int_B\mathsf{E}(W|V=t)\mathsf{P}_V(dt)}{\mathsf{P}(V\in B)}=\int_B\mathsf{E}(W|V=t)\widetilde{\mathsf{P}}_{V,B}(dt),$$

где $\widetilde{\mathsf{P}}_{V,B}(D)=\frac{\mathsf{P}(D\cap B)}{\mathsf{P}(B)}$, $D\in\mathcal{B}(\mathbb{R}^m)$. $\square$

2.4.7 Доказательство леммы 2.10

Достаточно показать, что для любого множества $B\in\mathcal{B}(\mathbb{R})$,

$$\int_B\mathsf{E}(\xi|\eta=t)\,\mathsf{P}_\eta(dt)=\sum_{r\in S}\int_B\mathsf{E}(\xi|\zeta=r,\eta=t)\mathsf{P}(\zeta=r|\eta=t)\,\mathsf{P}_\eta(dt).$$

Понятно, что

$$\int_B\mathsf{E}(\xi|\eta=t)\,\mathsf{P}_\eta(dt)=\mathsf{E}\xi\mathbb{I}(\eta\in B)=\sum_{r\in S}\mathsf{E}\xi\mathbb{I}(\eta\in B,\zeta=r)$$
$$=\sum_{r\in S}\int_{B\times\{r\}}\mathsf{E}(\xi|\eta=t,\zeta=v)d\mathsf{P}_{\eta,\zeta}(t,v),$$

где $d\mathsf{P}_{\eta,\zeta}(t,r)$ означает интегрирование по мере $\mathsf{P}(\eta\in\cdot,\zeta\in\cdot)$. Теперь покажем, что для измеримой функции $\varphi\colon\mathbb{R}\times S\to\mathbb{R}$, удовлетворяющей условию $\mathsf{E}|\varphi(\eta,\zeta)|<\infty$, верно следующее соотношение

$$\int_{B\times\{r\}}\varphi(t,v)d\mathsf{P}_{\eta,\zeta}(t,v)=\int_B\varphi(t,r)\mathsf{P}(\zeta=r|\eta=t)\mathsf{P}_\eta(dt).\tag{112}$$

Действительно, для $A\in\mathcal{B}(\mathbb{R})$ и $s\in S$, рассмотрим

$$\varphi(t,v):=\mathbb{I}\{t\in A,v=s\},\quad t\in\mathbb{R},\quad v\in S.$$

Очевидно, если $s\neq r$, то (112) верно. Если $s=r$, то

$$\int_{B\times\{r\}}\varphi(t,v)d\mathsf{P}_{\eta,\zeta}(t,v)=\mathsf{P}_{\eta,\zeta}((A\cap B)\times\{r\})=\mathsf{P}(\zeta=r,\eta\in A\cap B)$$
$$=\int_{A\cap B}\mathsf{P}(\zeta=r|\eta=t)\mathsf{P}_\eta(dt)=\int_B\mathbb{I}(t\in A,r=r)\mathsf{P}(\zeta=r|\eta=t)\mathsf{P}_\eta(dt)$$
$$=\int_B\varphi(t,r)\mathsf{P}(\zeta=r|\eta=t)\mathsf{P}_\eta(dt).$$

Таким образом, (112) выполняется для $\varphi(t,v) = \mathbb{I}\{t \in A, v \in E\}$, где можно взять произвольные $A \in \mathcal{B}(\mathbb{R})$ и $E \subset S$. Учитвая, что любая измеримая функция $\varphi : \mathbb{R} \times S \to \mathbb{R}$ может быть приближена конечной линейной комбинацией рассматрвиаемых функций типа $\mathbb{I}\{A\}\mathbb{I}\{E\}$, мы приходим к искомому утверждению (112). Также заметим, что $\mathsf{E}|\mathsf{E}(\xi|\eta,\zeta)| \leq \mathsf{E}|\xi| < \infty$. $\square$

3 Оценивание взаимной информации в смешанной модели

При подготовке данной главы диссертации использован материал публикации [102], выполненной автором в соавторстве с профессором А.В.Булинским. В публикации А.В.Булинскому принадлежит постановка задач и общий подход к их решению, им также доказана лемма 1 (лемма 3.9 в диссертации). Все остальные результаты принадлежат автору диссертации, им проведены все компьютерные симуляции.

3.1 Постановка задачи

Приведем основные определения, используемые далее. Рассмотрим два измеримых пространства $(S, \mathcal{B})$, $(M, \mathcal{A})$ и два случайных элемента (величины или вектора) X, Y, определенных на полном вероятностном пространстве $(\Omega, \mathcal{F}, \mathsf{P})$. Пусть $X : \Omega \to S$, $X \in \mathcal{F}|\mathcal{B}$ (т.е. для любых $B \in \mathcal{B}$ выполняется $X^{-1}(B) \in \mathcal{F}$), $Y : \Omega \to M$, $Y \in \mathcal{F}|\mathcal{A}$. Положим $(U, \mathcal{C}) := (S \times M, \mathcal{B} \otimes \mathcal{A})$, где σ-алгебра $\mathcal{B} \otimes \mathcal{A}$ порождена "прямоугольниками" $B \times A$, здесь $B \in \mathcal{B}$ и $A \in \mathcal{A}$. Пусть $\mathsf{P}_{X,Y}$ означает распределение случайного вектора (X, Y) на $\mathcal{C}$. Обозначим как P_X и P_Y законы распределения X и Y на $\mathcal{B}$ и $\mathcal{A}$, соответственно. Тогда взаимная информация X и Y определяется формулой

$$I(X, Y) = D_{KL}(\mathsf{P}_{X,Y}, \mathsf{P}_X \otimes \mathsf{P}_Y), \tag{113}$$

где D_{KL} — дивергенция Кульбака-Лейблера мер $\mathsf{P}_{X,Y}$ и $\mathsf{P}_X \otimes \mathsf{P}_Y$. Дивергенция Кульбака-Лейблера для веростностных мер $\mathbb{P}$ и $\mathbb{Q}$ на измеримом пространстве $(T, \mathcal{E})$ определяется следюущим образом:

$$D_{KL}(\mathbb{P}, \mathbb{Q}) := \begin{cases} \int_T \log\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right) d\mathbb{P}, & \text{если } \mathbb{P} \ll \mathbb{Q}, \\ \infty, & \text{иначе,} \end{cases}$$

где $\frac{d\mathbb{P}}{d\mathbb{Q}}$ — производная Радона-Никодима $\mathbb{P}$ относительно $\mathbb{Q}$. Далее все логарифмы будут браться по основанию e, поскольку константа не влияет на установленные результаты. Из формулы (113) следуют важные свойства взаимной информации: $I(X, Y) \geq 0$ и $I(X, Y) = 0$ тогда и только тогда, когда X и Y независимы.

В данной главе рассматриваются случайные элементы X и Y, обладающие следующими свойствами (см. главу 2, [44], [65], [73]). Предположим, что X принимает значения в пространстве $\mathbb{R}^d$, снабженном σ-алгеброй

$\mathcal{B}(\mathbb{R}^d)$ (таким образом, $S = \mathbb{R}^d$ и $\mathcal{B} = \mathcal{B}(\mathbb{R}^d)$), Y принимает значения во множестве M и $\mathcal{A} = 2^M$, т.е. состоит из всех подмножеств M. Пусть $\mathsf{P}(Y = y) > 0$ для всех $y \in M$. Предположим, что $\mathsf{P}_{X,Y} \ll \mu \otimes \lambda$, где μ — мера Лебега на $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, λ — считающая мера на $(M, 2^M)$. Такую модель мы будем называть смешанной. Отметим, что описанная модель включает в себя классическую логистическую регрессию (см., например, [67]).

В данной модели производная Радона-Никодима

$$\frac{d\mathsf{P}_{X,Y}}{d(\mu \otimes \lambda)} := f_{X,Y}, \tag{114}$$

поэтому существует $\frac{d\mathsf{P}_X}{d\mu} = f_X$, где $f_X(x) = \sum_{y \in M} f_{X,Y}(x, y)$ и $\frac{d\mathsf{P}_Y}{d\lambda}(y) = \mathsf{P}(Y = y)$, $x \in \mathbb{R}^d$, $y \in M$. Понятно, что $(\mathsf{P}_X \otimes \mathsf{P}_Y)$-п.н. выполняется

$$\frac{d\mathsf{P}_{X,Y}}{d(\mathsf{P}_X \otimes \mathsf{P}_Y)}(x, y) = \frac{f_{X,Y}(x, y)}{f_X(x)\mathsf{P}(Y = y)}, \quad x \in \mathbb{R}^d, \ y \in M$$

(здесь положим, что $0/0 := 0$). Следовательно,

$$I(X, Y) = \sum_{y \in M} \int_{\mathbb{R}^d} \left(\log \frac{f_{X,Y}(x, y)}{f_X(x)\mathsf{P}(Y = y)} \right) f_{X,Y}(x, y)\mu(dx), \tag{115}$$

где мы формально полагаем $0 \log 0 := 0$. Таким образом,

$$I(X, Y) = \sum_{y \in M} \int_{\mathbb{R}^d} (\log f_{X,Y}(x, y) - \log f_X(x) - \log \mathsf{P}(Y = y)) f_{X,Y}(x, y)\mu(dx). \tag{116}$$

Для взаимной информации, определяемой формулой (116), существует три представления (при условии, что все выражения в них определены):

$$I(X, Y) = H(X) + H(Y) - H(X, Y), \tag{117}$$

$$I(X, Y) = H(X) - H(X|Y), \tag{118}$$

$$I(X, Y) = H(Y) - H(Y|X). \tag{119}$$

Здесь энтропия Шеннона случайных объектов X, Y и (X, Y) определяется следующим образом:

$$H(X) = -\int_{\mathbb{R}^d} (\log f_X(x)) f_X(x)\mu(dx), \tag{120}$$

$$H(Y) = -\sum_{y\in M}(\log \mathsf{P}(Y=y))\mathsf{P}(Y=y), \tag{121}$$

$$H(X,Y) = -\sum_{y\in M}\int_{\mathbb{R}^d}(\log f_{X,Y}(x,y))f_{X,Y}(x,y)\mu(dx). \tag{122}$$

Условная энтропия Шеннона X (при условии Y) и Y (при условии X) задаются, соответственно, формулами

$$H(X|Y) = -\sum_{y\in M}\int_{\mathbb{R}^d}(\log f_{X|Y}(x|y))f_{X,Y}(x,y)\mu(dx), \tag{123}$$

$$H(Y|X) = -\sum_{y\in M}\int_{\mathbb{R}^d}(\log f_{Y|X}(y|x))f_{X,Y}(x,y)\mu(dx), \tag{124}$$

где условные плотности имеют вид

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{\mathsf{P}(Y=y)}, \quad f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}, \quad x\in\mathbb{R}^d, y\in M. \tag{125}$$

Отметим, что $0 \le H(Y) < \infty$ для дискретной случайной величины Y, принимающей значения в конечном множестве M. Более того, в смешанной модели $H(Y|X) < \infty$. Действительно, в силу (124)

$$H(Y|X) = -\sum_{y\in M}\int_{\mathbb{R}^d}(\log f_{Y|X}(y|x))f_{Y|X}(y|x)f_X(x)\mu(dx)$$
$$\le \sum_{y\in M}\int_{\mathbb{R}^d} f_X(x)\mu(dx) = \sharp(M),$$

где для $u\in[0,1]$ мы используем неравенство $-\log u \le \frac{1}{u}$ (формально полагаем $\log 0 := -\infty$, $\frac{1}{0} := \infty$), а $\sharp(M)$ означает мощность множества M. Также применим теорему Лебега о дифференцировании (см, например, [99], стр. 654) для того, чтобы показать, что $f_{Y|X}(y|x) \le 1$ для всех $y\in M$ и μ-почти всех $x\in\mathbb{R}^d$. Таким образом, $H(Y|X)\ge 0$.

Следовательно, в смешанной модели $I(X,Y)$ всегда конечна и (119) выполняется. Однако чтобы гарантировать выполнение (118), нужно потребовать выполнения особых условий, например, чтобы энтропия $H(X)$ была конечна.

Формулы (162) – (119) показывают, что существует множество способов оценить $I(X,Y)$ по выборке, состоящей из независимых одинаково распределенных наблюдений $(X_1,Y_1),\dots,(X_n,Y_n)$, где (X_1,Y_1) распределено так

же, как и (X, Y). Если (X, Y) имеет плотность относительно меры Лебега на пространстве $\mathbb{R}^m$ (таким образом, Y также имеет плотность), то можно воспользоваться статистическими оценками взаимной информации $I(X, Y)$, предложенными в работе [60]. Идея этих оценок состоит в использовании статистик k-ближайших соседей (для оценивания энтропии Шеннона абсолютно непрерывного вектора). Статистическая оценка $H(X)$ с помощью статистик ближайших соседей была впервые предложена в работе [5] и развивалась в ряде работ, см., например, [18], [27], [77] и приводимые в этих работах ссылки. Для смешанной модели оценка $I(X, Y)$, основанная на формуле (118) была предложена в работе [32], однако там не исследовались ее асимптотические свойства. В данной главе приведены условия асимптотической несмещенности и L_2-состоятельности упомянутой оценки.

В главе 3 также вводится новая оценка $I(X, Y)$, основанная на оценке условной энтропии, описанной в предыдущей главе.

3.2 Основной результат

Пусть $(X_1, Y_1), (X_2, Y_2), \ldots$ — независимые одинаково распределенные случайные велчины такие, что $Law(X_1, Y_1) = Law(X, Y)$, а распределения (X, Y) описывается смешанной моделью, определенной в предыдущей главе. Пусть $\zeta_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, $n \in \mathbb{N}$. Для $y \in M$ рассмотрим последовательность индексов $1 \leq j_1^y < j_2^y < \ldots$ таких, что $Y_{j_k^y} = y$. Тогда можно написать

$$\zeta_n^y = \{(X_{j_k^y}, y)\}_{k=1}^{N_{y,n}}, \quad N_{y,n} = \max\{k \in \mathbb{N} \colon j_k^y \leq n\}.$$

Предположим, что энтропия $H(X|Y)$ определена и конечна. Положим

$$\widehat{P}_n(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{Y_i = y\}, \;\; y \in M, \;\; n \in \mathbb{N}, \tag{126}$$

где $\mathbb{I}\{\cdot\}$ — индикаторная функция. Тогда $\widehat{P}_n(y) = \frac{N_{y,n}}{n}$, $y \in M$, $n \in \mathbb{N}$. В работе [32] была предложена следующая оценка взаимной информации $I(X, Y)$ для смешанной модели

$$\widehat{I}_{n,k}(X; Y) := \widehat{H}_{n,k}(X) - \sum_{y \in M} \widehat{H}_{n,k}(X|Y = y)\widehat{P}_n(y). \tag{127}$$

Здесь $\widehat{H}_{n,k}(X)$ — оценка энтропии $H(X)$, построенная по выборке ζ_n следующим образом:

$$\widehat{H}_{n,k}(X) = \frac{1}{n}\sum_{i=1}^{n} \eta_{n,k,i}(X_1, \ldots, X_n), \tag{128}$$

$$\eta_{n,k,i}(X_1, \ldots, X_n) := d \log \rho_{n,k,i}(X_1, \ldots, X_n) + \log n - \log k + \log V_d, \tag{129}$$

$$\rho_{n,k,i}(X_1, \ldots, X_n) := \|X_i - X_{i,(k)}\|, \tag{130}$$

где $V_d = \frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)}$ объем единичного шара в $\mathbb{R}^d$ и для $k \in \mathbb{N}$ $X_{i,(k)}$ является k ближайшим соседом X_i в выборке $\{X_1, \ldots, X_n\} \setminus \{X_i\}$, $\|\cdot\|$ обозначает евклидову норму в $\mathbb{R}^d$. Таким образом, $\rho_{n,k,i}(X_1, \ldots, X_n)$ является расстоянием от X_i до $X_{i,(k)}$. Пусть для $y \in M$ величина $\widehat{H}_{n,k}(X|Y=y)$ является оценкой

$$H(X|Y=y) = -\int_{\mathbb{R}^d} (\log f_{X|Y}(x|y)) f_{X|Y}(x|y) dx, \tag{131}$$

определенной полностью аналогично $\widehat{H}_{n,k}(X)$, но с ζ_n^y вместо ζ_n. Следовательно,

$$\widehat{H}_{n,k}(X|Y=y) = \begin{cases} \frac{1}{N_{y,n}} \sum_{i=1}^{N_{y,n}} \eta_{N_{y,n},k,i}(X_{j_i^y}, \ldots, X_{j_{N_{y,n}}^y}), & N_{y,n} \geq k, \\ 0, & \text{иначе.} \end{cases} \tag{132}$$

Отметим, что в отличие от (128) в выражении (132) мы имеем выборки случайного размера, исследование которых является отдельным направлением (см., например, [13, 49, 59]) Следовательно, вообще говоря, возможна ситуация, в которой $N_{y,n} < k$, а значит, нельзя найти $X_{i,(k)}$. В таком случае мы будем считать, что $\widehat{H}_{n,k}(X|Y=y) = 0$.

Нам понадобится следующий результат. Его доказательство вынесено в раздел 3.6.

Лемма 3.1 *Для всех $y \in M$ и всех $n \in \mathbb{N}$ случайные величины $N_{y,n}$, $X_{j_1^y}$, $X_{j_2^y}, \ldots$ независимы. Более того, для каждого $k \in \{0, \ldots, n\}$*

$$\mathsf{P}(N_{y,n} = k) = \binom{n}{k} \mathsf{P}(Y = y)^k \mathsf{P}(Y \neq y)^{n-k} \tag{133}$$

и $\mathsf{P}(X_{j_i^y} \in B) = \mathsf{P}(X \in B | Y = y)$, где $B \in \mathcal{B}(\mathbb{R}^d)$, $i \in \mathbb{N}$ и $y \in M$.

Замечание 3.2 . *Очевидно, что лемма 3.1 является более общей, чем лемма 2.8, поскольку последовательность $X_{j_1^y}, X_{j_2^y}, \ldots$ рассматривается для случайного $N_{y,n}$.*

Чтобы сформулировать основной результат, введем некоторые обозначения. Для $x \in \mathbb{R}$ и $r \geq 0$ положим $B(x,r) = \{v \in \mathbb{R}^d \colon \|x - v\| \leq r\}$. Аналогично работам [28] и [27], для $r > 0$, $R > 0$, $\varepsilon > 0$, $\nu > 0$ и вероятностной плотности f (относительно меры μ, заданной на $\mathbb{R}^d$) введем функционалы, принимающие значения в $[0, \infty]$:

$$I_f(x,r) := \frac{\int_{B(x,r)} f(z)\, dz}{r^d V_d}, \tag{134}$$

$$M_f(x,R) = \sup_{r \in (0,R]} I_f(x,r), \quad m_f(x,R) = \inf_{r \in (0,R]} I_f(x,r), \tag{135}$$

$$L_f(\nu) := \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |\log \|x - z\||^\nu f(x) f(z) dx\, dz, \tag{136}$$

$$Q_f(\varepsilon, R) := \int_{\mathbb{R}^d} M_f^\varepsilon(x,R) f(x)\, dx, \tag{137}$$

$$T_f(\varepsilon, R) := \int_{\mathbb{R}^d} m_f^{-\varepsilon}(x,R) f(x)\, dx. \tag{138}$$

Для упрощения обозначений далее будем писать dx, dz вместо $\mu(dx)$, $\mu(dz)$, соответственно. Отметим, что в [28] рассматриваются более общие функционалы, которые используются для изучения асимптотических свойств статистических оценок дивергенции Кульбака-Лейблера (а также дифференциальной энтропии Шеннона). Для $\varepsilon > 0$, будем писать $Q_f(\varepsilon) := Q_f(\varepsilon, \varepsilon)$ и $T_f(\varepsilon) := T_f(\varepsilon, \varepsilon)$. Для вероятностной плотности f на $\mathbb{R}^d$ и некоторого $\varepsilon_i > 0, R_i > 0$, $i = 1, 2$ конечность $Q_f(\varepsilon_1, R_1)$ и $T_f(\varepsilon_2, R_2)$, а также конечность $Q_f(\varepsilon)$ и $T_f(\varepsilon)$ обсуждается в следствии 1 [27]. Как обычно, предполагается, что если $\int_{\mathbb{R}^d} h(x) dx$ существует, то значение интеграла сохраняется, если положить $h(x) = \infty$ (или $-\infty$) для x, принадлежащих множеству $B \in \mathcal{B}(\mathbb{R}^d)$ такому, что $\mu(B) = 0$. Поэтому в формуле (136) не предполагается, что $x \neq y$.

Лемма 3.3 *Пусть для некоторых $\nu > 1$ и $\varepsilon > 0$, выполняются следующие неравенства:*

$$L_f(\nu) < \infty, \;\; Q_f(\varepsilon) < \infty, \;\; T_f(\varepsilon) < \infty,$$

если положить $f = f_X(\cdot)$ и $f = f_{X|Y}(\cdot|y)$, $y \in M$. Тогда для всех $k \in \mathbb{N}$

$$\mathsf{E}\widehat{I}_{n,k}(X,Y) \to I(X,Y), \;\; n \to \infty. \tag{139}$$

Доказательство. Согласно замечанию 2.4 [27], конечность $Q_f(\varepsilon)$ и $T_f(\varepsilon)$ для некоторого $\varepsilon > 0$ влечет неравенство

$$\int_{\mathbb{R}^d} |\log f(x)| f(x)\, dx < \infty$$

и, следовательно, $|H(X)| < \infty$. Применим этот результат к $f = f_{X|Y}(\cdot|y)$, $y \in M$. Таким образом, $|H(X|Y = y)| < \infty$ для всех $y \in M$. Следовательно, $|H(X|Y)| \le \sum_{y\in M} |H(X|Y = y)|\mathsf{P}(Y = y) < \infty$. Теорема 1 [28] утверждает, что $f = f_X$ для каждого фикисированного $k \in \mathbb{N}$ (т.е. не только для $k = 1$ как в работе [27]), выполняется соотношение $\mathsf{E}\widehat{H}_{n,k} \to H(X)$ при $n \to \infty$. Следовательно, для доказательства утверждения (139) достаточно проверить, что

$$\sum_{y\in M} \mathsf{E}\widehat{H}_{n,k}(X|Y = y)\widehat{P}_n(y) \to H(X|Y), \quad n \to \infty. \tag{140}$$

Верна следующая цепочка равенств:

$$\begin{aligned}\sum_{y\in M} \mathsf{E}\widehat{H}_{n,k}(X|Y = y)\widehat{P}_n(y) - H(X|Y) \\ = \sum_{y\in M} \mathsf{E}(\widehat{H}_{n,k}(X|Y = y)(\widehat{P}_n(y) - \mathsf{P}(Y = y))) \\ + \sum_{y\in M} \mathsf{E}((\widehat{H}_{n,k}(X|Y = y) - H(X|Y = y))\mathsf{P}(Y = y).\end{aligned} \tag{141}$$

В силу леммы 3.1 случайные элементы $N_{y,n}$ и $X_{j_1^y}, X_{j_2^y}, \ldots$ независимы. Таким образом, для всех $y \in M$

$$\mathsf{E}\widehat{H}_{n,k}(X|Y = y) = \sum_{m=k}^{n} \left(\frac{1}{m} \mathsf{E} \sum_{i=1}^{m} \eta_{m,k,i}(X_{j_1^y}, \ldots, X_{j_m^y}) \right) \mathsf{P}(N_{y,n} = m),$$

$$\begin{aligned}\mathsf{E}(\widehat{H}_{n,k}(X|Y = y) - H(X|Y = y)) \\ = \sum_{m=k}^{n} a_m \mathsf{P}(N_{y,n} = m) - H(X|Y = y) \sum_{m=0}^{k-1} \mathsf{P}(N_{y,n} = m),\end{aligned} \tag{142}$$

где $a_m := \frac{1}{m}\mathsf{E}\sum_{i=1}^{m} \eta_{m,k,i}(X_{j_1^y}, \ldots, X_{j_m^y}) - H(X|Y = y)$, $m \in \mathbb{N}$, $m \ge k$. Очевидно, что a_m зависит от y и k. Для всех $y \in M$, случайные векторы $X_{j_1^y}, X_{j_2^y}, \ldots$ независимы и имеют то же распределение, что и X при условии

$Y = y$. В силу [28] верно соотношение $a_m \to 0$ при $m \to \infty$. Следовательно, существует $A > 0$ такое, что $|a_m| \le A$ для всех $m \in \mathbb{N}$. Если случайная величина $S(n)$ имеет биномиальное распределение $B(n,p)$ с параметром $p \in (0,1)$, то согласно классическому неравенству Хефдинга [54] для всех $\delta \in (0,p)$

$$\mathsf{P}(|S(n) - np| \ge n\delta) \le 2\exp\{-2n\delta^2\}. \tag{143}$$

Таким образом, для всех достаточно больших n можно переписать и оценить выражение $|\sum_{m=k}^{n} a_m \mathsf{P}(N_{y,n} = m)|$ следующим образом:

$$\left| \sum_{m=k}^{[np-n\delta]} a_m \mathsf{P}(N_{y,n} = m) + \sum_{m=[np-n\delta]+1}^{[np+n\delta]} a_m \mathsf{P}(N_{y,n} = m) + \sum_{m=[np+n\delta]+1}^{n} a_m \mathsf{P}(N_{y,n} = m) \right| \le A\mathsf{P}(|N_{y,n} - np| \ge n\delta) + \sup_{m > n(p-\delta)} |a_m|.$$

Соотношение $a_m \to 0$ при $m \to \infty$ влечет, что $\sup_{m>n(p-\delta)} |a_m| \to 0$ при $n \to \infty$. Также для всех достаточно больших n

$$0 \le \sum_{m=0}^{k-1} \mathsf{P}(N_{y,n} = m) \le \mathsf{P}(|N_{y,n} - np| \ge n\delta).$$

Учитывая (133), (143) с $p = \mathsf{P}(Y = y)$ и $\delta \in (0, \min_{y \in M} \mathsf{P}(Y = y))$, получим

$$\sum_{y \in M} \mathsf{E}((\widehat{H}_{n,k}(X|Y = y) - H(X|Y = y))\mathsf{P}(Y = y) \to 0, \;\; n \to \infty.$$

Теперь можем написать, что

$$\mathsf{E}(\widehat{H}_{n,k}(X|Y = y)(\widehat{P}_n(y) - \mathsf{P}(Y = y))) = \sum_{m=k}^{n} \left(\frac{m}{n} - \mathsf{P}(Y = y) \right) b_m \mathsf{P}(N_{y,n} = m),$$

где $b_m := \frac{1}{m}\mathsf{E}\left(\sum_{i=1}^{m} \eta_{m,k,i}(X_{j_1^y}, \ldots, X_{j_m^y})\right)$. Отметим, что b_m зависит от y и k. Заметим, что $a_m = b_m - H(X|Y = y) \to 0$ при $m \to \infty$. Таким образом, для всех $m \in \mathbb{N}$ и некоторой положительной константы B имеем $|b_m| \le B$, $m \in \mathbb{N}$. Следовательно, для всех $y \in M$, $p = \mathsf{P}(Y = y)$, любого $\delta \in (0, \mathsf{P}(Y = y))$ и всех достаточно больших n

$$\left| \sum_{m=k}^{n} \left(\frac{m}{n} - p \right) b_m \mathsf{P}(N_{y,n} = m) \right| \le B \sum_{m=k}^{n} \left| \frac{m}{n} - p \right| \mathsf{P}(N_{y,n} = m)$$
$$\le B \left(\delta \mathsf{P}(|N_{y,n} - np| < n\delta) + \mathsf{P}(|N_{y,n} - np| \ge n\delta) \right) \le B(\delta + 2\exp\{-2n\delta^2\}).$$

Для всех $y \in M$ приходим к соотношению

$$\mathsf{E}(\widehat{H}_{n,k}(X|Y=y)(\widehat{P}_n(y) - \mathsf{P}(Y=y))) \to 0, \;\; n \to \infty.$$

Учитывая оценки, полученные для суммы, фигурирующей в (141), мы приходим к (140). $\square$

Теперь покажем, что можно упростить или не использовать некоторые условия леммы 3.3. Заметим, что мы можем рассмотреть функционалы, определенные в (134) - (138) не только для вероятностной плотности f, но и для любой неотрицательной функции, интегрируемой на $\mathbb{R}^d$ по мере μ.

Теорема 3.4 *Предположим, что для некоторых $\nu > 1$ и $\varepsilon > 0$ выполняются следующие неравенства:*

$$L_f(\nu) < \infty, \;\; Q_f(\varepsilon) < \infty, \;\; T_g(\varepsilon) < \infty, \tag{144}$$

если мы положим $f = f_X$ и $g(\cdot) = f_{X,Y}(\cdot, y)$, $y \in M$. Тогда для всех $k \in \mathbb{N}$ (139) справедливо.

Доказательство. Положим $f(x) = f_X(x)$, $f(x|y) = f_{X|Y}(x|y)$ и $f(x,y) = f_{X,Y}(x,y)$, где $x \in \mathbb{R}^d$, $y \in M$. Очевидно, что для (μ-почти всех) $x \in \mathbb{R}^d$ можно написать $f(x) = \sum_{y \in M} f(x|y)\mathsf{P}(Y=y)$. Таким образом, для всех $y \in M$ и $x \in \mathbb{R}^d$ выполняется неравенство $f(x|y) \le \frac{f(x)}{\mathsf{P}(Y=y)}$. Следовательно, для всех $y \in M$

$$L_{f(\cdot|y)}(\nu) \le L_f(\nu)\mathsf{P}(Y=y)^{-2} < \infty,$$

поскольку $L_f(\nu) < \infty$. Ясно, что для всех $y \in M$, $r > 0$ и $R > 0$ также верно соотношение

$$I_{f(\cdot|y)}(x,r) \le \frac{1}{\mathsf{P}(Y=y)} I_f(x,r),$$

а значит, и

$$M_{f(\cdot|y)}(x,R) \le \frac{1}{\mathsf{P}(Y=y)} M_f(x,R).$$

Следовательно,

$$Q_{f(\cdot|y)}(\varepsilon,R) \le (\mathsf{P}(Y=y))^{-(1+\varepsilon)} Q_f(\varepsilon,R) < \infty,$$

поскольку $Q_f(\varepsilon,R) < \infty$.

Для некоторых $\varepsilon > 0$, $R > 0$ и каждого $y \in M$ соотношение

$$\int_{\mathbb{R}^d} m^{-\varepsilon}_{f(\cdot|y)}(x,R) f(x|y)dx < \infty,$$

выполняется тогда и только тогда, когда $\int_{\mathbb{R}^d} m^{-\varepsilon}_{f(\cdot,y)}(x,R)f(x,y)dx < \infty$ поскольку $f(\cdot|y) = \frac{f(\cdot,y)}{\mathsf{P}(Y=y)}$. Отметим, что $f(x) = \sum_{q\in M} f(x,q)$ для (μ-почти всех) $x \in \mathbb{R}^d$. Таким образом, для $x \in \mathbb{R}^d$ и $R > 0$, имеем $m_f(x,R) \geq \sum_{q\in M} m_{f(\cdot,q)}(x,R)$. Значит, для $\varepsilon > 0$,

$$T_f(\varepsilon,R) \leq \sum_{y\in M}\int_{\mathbb{R}^d}\left(\sum_{q\in M} m_{f(\cdot,q)}(x,R)\right)^{-\varepsilon} f(x,y)dx$$
$$\leq \sum_{y\in M}\int_{\mathbb{R}^d} m^{-\varepsilon}_{f(\cdot,y)}(x,R)f(x,y)dx = \sum_{y\in M} T_{f(\cdot,y)}(\varepsilon,R) < \infty,$$

поскольку $T_{f(\cdot,y)}(\varepsilon,R) < \infty$ при всех $y \in M$. $\square$

Теорема 3.5 *Пусть выполнены условия теоремы 3.4 и $L_\nu(f) < \infty$ для некоторого $\nu > 2$ (вместо $\nu > 1$). Тогда для всех $k \in \mathbb{N}$*

$$\mathsf{E}(\widehat{I}_{n,k}(X;Y) - I(X;Y))^2 \to 0, \;\; n \to \infty.$$

Доказательство. Условия теоремы влекут конечность функционалов $L_{f_X}(\nu)$, $Q_{f_X}(\varepsilon)$ и $T_{f_X}(\varepsilon)$. Значит, в силу теоремы 2 [28] для всех $k \in \mathbb{N}$ верно соотношение $\mathsf{E}(\widehat{H}_{n,k}(X) - H(X))^2 \to 0$ при $n \to \infty$. Из формулы, аналогичной (141), следует, что достаточно проверить для всех $y \in M$

$$\mathsf{E}\left(\widehat{H}_{n,k}(X|Y=y)(\hat{P}_n(y) - \mathsf{P}(Y=y))\right)^2 \to 0, \;\; n \to \infty, \tag{145}$$

$$\mathsf{E}\left((\widehat{H}_{n,k}(X|Y=y) - H(X|Y=y))\right)^2 \to 0, \;\; n \to \infty. \tag{146}$$

Аналогично (142) для всех $y \in M$ мы имеем

$$\mathsf{E}\left(\widehat{H}_{n,k}(X|Y=y)(\hat{P}_n(y) - \mathsf{P}(Y=y))\right)^2$$
$$= \sum_{m=k}^{n}\left(\frac{m}{n} - \mathsf{P}(Y=y)\right)^2 c_m \mathsf{P}(N_{y,n} = m),$$

где $c_m := \frac{1}{m^2}\mathsf{E}\left(\sum_{i=1}^m \eta_{m,k,i}(X_{j_1^y},\ldots,X_{j_m^y})\right)^2$, c_m также зависит от y и k. Согласно лемме 3.3 для всех $y \in M$ случайные векторы $X_{j_1^y}, X_{j_2^y}, \ldots$ независимы и распределены также как X при условии $Y = y$. Условие (144) (с $\nu > 2$) обеспечивает конечность функционалов $L_f(\nu)$, $Q_f(\varepsilon)$ и $T_f(\varepsilon)$ для

$f = f_{X|Y}(\cdot|y)$, $y \in M$. Следовательно, теорема 2 [28] гарантирует выполнение

$$d_m := \mathsf{E}\left(\frac{1}{m}\sum_{i=1}^{m}\eta_{m,k,i}(X_{j_1^y},\dots,X_{j_m^y}) - H(X|Y=y)\right)^2 \to 0, \quad m\to\infty. \tag{147}$$

Поэтому $c_m \to (H(X|Y=y))^2$ при $m\to\infty$. Значит, $0 \le c_m \le C$ для всех $m \in \mathbb{N}$ и некоторой положительной константы C. Аналогично доказательству теоремы 3.4 для всех $n > k$, $y \in M$, $\delta \in (0,p)$, где $p = \mathsf{P}(Y=y)$ мы имеем

$$\sum_{m=k}^{n}\Big(\frac{m}{n}-p\Big)^2 c_m \mathsf{P}(N_{y,n}=m) \le C(\delta^2 + 2\exp\{-2n\delta^2\}). \tag{148}$$

Таким образом, мы приходим к соотношению (145). Похожим образом для (142) можно установить, что

$$\mathsf{E}\Big((\widehat{H}_{n,k}(X|Y=y) - H(X|Y=y))\Big)^2 = \sum_{m=k}^{n} d_m \mathsf{P}(N_{y,n}=m) + (H(X|Y=y))^2 \sum_{m=0}^{k-1}\mathsf{P}(N_{y,n}=m).$$

Соотношение (147) обеспечивает существование положительной константы D такой, что $0 \le d_m \le D$ для всех $m \in \mathbb{N}$. Отсюда следует, что для всех $y \in M$, любого $\delta \in (0,p)$, где $p = \mathsf{P}(Y=y)$, и всех достаточно больших n (т.е. $np - \delta n > k$)

$$\sum_{m=k}^{n} d_m \mathsf{P}(N_{y,n}=m) \le D\mathsf{P}(|N_{y,n}-np|>\delta n) + \sup_{m>np-\delta n} d_m + (H(X|Y=y))^2\mathsf{P}(|N_{y,n}-np|>\delta n).$$

Последнее неравенство и (148) влекут выполнение (146). Теорема полностью доказана. □

3.3 Сравнение оценок взаимной информации

Рассмотрим пять методов статистического оценивания взаимной информации по формулам (115), (118) и (119) (также включающих (120)

— (125)). Воспользуемся статистиками Козаченко-Леоненко для оценивания $H(X)$ или, более точно, их обобщением на случай k ближайших соседей, предложенным в работе [60]. Асимптотическая несмещенность и L_2-состоятельность таких оценок была установлена в работе [28]. Оценки энтропии $H(Y)$ методом «plug-in» имеют вид

$$\widehat{H}_n(Y) := -\sum_{y\in M} \widehat{\mathsf{P}}_n(y)\log \widehat{\mathsf{P}}_n(y), \tag{149}$$

где $\widehat{\mathsf{P}}_n(y)$ введено в (126). Таким образом, следуя методу, предложенному в работе [32], для фиксированного $k \in \mathbb{N}$ и для $n \in \mathbb{N}$, мы определяем $\widehat{I}_{n,k}(X,Y)$ с помощью (127), а $\widehat{H}_n(Y)$ задается формулой (149).

В силу уравнения (120) и результатов главы 2 мы можем ввести оценку

$$\widehat{I}^{(1)}_{n,k}(X,Y) = \widehat{H}_n(Y) - \widehat{H}_{n,k}(Y|X), \tag{150}$$

где

$$\widehat{H}_{n,k}(Y|X) = \frac{1}{n}\sum_{i=1}^{n} \widehat{H}_{n,k,i}(Y|X).$$

Здесь для $n \in \mathbb{N}$, $n > 1$, $k = k(n) \in \{1,\dots,n-1\}$ и $Z_i = (X_i, Y_i)$, $i \in \mathbb{N}$,

$$\widehat{H}_{n,k,i}(Y|X) = -\log(\xi_{n,k,i}(Z_1,\dots,Z_n)+1) + \log k,$$

$$\begin{aligned}&\xi_{n,k,i}(Z_1,\dots,Z_n)\\ &:= \sharp\{j \in \{1,\dots,n\}\setminus\{i\}\colon Y_j = Y_i, \|X_i - X_j\| \le \rho_{n,k,i}(X_1,\dots,X_n)\}.\end{aligned} \tag{151}$$

Напомним, что $\sharp$ обозначает мощность конечного множества, $\|\cdot\|$ — евклидову норму в пространстве $\mathbb{R}^d$ и $\rho_{n,k,i}(X_1,\dots,X_n)$ определяется в (130). Очевидно, что случайная величина $\xi_{n,k,i}(Z_1,\dots,Z_n)$ принимает значения $0,1,\dots,k$. Заметим, что с вероятностью 1 все точки $X_1,\dots,X_n$ попарно не совпадают, поскольку X имеет плотность относительно меры Лебега.

Таким образом, как было объяснено в главе 2, в отличие от широко известных оценок Козаченко-Леоненко безусловной энтропии Шеннона случайного вектора (или их обобщения, определяемого уравнением (129)) ключевую роль здесь играют случайные величины $\xi_{n,k,i}$, $i = 1,\dots,n$. Более точно, сначала мы получим выражение для случайного множества $J \subset \{1,\dots,n\}$, состоящего из всех таких индексов $j \in \{1,\dots,n\}\setminus\{i\}$, для которых X_j принадлежит шару $B(X_i, \rho_{n,k,i})$ со случайным центром и случайным радиусом. Затем из набора наблюдений $\{(X_j, Y_j), j \in J\}$ выберем

$\{(X_j, Y_j), j \in I_i\}$, т.е. $I_i := \{j \in J : Y_j = Y_i\}$, где I_i также случайное множество. Мощность такого множества I_i, т.е. $\sharp I_i$ равняется значению случайное величины $\xi_{n,k,i}$, $i = 1, \ldots, n$.

Чтобы сформулировать асимптотический результат для $\widehat{I}^{(1)}_{n,k_n}(X, Y)$, где $k_n \to \infty$ при $n \to \infty$, мы будем использовать некоторые условия на гладкость вероятностной плотности, а именно, будем предполагать, что функция $g\colon \mathbb{R}^d \to \mathbb{R}$ является C_0-стягиваемой.

Теорема 3.6 *Пусть для всех $y \in M$ функция $f_{X,Y}(\cdot, y)$ является C_0-стягиваемой и строго положительной. Предположим, что для $f = f_X$ и некоторого $\varepsilon > 0$ выполнено*

$$\mathsf{E}|\log f(X)|^{1+\varepsilon} < \infty. \tag{152}$$

Тогда для всех $\alpha \in (0, 1)$ и $k_n \propto n^\alpha$ имеет место следующее соотношение:

$$\mathsf{E}\widehat{I}^{(1)}_{n,k_n}(X, Y) \to I(X, Y), \quad n \to \infty,$$

т.е. оценка $\widehat{I}^{(1)}_{n,k_n}(X, Y)$ является асимптотически несмещенной.

Теорема 3.7 *Пусть условие* (152) *теоремы 3.6 заменяется следующим. Предположим, что для $\varepsilon > 0$*

$$\mathsf{E}|\log f(X)|^{2+\varepsilon} < \infty.$$

Тогда для всех $\alpha \in (0, 1)$ и $k_n \propto n^\alpha$ справедливо соотношение

$$\mathsf{E}(\widehat{I}^{(1)}_{n,k_n}(X, Y) - I(X, Y))^2 \to 0, \quad n \to \infty,$$

т.е. оценка $\widehat{I}^{(1)}_{n,k_n}(X, Y)$ является L_2-состоятельной.

Доказательство. Нетрудно показать, что $\widehat{H}_n(Y) \to H(Y)$ в L_2 при $n \to \infty$. Теоремы 2.3 и 2.4 обеспечивают асимптотическую несмещенность и L_2-состоятельность оценок $\widehat{H}_{n,k_n}$ условной энтропии $H(Y|X)$ при $n \to \infty$, когда выполнены условия теорем 3.6 и 3.7, соответственно. Следовательно, мы приходим к искомым соотношениям. $\square$

Также введем модификации оценок $\widehat{I}^{(1)}_{n,k_n}(X, Y)$. Пусть оценки $\widehat{I}^{(2)}_{n,k_n}(X, Y)$ и $\widehat{I}^{(3)}_{n,k_n}(X, Y)$ определены формулами (150) — (151), где вместо $\widehat{H}_{n,k,i}(Y|X)$ используются

$$\widetilde{H}_{n,k,i}(Y|X) := -\log(\xi_{n,k,i}(Z_1, \ldots, Z_n) + 1) + \log(k + 1) \tag{153}$$

и

$$\overline{H}_{n,k,i}(Y|X) := -\log(\xi_{n,k,i}(Z_1,\dots,Z_n)+1)+\psi(k), \tag{154}$$

соответственно. В последней формуле $\psi(k)$ обозначает дигамма функцию. Формула (153) позволяет установить соотношение

$$\frac{\xi_{n,k,i}+1}{k+1} \in [0,1].$$

Использование $\psi(k)$ в (154) естественно для метода ближайших соседей (см, напр., [28] и ссылки в ней). Эти модификации не меняют асимптотических свойств оценок, поскольку $|\widehat{I}^{(2)}_{n,k}(X,Y) - \widehat{I}^{(1)}_{n,k}(X,Y)| = \log\left(\frac{k+1}{k}\right) \to 0$ и $|\widehat{I}^{(3)}_{n,k}(X,Y) - \widehat{I}^{(1)}_{n,k}(X,Y)| = |\log k - \psi(k)| \to 0$, при $k = k_n \to \infty$, $n \to \infty$. Следовательно, теоремы 3.6 и 3.7 остаются верными и для $\widehat{I}^{(2)}_{n,k_n}(X,Y)$ и $\widehat{I}^{(3)}_{n,k_n}(X,Y)$, $n \in \mathbb{N}$.

Рассмотрим

$$\overline{\rho}_{n,k,i} := \rho((X_i,Y_i),(X_{i,(k)},Y_{i,(k)})),$$

где $\overline{\rho}((x_1,y_1),(x_2,y_2)) = \max\{\|x_1-x_2\|, \|y_1-y_2\|\}$ и $(X_{i,(k)},Y_{i,(k)})$ — расстояния до k-ближайшего соседа (X_i,Y_i) в метрике $\overline{\rho}$, $\|\cdot\|$ обозначает l^2-норму в евклидовом пространстве (X и Y принимают значения в $\mathbb{R}^d$ и $\mathbb{R}^m$, соответственно). Ясно, что $\overline{\rho}_{n,k,i} = \overline{\rho}_{n,k,i}((X_1,Y_1),\dots,(X_n,Y_n))$. Для $i = 1,\dots,n$, положим $k(i) := k$, если $\overline{\rho}_{n,k,i} > 0$, и пусть

$$k(i) := \sharp\{j \in J_{n,i} \colon \rho((X_i,Y_i),(X_j,Y_j)) = 0\}$$

иначе, где $J_{n,i} := \{1,\dots,n\}\setminus\{i\}$. Оценка взаимной информации была предложена в [44] и определяется следующим образом:

$$\hat{I}^{(4)}_{n,k}(X,Y) := \frac{1}{n}\sum_{i=1}^{n}\gamma_{n,k(i),i},$$

здесь

$$\gamma_{n,k(i),i} = \psi(k(i)) + \log n - \log(n_i+1) - \log(m_i+1),$$
$$n_i = \sharp\{j \in J_{n,i} \colon \|X_i - X_j\| \le \overline{\rho}_{n,k(i),i}\},$$
$$m_i = \sharp\{j \in J_{n,i} \colon \|Y_i - Y_j\| \le \overline{\rho}_{n,k(i),i}\}, \quad i = 1,\dots,n.$$

3.4 Сравниваемые модели

Для сравнения оценок взаимной информации мы используем широко известную модель логистической регрессии, описывающей зависимость

между вектором X и переменной отклика Y с масштабирующим параметром $c \in \mathbb{R}$. Условное распределение случайной величины Y со значениями в множестве $M = \{0, c\}$ при условии X определяется формулой

$$\mathsf{P}(Y = c|X = x) = \frac{1}{1 + \exp\{-(w, x) - b\}}, \quad x \in \mathbb{R}^d, \quad w \in \mathbb{R}^d, \quad b \in \mathbb{R}, \quad c \in \mathbb{R},$$
$$\mathsf{P}(Y = 0|X = x) = 1 - \mathsf{P}(Y = c|X = x),$$

где $(\cdot, \cdot)$ — скалярное произведение в пространстве $\mathbb{R}^d$. Мы будем рассматривать две возможности маргинального распределения компонент вектора X:

1. $X \sim N(\nu, \Sigma)$, $\nu \in \mathbb{R}^d$, Σ строго положительно определенная матрица размера $d \times d$,
2. $X = (X_1, \ldots, X_d)$, $\{X_i\}_{i=1}^d$ — независимые одинаково распределенные случайные величины, $X_1 \sim C_1(\nu, \gamma)$, $\nu \in \mathbb{R}$, $\gamma > 0$,

где $X_1 \sim C_1(\nu, \gamma)$ означает, что X_1 имеет плотность Коши:

$$f_{X_1}(x) = \frac{1}{\pi\gamma\left(1 + \left(\frac{x-\nu}{\gamma}\right)^2\right)}, \quad x \in \mathbb{R}.$$

Для вектора $X = (X_1, \ldots, X_d)$ мы будем писать $X \sim C_d(\nu, \gamma)$.

Положим $f(x) = f_X(x)$, $g_y(x) = f_{X,Y}(x, y)$, $x \in \mathbb{R}^d$, $y \in M$. Теперь убедимся, что оба рассматриваемых случая удовлетворяют условиям теорем 3.4 и 3.6, а значит, имеет место асимптотическая несмещенность и L_2-состоятельность оценок $\widehat{I}_{n,k}(X, Y)$ и $\widehat{I}^{(i)}_{n,k}(X, Y)$, $i = 1, 2, 3, 4$. Проверим следующие условия.

(A) $L_f(\nu) < \infty$, $G_f(\varepsilon) < \infty$, $T_{g_y}(\varepsilon) < \infty$ для всех $y \in M$ и некоторого $\nu > 2$, $\varepsilon > 0$,

(B) для всех $y \in M$, функция g_y является C_0-стягиваемой, строго положительной и для некоторого $\varepsilon > 0$, $\mathsf{E}|\log f(X)|^{2+\varepsilon} < \infty$,

(C) $\mathsf{E}\left|\log \frac{f_{X,Y}(X,Y)}{f_X(X)f_Y(Y)}\right| < \infty$ и $\sharp\{(x, y) \in \mathbb{R}^d \times M \colon P_{X,Y}(A_{x,y,0}) > 0\} < \infty$, где $A_{x,y,r} = \{(a, b) \in \mathbb{R}^d \times M \colon \|a - x\| \leq r, \|b - y\| \leq r\}$, $r \geq 0$.

В силу теорем 3.4-3.7 условие (A) влечет указанные выше свойства оценок $\widehat{I}_{n,k}(X, Y)$. Условие (B) гарантирует наличие желаемых свойств для оценок $\widehat{I}^{(i)}_{n,k}(X, Y)$ для $i = 1, 2, 3$. Согласно теоремам 1 и 2 [44] для $\widehat{I}^{(4)}_{n,k}(X, Y)$ достаточно показать, что (C) выполняется.

Лемма 3.8 *В модели логистической регрессии с* $X \sim N(\nu, \Sigma)$ *справедливы условия (A), (B) и (C).*

Лемма 3.9 *В модели логистической регрессии с* $X \sim C_d(\nu, \gamma)$ *справедливы условия (A), (B) и (C).*

Доказательства лемм приводятся в разделе 3.6.

3.5 Результаты компьютерных симуляций

Для сравнения пяти способов оценивания взаимной информации, описанных выше, были проведены численные эксперименты для модели логистической регрессии с различными комбинациями параметров. Для сравнения качества различных методом применяется следующая процедура.

1. Для $n, N \in \mathbb{N}$ независимо генерируются выборки $\{(X_{i,j}, Y_{i,j})\}_{i=1}^{n}$, $j = 1, \ldots, N$, состоящие из независимых одинаково распределенных случайных векторов, имеющих плотность $f_{X,Y}(x, y)$, которая вводилась формулой (114).

2. Для некоторого целого числа $n > 1$ определим множества $K_{1,n} = \{1, 2, \ldots, \min\{10, n-1\}\}$ и

$$K_{2,n} = \begin{cases} \{k = 5q, q \in \mathbb{N}\colon 15 \le k < n\}, & \text{если } n > 15, \\ \varnothing, & \text{иначе.} \end{cases}$$

   Тогда для всех $k \in K_{1,n} \cup K_{2,n} \cup \{n-1\}$ вычислим $\widehat{I}_{n,k,j}(X, Y)$ и $\widehat{I}^{(r)}_{n,k,j}(X, Y)$, $r = 1, 2, 3, 4$, где индекс j означает, что оценка построена по выборке $\{(X_{i,j}, Y_{i,j})\}_{i=1}^{n}$. Иными словами, мы вычисляем значения оценок для всех возможных значений k, не превышающих 10, затем для всех значений с шагом 5 и для максимально возможного значения $k = n - 1$.

3. Вычислим оцененные значения среднеквадратичной ошибки:

$$EMSE_{n,k,N} = \frac{1}{N} \sum_{j=1}^{N} (I(X;Y) - \widehat{I}_{n,k,j}(X, Y))^2,$$

$$EMSE^{r}_{n,k,N} = \frac{1}{N} \sum_{j=1}^{N} (I(X;Y) - \widehat{I}^{(r)}_{n,k,j}(X, Y))^2, \quad r \in \{1, 2, 3, 4\}.$$

   Очевидно, что $EMSE_{n,k,N}$ и $EMSE^{r}_{n,k,N}$, $r = 1, 2, 3, 4$ зависят от $\{(X_{i,j}, Y_{i,j})\}_{i=1}^{n}$, $j = 1, \ldots, N$.

Рисунок	d	X	w	b	c	n	I(X;Y)
1	1	$N(0,1)$	1	0	1	$\{100, 500\}$	0.0937
2	2	$N(0, I_2)$	$(1,1)$	0	1	$\{100, 500\}$	0.1536
3	1	$C_1(0,1)$	1	0	1	$\{100, 500\}$	0.2269
4	2	$C_2(0,1)$	$(1,1)$	0	1	$\{100, 500\}$	0.6931
5	2	$N(0, I_2)$	$(1,\ 1)$	0	$\{0.1, 10\}$	500	0.1536

Таблица 2: Комбинации параметров и соответствующие значения взаимной информации.

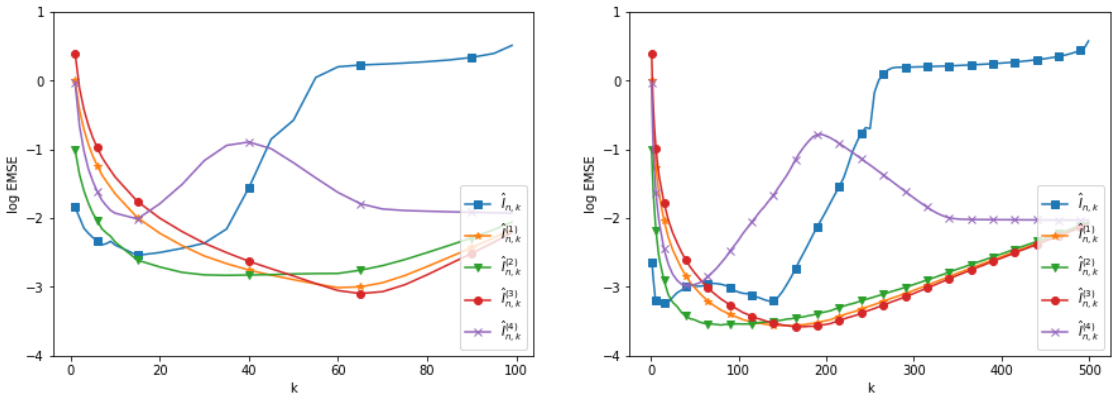

Рис. 3: Слева: $n = 100$. Справа: $n = 500$.

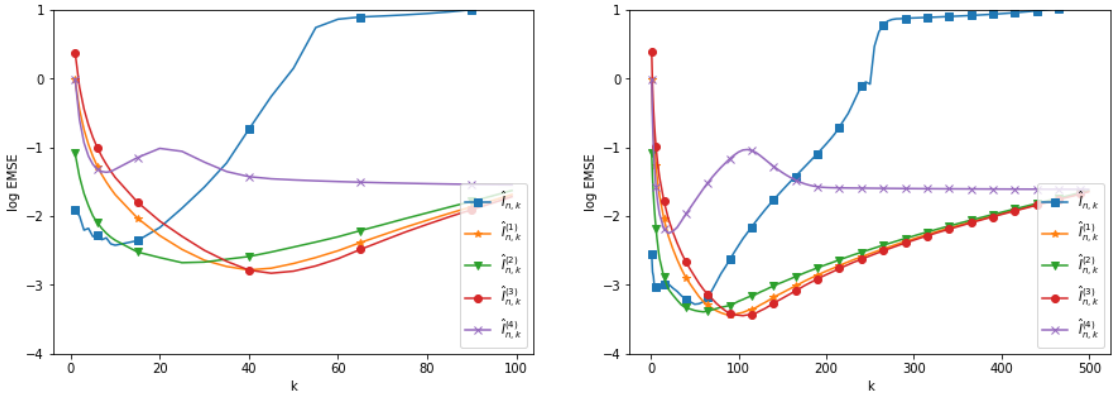

Рис. 4: Слева: $n = 100$. Справа: $n = 500$.

Значения параметров для модели приводятся в таблице 2. I_2 означает единичную матрицу размера 2×2. Каждый раз мы полагаем $N = 100$. В симуляциях, которым соответствуют рисунки 3—6, для указанных комбинаций параметров мы варьируем величину выборки, полагая $n = 100$ и $n = 500$. В последней группе симуляций (см. рисунок 7) мы полагаем $n = 500$, но при этом варьируем масштабирующий параметр c. Значения

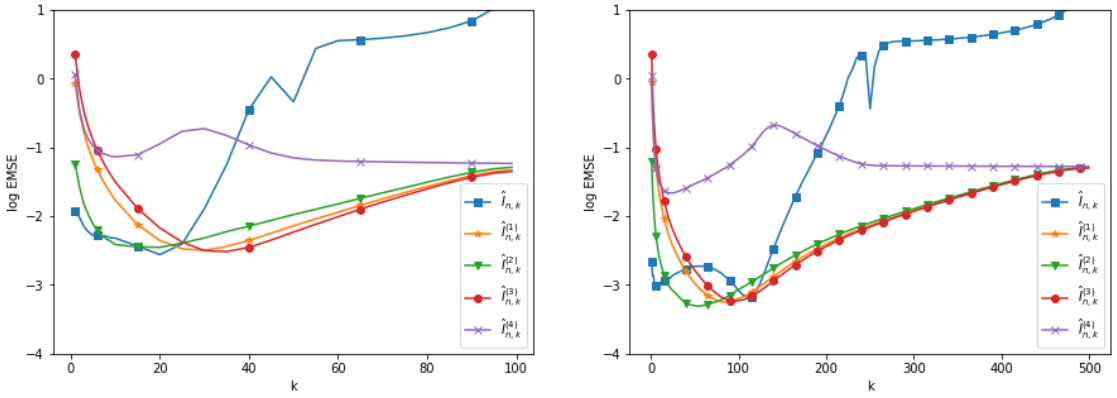

Рис. 5: Слева: $n = 100$. Справа: $n = 500$.

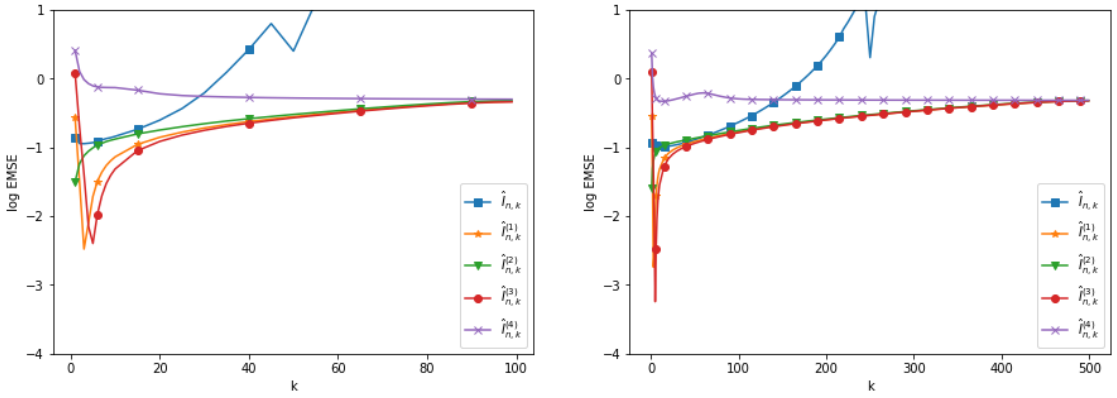

Рис. 6: Слева: $n = 100$. Справа: $n = 500$.

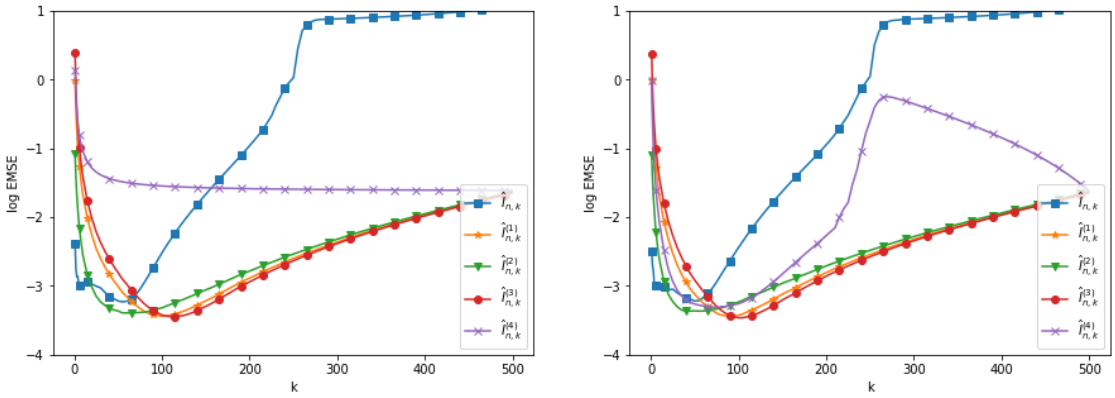

Рис. 7: Слева: $c = 0.1$, Справа: $c = 10$.

$I(X, Y)$ получены по формуле (119), где $H(Y) = \log 2$, поскольку $b = 0$, а значит, $\mathsf{P}(Y = c) = \mathsf{P}(Y = 0) = \frac{1}{2}$. Значение $H(Y|X)$ для каждой комбинации параметров и фиксированной абсолютной ошибки было вычислено численно методом nquad, реализованным в Python библиотеке scipy (основанной на Fortran библиотеке QUADPACK, см., например, [40]). Абсо-

лютные ошибки не превышают 10^{-7}. Отметим, что значения параметров n и c не влияют на значение взаимной информации. Мы применяем $\log_{10}$ вместо $\log_e$ для того, чтобы показать порядок EMSE (поскольку ошибки отображаются в десятичной логарифмической шкале).

Рисунки 3—7 показывают, что во всех рассматриваемых случаях минимальное значение EMSE достигается на оценке $I^{(3)}_{k,n}$ при некотором значении k. Это значение k зависит от комбинации параметров и n. Например, для первой схемы минимум EMSE достигается для $I^{(3)}_{k,n}$ при $k = 65$, если $n = 100$, и при $k = 175$, если $n = 500$. Для четвертой схемы $I^{(3)}_{k,n}$ также показывает лучшее качество при $k = 5$ для $n = 100$ или $n = 500$. Такой эффект обусловлен типом распределения вектора предикторов. Точнее, в первой схеме $\{X_i\}_{i=1}^n$ располагаются в основном очень плотно, поэтому для оценивания плотности в точках наблюдений мы можем брать очень большое количество соседей. В четвертой схеме рассматривается распределение Коши с тяжелыми хвостами, поэтому увеличение значения k не приводит к более точному оцениванию плотности. Отметим, что согласно результатам экспериментов, приведенным на рисунках 3—7, стоит использовать $I^{(3)}_{k,n}$ для $k \geq 0.4n$ во всех рассматриваемых случаях.

Эксперименты с различными масштабирующими параметрами (см. рисунок 7 и правую часть рисунка 4) показывают, что только $I^{(4)}_{n,k}$ чувствительна к изменению масштаба переменной Y. Эта оценка более точна в смысле EMSE при $c = 0.1$, чем при $c = 10$. Как было отмечено ранее, взаимная информация не зависит от взаимно однозначных отображений переменной Y, поэтому инвариантность ее оценок относительно таких преобразований также является желаемым свойством.

3.6 Вспомогательные результаты

Доказательство леммы 3.1. Формула (133) становится очевидной, если рассмотреть схему экспериментов Бернулли с вероятностью успеха $\mathsf{P}(Y = y)$. Имеем $\{N_{y,n} = 0\} = \{Y_q \neq y, q = 1, \ldots, n\}$ и

$$\begin{aligned}\{N_{y,n} = k\} &= \cup_I \{j^y_1 = i_1, \ldots, j^y_k = i_k\} \\ &= \cup_I \left(\{Y_i = y, i \in I\} \cap \{Y_q \neq y, q \in \{1, \ldots, n\} \setminus I\}\right) \quad (155)\end{aligned}$$

для $k = 1, \ldots, n$, где $\cup_I$ берется по множеству индексов $I = \{i_1, \ldots, i_k\}$ таких, что $1 \leq i_1 < \ldots < i_k \leq n$.

Таким образом, для $m \in \mathbb{N}$, $m > k$ и любых $B_i \in \mathcal{B}(\mathbb{R}^d)$, $i = 1, \dots, m$,

$$
\begin{aligned}
&\mathsf{P}(X_{j_1^y} \in B_1, \dots, X_{j_m^y} \in B_m, N_{y,n} = k) \\
&= \sum_{\substack{J=\{i_1,\dots,i_k,i_{k+1},\dots,i_m\}: \\ 1\le i_1<\dots<i_k\le n<i_{k+1}<\dots<i_m}} \mathsf{P}(X_{i_1} \in B_1, \dots, X_{i_m} \in B_m, j_1^y = i_1, \dots, j_m^y = i_m) = \\
&\qquad = \sum_{p=n+m-k}^{\infty} \sum_{J} \mathsf{P}(X_{i_1} \in B_1, \dots, X_{i_m} \in B_m, j_1^y = i_1, \dots, j_m^y = i_m)
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{p=n+m-k}^{\infty} \sum_{J} \mathsf{P}(X_{i_1}\in B_1,\dots, X_{i_m}\in B_m, \{Y_i=y, i\in J\}, \{Y_q\neq y, q\in\{1,\dots,p\}\setminus J\}) \\
&\qquad = \sum_{p=n+m-k}^{\infty} \sum_{J} \prod_{i=1}^{m} \mathsf{P}(X \in B_i, Y = y) \prod_{q\in\{1,\dots,p\}\setminus J} \mathsf{P}(Y \neq y) \\
&\qquad\qquad = \prod_{i=1}^{m} \mathsf{P}(X \in B_i, Y = y) \sum_{p=n+m-k}^{\infty} \sum_{J} \mathsf{P}(Y \neq y)^{p-m},
\end{aligned}
$$

где для $p \ge n + m - k$ $\sum_J$ обозначает сумму по множеству J, имеющему вид

$$
\begin{aligned}
&J(p, n, m, k) \\
&= \{\{i_1, \dots, i_k, i_{k+1}, \dots, i_m\} \colon 1 \le i_1 < \dots < i_k \le n < i_{k+1} < \dots < i_m = p\}.
\end{aligned} \tag{156}
$$

Для $k=0$ сумма $\sum_J$ берется по индексам $i_1, \dots, i_m$ таким, что $n < i_1 < \dots < i_m$. Если $J = \{i_1, \dots, i_m\} \subset \{1, \dots, p\}$, где $i_m = p$, тогда $\{j_1^y = i_1, \dots, j_m^y = p\} = \{Y_i = y, i \in J\} \cap \{Y_q \neq y, q \in \{1, \dots, p\} \setminus J\}$. Для рассматриваемых p, n, m и k мощность всех множеств $J(p, n, m, k)$ равна $\binom{n}{k}\binom{p-n-1}{m-k-1}$. Следовательно,

$$
\begin{aligned}
&\mathsf{P}(X_{j_1^y} \in B_1, \dots, X_{j_m^y} \in B_m, N_{y,n} = k) \\
&\qquad = \binom{n}{k} \prod_{i=1}^{m} \mathsf{P}(X \in B_i, Y = y) \sum_{p=n+m-k}^{\infty} \binom{p-n-1}{m-k-1} \mathsf{P}(Y \neq y)^{p-m},
\end{aligned}
$$

где учитывается тот факт, что $(X_1, Y_1), (X_2, Y_2), \dots$ являются независимыми одинаково распределенными случайными векторами, имеющими такое

же распределение, как и (X, Y). Положим $l = p - (n + m - k)$. Тогда

$$\sum_{p=n+m-k}^{\infty} \binom{p-n-1}{m-k-1} \mathsf{P}(Y \neq y)^{p-m} = \sum_{l=0}^{\infty} \binom{l+m-k-1}{m-k-1} \mathsf{P}(Y \neq y)^{l+n-k}$$

$$= \mathsf{P}(Y \neq y)^{n-k} \mathsf{P}(Y = y)^{-(m-k)} \sum_{l=0}^{\infty} \binom{l+m-k-1}{l} \mathsf{P}(Y \neq y)^{l} \mathsf{P}(Y = y)^{m-k}$$

$$= \mathsf{P}(Y \neq y)^{n-k} \mathsf{P}(Y = y)^{k} \mathsf{P}(Y = y)^{-m},$$

поскольку $\sum_{l=0}^{\infty} \binom{l+m-k-1}{l} \mathsf{P}(Y \neq y)^{l} \mathsf{P}(Y = y)^{m-k} = 1$ для отрицательного биномиального распределения. Таким образом,

$$\mathsf{P}(X_{j_1^y} \in B_1, \ldots, X_{j_m^y} \in B_m, N_{y,n} = k)$$

$$= \binom{n}{k} \mathsf{P}(Y = y)^{k} \mathsf{P}(Y \neq y)^{n-k} \prod_{i=1}^{m} \mathsf{P}(X \in B_i | Y = y). \quad (157)$$

Последняя формула дает утверждение леммы при $m > k$.

Пусть теперь $1 \leq m \leq k$, где $k \in \{0, \ldots, n\}$. Возможен лишь случай $k > 0$, поэтому

$$\mathsf{P}(X_{j_1^y} \in B_1, \ldots, X_{j_m^y} \in B_m, N_{y,n} = k)$$

$$= \sum_{\substack{I=\{i_1,\ldots,i_k\}: \\ 1 \leq i_1 < \ldots < i_k \leq n}} \mathsf{P}(X_{i_1} \in B_1, \ldots, X_{i_m} \in B_m, j_1^y = i_1, \ldots, j_m^y = i_m, \ldots, j_k^y = i_k)$$

$$= \binom{n}{k} \prod_{i=1}^{m} \mathsf{P}(X \in B_i, Y = y) \mathsf{P}(Y = y)^{k-m} \mathsf{P}(Y \neq y)^{n-k}$$

$$= \binom{n}{k} \mathsf{P}(Y = y)^{k} \mathsf{P}(Y \neq y)^{n-k} \prod_{i=1}^{m} \mathsf{P}(X \in B_i | Y = y).$$

Это равенство влечет выполнение всех предположений леммы в рассматриваемом случае $\square$

Доказательство леммы 3.6. Без ограничения общности мы можем положить $c = 1$. Действительно, $I(X, Y) = I(X, aY)$ для всех $a \neq 0$, поскольку для всех взаимно однозначных отображений $\phi\colon M \to T$, где $\sharp(M) = \sharp(T) = m$, выполняется равенство $I(X; \phi(Y)) = I(X, Y)$.

Условие (B) выполняется в силу следствия 2.5.

Проверим условие (C). Поскольку $P_X \ll \mu$, то для всех $x \in \mathbb{R}^d$ и $y \in M$,

$$0 \leq P_{X,Y}(A_{x,y,0}) = \mathsf{P}(X = x, Y = y) \leq \mathsf{P}(X = x) = 0.$$

Более того,

$$\mathsf{E}\left|\log\frac{f_{X,Y}(X,Y)}{f_X(X)f_Y(Y)}\right| = \sum_{y\in M}\mathsf{E}\left(\left|\log\frac{f_{X,Y}(X,Y)}{f_X(X)f_Y(Y)}\right|\Big|Y=y\right)\mathsf{P}(Y=y)$$
$$\leq \mathsf{E}\left|\log f_Y(Y)\right| + \mathsf{E}\left|\log f_{Y|X}(Y|X)\right| = H(Y) + H(Y|X) < \infty.$$

Здесь мы используем, что $0 \leq f_{Y|X}(y|x) \leq 1$ для всех $y \in M$, а μ-почти всех $x \in \mathbb{R}^d$ и $0 \leq f_Y(y) = \mathsf{P}(Y = y) \leq 1$. Таким образом, (C) выполнено.

Следствие 2.7 из [27] влечет, и что для $\nu > 2$, $\varepsilon > 0$ верны неравенства $L_f(\nu) < \infty$ и $G_f(\varepsilon) < \infty$. Остается показать, что $T_{g_y}(\varepsilon) < \infty$ для $y \in M$.

Согласно формуле (4.17) [27]

$$f_X(z) \geq f_X(x)e^{-\frac{\|z-x\|^2}{2\lambda_{min}}+(\Sigma^{-1}(\nu-x),z-x)},$$

где $\lambda_{min} > 0$ — минимальное собственной значение матрицы Σ. Немедленно получаем

$$g_1(z) \geq g_1(x)\frac{1+e^{-(w,x)-b}}{1+e^{-(w,z)-b}}e^{-\frac{\|z-x\|^2}{2\lambda_{min}}+(\Sigma^{-1}(\nu-x),z-x)},$$

$$g_0(z) \geq g_0(x)\frac{1+e^{(w,x)+b}}{1+e^{(w,z)+b}}e^{-\frac{\|z-x\|^2}{2\lambda_{min}}+(\Sigma^{-1}(\nu-x),z-x)}.$$

Вначале установим нижнюю границу для $\frac{1+e^{-(w,x)-b}}{1+e^{-(w,z)-b}}$, где $\|x-y\| \leq r$. Положим $u = -(w,x) - b$. Тогда

$$\frac{1+e^{-(w,x)-b}}{1+e^{-(w,z)-b}} \geq \frac{1+e^{-\|w\|r+u}}{1+e^u} = 1 - \frac{e^u}{1+e^u}(1-e^{-\|w\|r}) \geq e^{-\|w\|r}. \qquad (158)$$

Подобным образом получим оценку $\frac{1+e^{(w,x)+b}}{1+e^{(w,z)+b}} \geq e^{-\|w\|r}$, где $\|x-z\| \leq r$. Поэтому $\|x-z\| \leq r$, имеем

$$\int_{B(x,r)} f_{X,Y}(z,y)\,dz \geq f_{X,Y}(x,y)e^{-\|w\|r-\frac{r^2}{2\lambda_{min}}}\int_{B(0,r)} e^{(\Sigma^{-1}(\nu-x),u)}\,du$$
$$\geq f_{X,Y}(x,y)e^{-\|w\|r-\frac{r^2}{2\lambda_{min}}}\int_{B(0,r)}(1+(\Sigma^{-1}(\nu-x),u))\,du$$
$$= f_{X,Y}(x,y)e^{-\|w\|r-\frac{r^2}{2\lambda_{min}}}r^dV_d \geq f_{X,Y}(x,y)e^{-\|w\|R-\frac{R^2}{2\lambda_{min}}}r^dV_d.$$

Поэтому, для $C = e^{-\|w\|R-\frac{R^2}{2\lambda_{min}}}$, $y \in M$, $x \in \mathbb{R}^d$ и $0 < r < R$,

$$m_{f_{X,Y}}(x,r) \geq Cf_{X,Y}(x,y), \quad . \qquad (159)$$

Для всех $\varepsilon \in (0,1)$ и $y \in M$,

$$\int_{\mathbb{R}^d} f_{X,Y}^{1-\varepsilon}(x,y)\,dx \le \int_{\mathbb{R}^d} f_X^{1-\varepsilon}(x)\,dx < \infty,$$

следовательно, $T_{g_y}(\varepsilon,R) = \int_{\mathbb{R}^d} m_{f_{X,Y}}^{-\varepsilon}(x,R) f_{X,Y}(x,y)\,dx < \infty$. В силу леммы 2.5 [27] мы получаем, что $T_{g_y}(\varepsilon) := T_{g_y}(\varepsilon,\varepsilon) < \infty$ при всех достаточно малых ε. Лемма полностью доказана. $\square$

Доказательство леммы 3.7. Аналогично доказательству леммы 3.6, получим, что условие (C) выполнено. Теперь перейдем к условию (A). Очевидно, $f(x) \le \frac{1}{(\pi\gamma)^d}$ для $x \in \mathbb{R}^d$, таким образом, $Q_f(\varepsilon) < \infty$ при $\varepsilon \in (0,1)$. Докажем, что $L_f(\nu) < \infty$ для некоторого $\nu > 2$. Применим линейную замену переменных в интегральном представлении $L_f(\nu)$ и заметим, что $|\log\|y-x\| - \log\gamma|^\nu \le 2^{\nu-1}(|\log\|y-x\||^\nu + |\log\gamma|^\nu)$. Тогда получим, что $L_f(\nu) < \infty$, если

$$\int_{\mathbb{R}^{2d}} |\log\|y-x\||^\nu \prod_{i=1}^d \frac{1}{(1+x_i^2)}\frac{1}{(1+y_i^2)}\,dxdy < \infty.$$

Положим $u = \frac{y-x}{\sqrt{2}}, v = \frac{x+y}{\sqrt{2}}$. Тогда для $u = (u_1,\dots,u_d)$ и $v = (v_1,\dots,v_d)$ исследуем сходимость следующих интегралов:

$$\int_{\mathbb{R}^{2d}} \Big|\log(\sqrt{2}\|u\|)\Big|^\nu \prod_{i=1}^d \frac{1}{\left(1+\frac{(u_i-v_i)^2}{2}\right)}\frac{1}{\left(1+\frac{(u_i+v_i)^2}{2}\right)}\,dudv$$
$$= (\sqrt{2}\pi)^d \int_{\mathbb{R}^d} |\log(\sqrt{2}\|u\|)|^\nu \prod_{i=1}^d \frac{1}{u_i^2+2}du,$$

поскольку для $z \in \mathbb{R}$, $I(z) = \int_{\mathbb{R}} \frac{1}{\left(1+\frac{(z-s)^2}{2}\right)}\frac{1}{\left(1+\frac{(z+s)^2}{2}\right)}\,ds = \frac{\sqrt{2}\pi}{z^2+2}$. Введем $v_i = u_i^2$, $i = 1,\dots,d$. Итак, достаточно показать, что $J(d) = \int_{(0,\infty)^d} h_d(v)dv < \infty$, где

$$h_d(v) := |\log(v_1+\ldots+v_d)|^\nu \prod_{i=1}^d \frac{1}{v_i^{1/2}(v_i+2)}, \quad v \in (0,\infty)^d.$$

Нетрудно убедиться, что $J(1) < \infty$. Рассмотрим $d \ge 2$. Рассмотрим представление $J(d) = J_1(d) + J_2(d)$, где $J_1(d)$ и $J_2(d)$ — интегралы функции $h_d(v)$, взятые по множествам $B_1 := (0,\infty)^d \cap \{v_1+\ldots+v_d < 1\}$ и

$B_2 := (0,\infty)^d \cap \{v_1 + \ldots + v_d \geq 1\}$, соответственно. Имеем

$$J_1(d) = \int_{B_1 \cap \{v_1 < v_2 + \ldots + v_d\}} h_d(v)dv + \int_{B_1 \cap \{v_1 \geq v_2 + \ldots + vx_d\}} h_d(v)dv$$
$$= J_{1,1}(d) + J_{1,2}(d),$$

$$J_{1,1}(d) \leq \int_{v_1>0} \frac{|\log v_1|^\nu}{\sqrt{v_1}(v_1+2)} dv_1 \int_{v_2>0,\ldots,v_d>0} \prod_{j=2}^{d} \frac{1}{\sqrt{v_j}(v_j+2)}\, dv_2 \ldots dv_d < \infty,$$

$$J_{1,2}(d) \leq \int_{v_1>0} \frac{1}{\sqrt{v_1}(v_1+2)}\, dv_1 \int_{v_2>0,\ldots,v_d>0} \frac{|\log(v_2 + \ldots + v_d)|^\nu}{\prod_{j=2}^{d} \sqrt{v_j}(v_j+2)}\, dv_2 \ldots dv_d < \infty,$$

где используется математическая индукция по d. Аналогичным образом можно проверить, что $J_2(d) < \infty$. Доказательство конечности $J(d)$ для всех $d \in \mathbb{N}$ и $\nu > 2$ окончено, поэтому $L_f(\nu) < \infty$ при каждом $\nu > 2$. Теперь покажем, что для $y \in M$, $x \in \mathbb{R}^d$, $R > 0$ и $0 \leq r \leq R$, неравенство (159) выполняется. Здесь $C > 0$ не зависит от x, y и r.

В силу (158) для всех $y \in M$ и $x, z \in \mathbb{R}^d$ таких, что $\|x - z\| \leq r$ $(r \geq 0)$,

$$\frac{f_{X,Y}(z,y)}{f_{X,Y}(x,y)} \geq e^{-\|w\|r} \prod_{i=1}^{d} \frac{1 + \left(\frac{x_i - \nu)}{\gamma}\right)^2}{1 + \left(\frac{z_i - \nu)}{\gamma}\right)^2}. \tag{160}$$

Для $d \in \mathbb{N}$ и $x, z \in \mathbb{R}^d$ положим $F_d(x,z) := \prod_{i=1}^{d} \frac{1+x_i^2}{1+z_i^2}$, $x = (x, \ldots, x_d)$, $z = (z_1, \ldots, z_d)$. Теперь рассмотрим $x, z \in \mathbb{R}^d$ такие, что $\|x - z\| \leq \gamma r$. Возьмем $R < 1/\gamma$. Тогда

$$F_d(x,z) \geq \prod_{i=1}^{d} \frac{1}{1 + \frac{(\gamma r)^2}{1+x_i^2} + \frac{2x_i}{1+x_i^2}(z_i - x_i)}.$$

Очевидно, что $\frac{1}{s+1} \geq 1 - s$ для $s > -1$. Тогда получаем, что

$$F_d(x,z) \geq \prod_{i=1}^{d} (a_i - b_i(z_i - x_i)),$$

где $a_i = 1 - \frac{(\gamma r)^2}{1+x_i^2} \geq 1 - (\gamma r)^2 \geq 1 - (\gamma R)^2$, $b_i = \frac{2x_i}{1+x_i^2}$, $i = 1, \ldots, d$. Используя метод математической индукции, можно показать, что для всех $r \geq 0$, $A_i, B_i \in \mathbb{R}$, $i = 1, \ldots, d$,

$$\int_{B(0,r)} \prod_{i=1}^{d} (A_i - B_i u_i) du = V_d r^d \prod_{i=1}^{d} A_i.$$

Вернемся к (160). Если $\|x-z\| \le r \le R$, где $0 < R < 1/\gamma$, тогда

$$\int_{B(x,r)} f(z,y)\,dz \ge f(x,y)e^{-\|w\|r}V^d r^d \left(1-(\gamma R)^2\right)^d.$$

Следовательно, (159) выполняется для $C = \left(1-(\gamma R)^2\right)^d e^{-\|w\|R}$. Значит, для всех $\varepsilon \in (0,1)$ и $y \in M$

$$\int_{\mathbb{R}^d} f_{X,Y}^{1-\varepsilon}(x,y)\,dx \le \int_{\mathbb{R}^d} f_X^{1-\varepsilon}(x)\,dx < \infty,$$

и $T_{g_y}(\varepsilon,R) = \int_{\mathbb{R}^d} m_{f_{X,Y}}^{-\varepsilon}(x,R) f_{X,Y}(x,y)\,dx < \infty$.

Остается проверить условие (B). Функция $g_1(x) = \frac{1}{1+e^{-(w,x)-b}} f(x)$ строго положительна. Более того,

$$\nabla g_1(x) = \frac{f(x)}{(1+e^{-(w,x)-b})^2} e^{-(w,x)-b} w + \frac{1}{1+e^{-(w,x)-b}} \nabla f(x),$$

$$\nabla f(x) = -\frac{2}{\gamma^2} f(x) \left(\frac{x_1-\nu}{\left(1+\left(\frac{x_1-\nu}{\gamma}\right)^2\right)}, \ldots, \frac{x_d-\nu}{\left(1+\left(\frac{x_d-\nu}{\gamma}\right)^2\right)} \right),$$

$$\|\nabla g_1(x)\| \le \frac{1}{4}|f(x)|\|w\| + \frac{1}{1+e^{-(w,x)-b}}\|\nabla f(x)\| \le \frac{1}{4(\pi\gamma)^d}\|w\| + \|\nabla f(x)\|.$$

Функция $s(u) = \frac{u}{(1+u^2)^2}$ непрерывна и $\lim_{u\to\infty} s(u) = \lim_{u\to-\infty} s(u) = 0$, значит $\max_{u\in\mathbb{R}} |(u-\nu)| \left(1+((u-\nu)/\gamma)^2\right)^{-1} < \infty$. Таким образом,

$$\max_{x\in\mathbb{R}^d} \|\nabla g_1(x)\| < \infty$$

и $g_1(x)$ удовлетворяет условию Липшица с некоторой константой $C_0 > 0$ для всех $x \in \mathbb{R}^d$. Согласно замечанию 2.2 мы можем заключить, что g_1 является C_0-стягиваемой. Аналогичное рассуждение верно и для g_0.

Для независимых одинаково распределенных случайных величин $X_1, \ldots, X_d$ неравенство Минковского дает

$$\begin{aligned}
\left(\mathsf{E}|\log f(X)|^{2+\varepsilon}\right)^{1/(2+\varepsilon)} &\le \sum_{i=1}^{d} \left(\mathsf{E}\,|\log f_{X_i}(X_i)|^{2+\varepsilon}\right)^{1/(2+\varepsilon)} \\
&= d\left(\mathsf{E}\,|\log f_{X_1}(X_1)|^{2+\varepsilon}\right)^{1/(2+\varepsilon)} \\
&= d\left(\int_{\mathbb{R}} \frac{1}{\pi\left(1+x^2\right)} \left|\log \pi\gamma\left(1+x^2\right)\right|^{2+\varepsilon} dx\right)^{1/(2+\varepsilon)} < \infty.
\end{aligned}$$

Таким образом, (B) выполнено. Доказательство леммы окончено. $\square$

4 Выбор значимых факторов

При подготовке данной главы диссертации использован материал публикации [103], выполненной автором без соавторов.

4.1 Основные определения

Выбор значимых переменных (факторов) играет важную роль в различных областях (см., например, монографии [20], [61], [62] и ссылки в них). Данная глава посвящена процедуре выбора значимых факторов, основанной на информационном подходе (см., например, [95]). В рамках смешанной модели мы применяем оценки условной энтропии и взаимной информации, исследованные в данной главе, для идентификации значимых переменных.

Далее в главе мы используем следующие обозначения. Для вектора $Z = (Z_1, \ldots, Z_d)$ со значениями в $\mathbb{R}^d$ его компонентами являются $Z_1, \ldots, Z_d$. Для множества $L = \{l_1, \ldots, l_m\}$, где натуральные числа $l_1, \ldots, l_m$ таковы, что $1 \leq l_1 < \ldots < l_m \leq d$ и $m \in \{1, \ldots, d\}$, положим $Z_L := (Z_{l_1}, \ldots, Z_{l_m})$. Другими словами, мы рассматриваем подвектор вектора Z с компонентами $Z_{l_1}, \ldots, Z_{l_m}$. Совместная плотность $f_{X_L,Y}$ вектора (X_L, Y) и плотность f_{X_L} вектора X_L легко находятся по плотностям $f_{X,Y}$ и f_X.

Переменная отклика во многих ситуациях зависит не от всего набора объясняющих переменных, а лишь от некоторых компонент вектора X. Формализуем это следующим образом.

Определение 4.1 *Набор индексов $S = \{s_1, \ldots s_m\}$ и набор переменных $X_S := (X_{s_1}, \ldots, X_{s_m})$, где $1 \leq s_1 < \ldots, < s_m \leq d$, называются значимыми, если при каждом $y \in M$ и μ-почти всех $x \in \mathbb{R}^d$ для условных плотностей выполняется соотношение*

$$f_{Y|X}(y|x) = f_{Y|X_S}(y|x_S). \tag{161}$$

Условная плотность Y при условии X определяется формулой (37). Отметим, что различные подходы к определению значимости, избыточности и взаимодополняемости признаков обсуждаются, например, в [95] (см. там же список литературы).

Важную роль в задаче отбора значимых переменных играют разные меры зависимости между X и Y. Метод, рассматриваемый в статье, опирается на понятие взаимной информации между X и Y (см. (113))

Взаимная информация в смешанной модели, а также сопутствующие определения, задаются формулами (115)-(125).

4.2 Метод идентификации значимых признаков

Пусть независимые векторы (X^i, Y^i), $i \in \mathbb{N}$, имеют такое же распределение, как вектор (X, Y). Рассмотрим выборку $\zeta_n = \{(X^i, Y^i)\}_{i=1}^n$. Пусть

$$Q_m = \{L := (l_1, \ldots, l_m) : 1 \le l_1 < \ldots < l_m \le d\},$$

т.е. Q_m — набор всех подмножеств $\{1, \ldots, d\}$, содержащих ровно m элементов. Для любого $L \in Q_m$ определим $\zeta_{n,L} = \{(X_L^i, Y^i)\}_{i=1}^n$ и оценим взаимную информацию $I(X_L; Y)$ для каждой выборки $\zeta_{n,L}$. Для этого используем оценку вида (150), при построении которой вместо ζ_n берется $\zeta_{n,L}$. Будем писать $\widehat{I}_{n,k,L} := \widehat{I}_{n,k}(X_L; Y)$, где $k = k(n)$ некоторая функция, $k(n) \in \{1, \ldots, n-1\}$.

Введем набор случайных множеств

$$\widehat{\mathbb{S}}_{n,k}(\omega) = \arg\max_{L \in Q_m} \widehat{I}_{n,k,L}(\omega).$$

Таким образом, $\widehat{\mathbb{S}}_{n,k}$ — набор таких множеств $\widehat{S}_{n,k}$, что

$$\max_{L \in Q_m} \widehat{I}_{n,k,L} = \widehat{I}_{n,k,\widehat{S}_{n,k}}$$

для каждого $\widehat{S}_{n,k} \in \widehat{\mathbb{S}}_{n,k}$. Отметим, что мы можем упростить вычисления, используя оценки условной энтропии для нахождения $\widehat{\mathbb{S}}_{n,k}(\omega)$. Действительно,

$$\widehat{\mathbb{S}}_{n,k}(\omega) = \operatorname*{argmax}_{L \in Q_m} \left(\widehat{H}_n(Y) - \frac{1}{n} \sum_{i=1}^n \widehat{H}_{n,k,i}(Y | X_L^i) \right) = \operatorname*{argmin}_{L \in Q_m} \widehat{H}_{n,k}(Y | X_L),$$

поскольку $\widehat{H}_n(Y)$ не зависит от $L \in Q_m$.

4.3 Основные результаты

Теорема 4.2 *Предположим, что для некоторого $m \in \{1, \ldots, n-1\}$ существует непустое множество $\mathbb{S}_m$, состоящее из всех наборов значимых индексов S мощности $|S| = m$. Пусть имеется строго положительная версия плотности $f_{X,Y}$. Пусть также для всех $L \in Q_m$ и $y \in M$ плотность $f_{X_L,Y}(\cdot, y)$ является C_0-стягиваемой и $\mathsf{E}|\log f_{X_L}(X_L)|^{2+\varepsilon} < \infty$ для некоторого $\varepsilon > 0$. Тогда для всех $\alpha \in (0,1)$ и $k = k(n) \propto n^\alpha$,*

$$\mathsf{P}(\widehat{\mathbb{S}}_{n,k} \subset \mathbb{S}_m) \to 1 \ \text{ при } \ n \to \infty.$$

В частности, если набор $\mathbb{S}_m$ состоит из единственного множества S_m, то $\mathsf{P}(\widehat{\mathbb{S}}_{n,k} = S_m) \to 1$, $n \to \infty$.

Условие $\mathsf{E}|\log f_{X_L}(X_L)|^{2+\varepsilon}<\infty$ означает, что рассматривается $|\log f_{X_L}(u)|^{2+\varepsilon}$, вместо u подставляется X_L, а затем берется математическое ожидание. Таким образом, в данном условии нижний индекс X_L не связан с усреднением.

Доказательство. Прежде всего, покажем, что если $S \in \mathbb{S}_m$, то

$$I(X_S; Y) = \max_{L \in Q_m} I(X_L; Y). \tag{162}$$

Достаточно проверить, что

$$H(Y|X_S) = \min_{L \in Q_m} H(Y|X_L).$$

Для $x \in \mathbb{R}^d$ и $U = \{j_1, \ldots, j_q\}$, где $1 \leq j_1 < \ldots < j_q \leq d$, пишем $x_U = (x_{j_1}, \ldots, x_{j_q})$. Таким образом, для $L \in Q_m$ и $S \in \mathbb{S}_m$ имеем

$$H(Y|X_L) - H(Y|X_S) = \sum_{y \in M} \int_{\mathbb{R}^d} f_{X,Y}(x, y) \log \frac{f_{Y|X_S}(y|x_S)}{f_{Y|X_L}(y|x_L)}\, dx.$$

Последний интеграл существует, так как функция

$$f_{X,Y}(x, y) \log \frac{f_{Y|X_S}(y|x_S)}{f_{Y|X_L}(y|x_L)}$$

определена для каждого $x \in \mathbb{R}^d$ и $y \in M$. Действительно, $f_{Y|X_L}(y|x_L) > 0$ при всех $x \in \mathbb{R}^d$, $y \in M$ и $L \in Q_m$, поскольку

$$f_{X_L,Y}(x_L, y) = \int_{\mathbb{R}^m} f_{X,Y}(x, y)\, dx_{\overline{L}} = f_{Y|X_L}(y|x_L) f_{X_L}(x_L),$$

где $\overline{L} = \{1, \ldots, d\} \setminus L$. Для всех $y \in M$ функция $f_{X,Y}(\cdot, y)$ является строго положительной. Значит, $f_{X_L,Y}(x_L, y) > 0$ для каждого x_L. Следовательно,

$$f_{Y|X_L}(y|x_L) f_{X_L}(x_L) > 0$$

и

$$f_{X_L}(x_L) = \sum_{y \in M} f_{X_L,Y}(x_L, y) > 0.$$

Отсюда $f_{Y|X_L}(y|x_L) > 0$ при всех $x \in \mathbb{R}^d$, $y \in M$ и $L \in Q_m$.

Согласно (161)

$$H(Y|X_L) - H(Y|X_S) = \int_{\mathbb{R}^d} f_X(x) \sum_{y\in M} f_{Y|X}(y|x) \log \frac{f_{Y|X_S}(y|x_S)}{f_{Y|X_L}(y|x_L)}\, dx$$
$$= \int_{\mathbb{R}^d} f_X(x) \sum_{y\in M} f_{Y|X}(y|x) \log \frac{f_{Y|X}(y|x)}{f_{Y|X_L}(y|x_L)}\, dx$$
$$= \int_{\mathbb{R}^d} f_X(x) D_{KL}(\mathsf{P}_{Y|X=x}||\mathsf{P}_{Y|X_L=x_L})\, dx,$$

где дивергенция Кульбака-Лейблера $D_{KL}(\mathsf{P}_{Y|X=x}||\mathsf{P}_{Y|X_L=x_L})$ находится для вероятностных мер на $(M, 2^M)$, имеющих соответственно плотности (относительно считающей меры) $f_{Y|X}(\cdot|x)$ и $f_{Y|X_L}(\cdot|x_L)$. Дивергенция Кульбака-Лейблера неотрицательна. Значит, для всех множеств индексов $S \in \mathbb{S}_m$ и $L \in Q_m$ получаем неравенство

$$H(Y|X_S) - H(Y|X_L) \geq 0.$$

Таким образом, (162) установлено.

Рассмотрим функцию $h : G \to \mathbb{R}$, где G — конечное множество. Введем

$$T := \arg\max_{t\in G} h(t).$$

Пусть $h_n : G \to \mathbb{R}$, $n \in \mathbb{N}$ — такая последовательность случайных функций, что $h_n(t) \xrightarrow{\mathsf{P}} h(t)$ для каждого $t \in G$ при $n \to \infty$. Возьмем случайное множество

$$T_n := \arg\max_{t\in G} h_n,\ n \in \mathbb{N}.$$

Иначе говоря, для каждого $n \in \mathbb{N}$ и любого $\omega \in \Omega$ выполняется равенство

$$\max_{t\in G} h_n(t, \omega) = h_n(u, \omega)$$

при всех $u \in T_n$. Тогда $\mathsf{P}(T_n \subset T) \to 1$, $n \to \infty$. Действительно, для любого $\varepsilon > 0$ и достаточно больших n

$$\mathsf{P}\left(\bigcap_{t\in G} |h(t) - h_n(t)| < \varepsilon\right) > 1 - \varepsilon.$$

Очевидно, можно выбрать $\varepsilon < \frac{1}{2} \max_{t\in G\setminus T} |h(t) - h(u)|$, где $u \in T$. Тогда для всех достаточно больших n (с вероятностью близкой к единице) выполняется неравенство $\max_{t\in G\setminus T} h_n(t) < \max_{t\in T} h_n(t)$. Следовательно, $T_n \subset T$.

Другими словами, с вероятностью близкой к единице любой элемент T_n может рассматриваться как точка максимума функции h, когда n достаточно велико. В частности, если $|T| = 1$ (т.е. T — одноэлементное множество), то $\mathsf{P}(T_n = T) \to 1$ при $n \to \infty$.

Теперь мы можем применить это рассуждение к нашим функциям $h(L) := I(X_L, Y)$, $L \in G := Q_m$, и $h_n(L) := \widehat{I}_{n,k(n),L}$. В силу теоремы 3.7 оценка $\widehat{I}_{n,k,L}$ является L_2-состоятельной, а, значит, и состоятельной оценкой $I(X_L, Y)$ для каждого $L \in Q_m$. Доказательство завершено. □

Рассмотрим модель логистической регрессии (см, напр., [67]). Напомним, что в рамках этой модели Y и X принимают значения во множествах $M = \{0, 1\}$ и $\mathbb{R}^d$, соответственно. Кроме того, выполняются следующие равенства:

$$\mathsf{P}(Y = 1|X = x) = \frac{1}{1 + \exp\{-(w, x) - b\}}, \quad x \in \mathbb{R}^d, \quad w \in \mathbb{R}^d, \quad b \in \mathbb{R}, \tag{163}$$

$$\mathsf{P}(Y = 0|X = x) = 1 - \mathsf{P}(Y = 1|X = x), \tag{164}$$

где $(\cdot, \cdot) := (\cdot, \cdot)_d$ — скалярное произведение в $\mathbb{R}^d$. Далее пишем (u, u) вместо $(u, u)_k$, когда ясно, что $u \in \mathbb{R}^k$. Отметим, что смешанная модель является более общей по сравнению с широко используемой логистической регрессией. Следующее утверждение оказывается полезным.

Теорема 4.3 *Пусть условное распределение Y при заданном $X = (X_1, \ldots, X_d)$ описывается формулами* (163) *и* (164). *Предположим, что $X_1, \ldots, X_d$ — независимые случайные величины. Тогда множество индексов*

$$S := \{i : w_i \neq 0, 1 \leq i \leq d\}$$

является минимальным набором значимых факторов, т.е. S является значимым и если L – другое значимое множество индексов, то его мощность $|L| \geq |S|$.

Доказательство. Множество индексов $L \subset \{1, \ldots, d\}$ является значимым тогда и только тогда, когда для всех $x_L \in \mathbb{R}^k$,

$$f_{Y|X_L}(1|x_L) = f_{Y|X}(1|x). \tag{165}$$

Левая часть этого равенства может быть переписана следующим образом:

$$f_{Y|X_L}(1|x_L) = \frac{f_{X_L,Y}(x_L, 1)}{f_{X_L}(x_L)}$$

$$= \frac{\int_{\mathbb{R}^{d-|L|}} (1+\exp\{-(w_S, x_S)-b\})^{-1} f_{X_L}(x_L) f_{X_{\overline{L}}}(x_{\overline{L}})\, dx_{\overline{L}}}{f_{X_L}(x_L)}$$

$$= \int_{\mathbb{R}^{d-|L|}} \frac{f_{X_{\overline{L}}}(x_{\overline{L}})}{1+\exp\{-(w_S, x_S)-b\}}\, dx_{\overline{L}}. \tag{166}$$

Отметим, что $(w_S, x_S) = (w_{S\cap L}, x_{S\cap L}) + (w_{S\cap\overline{L}}, x_{S\cap\overline{L}})$, поэтому последнее выражение в (166) зависит только от $x_{S\cap L}$. Кроме того, нетрудно убедиться, что если $S\cap L \neq \varnothing$, то эта функция от $x_{S\cap L}$ не является константой на $\mathbb{R}^{|S\cap L|}$. Действительно, мы можем сравнить значения функции под знаком интеграла в точках $x_{S\cap L}$ и $z_{S\cap L}$ ($x, z \in \mathbb{R}^d$). Заметим, что если функция g интегрируема на измеримом множестве $A \subset \mathbb{R}^m$ относительно меры Лебега μ, а также, если $g(u) > 0$ для μ-почти всех $u \in A$ и $\mu(A) > 0$, то $\int_A g(u)\mu(du) > 0$. Понятно, что когда $S \setminus L \neq \varnothing$, функция

$$f_{Y|X}(1|x) = \frac{1}{1+\exp\{-(w,x)-b\}} = \frac{1}{1+\exp\{-(w_S, x_S)-b\}}$$

не может принимать одно и то же значение для всех $x_{S\setminus L}$. Следовательно, $S \subset L$ обеспечивает выполнение (165). Очевидно, если $L = S$, то в силу (166)

$$f_{Y|X_S}(1|x_S) = \frac{1}{1+\exp\{-(w_S, x_S)-b\}} \int_{\mathbb{R}^{d-|S|}} f_{X_{\overline{S}}} dx_{\overline{S}} = \frac{1}{1+\exp\{-(w_S, x_S)-b\}}.$$

Значит, (161) выполняется. Доказательство завершено.

4.4 Численные эксперименты

В рамках модели (163) – (164) мы предполагаем $X \sim N(0, I_d)$, где I_d — единичная матрица порядка d. Тогда в силу леммы 3.8 условия теоремы 4.2 выполнены. Для статистического оценивания $\mathsf{P}(\widehat{\mathbb{S}}_{n,k} \subset \mathbb{S}_m)$ мы реализуем следующую процедуру.

1. Выберем n, m, d и сгенерируем N ($N \in \mathbb{N}$) независимых случайных выборок $\zeta^{(N)}_{j,n,d} = \{(X^{i,j}, Y^{i,j})\}_{i=1}^n$, $j = 1, \ldots, N$, состоящих из независимых одинаково распределенных случайных пар наблюдений, распределенных как (X, Y). Положим $b = 0$ и

$$w = (w_1, \ldots, w_d), \quad w_i = \begin{cases} 1, & i \le m, \\ 0, & i > m. \end{cases}$$

Другими словами, мы рассматриваем случай, когда m первых компонент вектора w не равны 0, а остальные равны 0. Согласно теореме 2, множество $S_m = \{1, \ldots, m\}$ является единственным набором значимых индексов мощности m.

2. Для каждой выборки $\zeta^{(N)}_{j,n,d}$ и каждого подмножества индексов $L \in Q_m$ мы находим $\widehat{\mathbb{S}}_{n,k}(\zeta^{(N)}_{j,n,d})$, т.е. оцениваем набор значимых факторов для каждой сгенерированной выборки.

3. Вычислим

$$ACC^{(N)}_{n,d,k} = \widehat{\mathsf{P}}^{(N)}_n(\widehat{\mathbb{S}}_{n,k} \subset \mathbb{S}_m) = \frac{1}{N}\sum_{j=1}^{N} \mathbb{I}(\widehat{\mathbb{S}}_{n,k}(\zeta^{(N)}_{j,n,d}) = S_m).$$

В силу усиленного закона больших чисел можно установить, что

$$\widehat{\mathsf{P}}^{(N)}_n(\widehat{\mathbb{S}}_{n,k} = S_m) \to \mathsf{P}(\widehat{\mathbb{S}}_{n,k} = S_m) \text{ п.н., } N \to \infty.$$

Для каждой конфигурации параметров возьмем $N = 100$. На рисунках 8-10 изображены графики $ACC^{(N)}_{n,d,k}$ как функции от n для различных комбинаций параметров $d \in \{10, 50\}$, $m \in \{1, 2, 3\}$ и $k \in \{5, 10, 20, 30\}$. Эксперименты показывают, что во всех рассматриваемых случаях значение $ACC^{(N)}_{n,d,k}$ может быть равно 1 при конечных n. Таким образом, мы продемонстрировали, что даже для довольно малых значений k и умеренно больших n предложенная процедура позволяет точно идентифицировать значимые факторы.

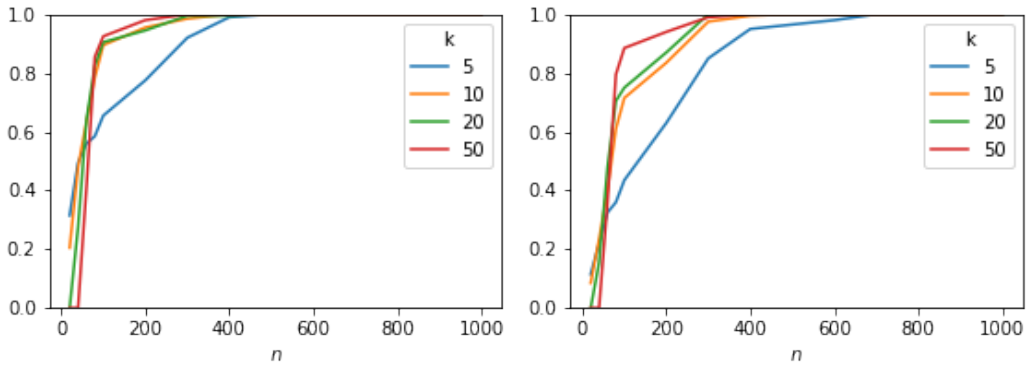

Рис. 8: $m = 1$. Слева: $d = 10$. Справа: $d = 50$

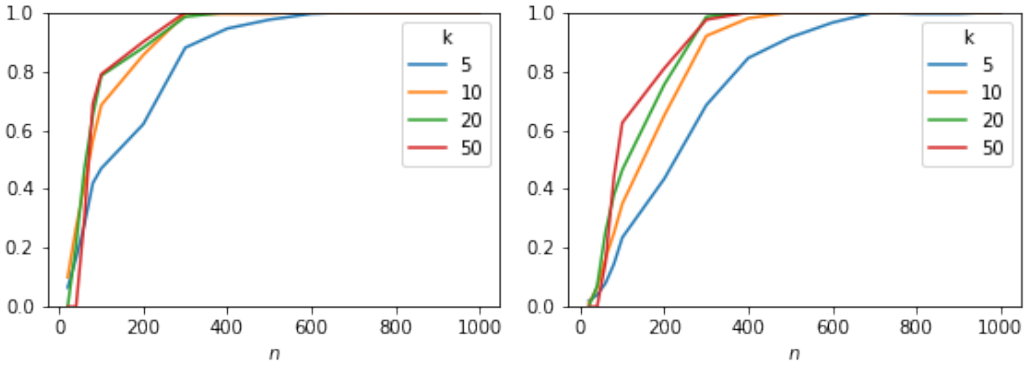

Рис. 9: $m = 2$. Слева: $d = 10$. Справа: $d = 50$

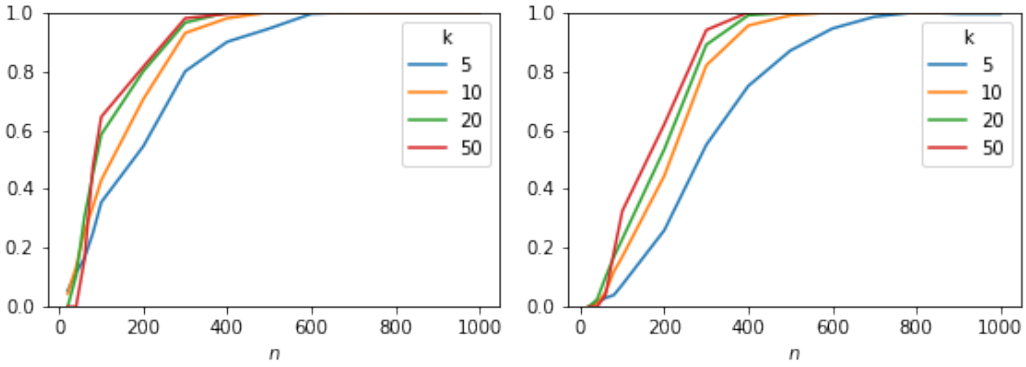

Рис. 10: $m = 3$. Слева: $d = 10$. Справа: $d = 50$

Заключение

Перечислим кратко основные результаты диссертационной работы.

Предложена модификация оценки MDR-EFE для стратифицированной выборки в модели с дискретными предикторами и дискретной переменной отклика, доказана её сильная состоятельность. Введен новый стоимостной подход, позволяющий сравнивать предложенную оценку для стратифицированной выборки с оценкой MDR-EFE для выборки из независимых одинаково распределённых случайных величин. Приведены результаты численных экспериментов по сравнению оценок в рамках предложенной обобщенной XOR-модели, продемонстрировавшие преимущества новой оценки.

Введена оценка условной энтропии дискретной случайной величины при условии случайного вектора, имеющего абсолютно непрерывное распределение. Доказана её асимптотическая несмещённость и L_2-состоятельность.

Аналогичные свойства доказаны для оценки взаимной информации, введённой в работе [32], а также для оценки взаимной информации, по-

строенной на основе оценки условной энтропии, предложенной в данной диссертационной работе.

На основе предложенных оценок условной энтропии и взаимной информациии разработана процедура отбора значимых факторов в предположении, что число значимых признаков известно, доказана состоятельность этой процедуры.

Таким образом, полученные в рамках диссертации результаты могут быть использованы для дальнейших исследований в области отбора значимых признаков и применены для решения практических задач, например, для обнаружения факторов, влияющих на возникновение некоторого заболевания.

В дальнейшем представляло бы интерес развитие изложенных методов для случая, когда количество значимых факторов заранее неизвестно. В настоящий момент в научной литературе этот вопрос мало изучен (см., например, [32]).

Список литературы

[1] А. М. Зубков, А. А. Серов. Полное доказательство универсальных неравенств для функции распределения биномиального закона. *Теория вероятн. и ее примен.*, 57(3):597–602, 2012.

[2] В.И. Богачев. *Основы теории меры. Т.1,2, 2-е изд.* НИЦ «Регулярная и хаотическая динамика», 2003.

[3] А.В. Булинский. К основам метода понижения размерности объясняющих переменных. *Учреждение Российской академии наук Санкт-Петербургское отделение Математического института им. В.А. Стеклова РАН*, 408:84–101, 2012.

[4] А.И. Галушкин. *Синтез многослойных систем распознавания образов.* Энергия, 1974.

[5] Л. Ф. Козаченко, Н. Н. Леоненко. О статистической оценке энтропии случайного вектора. *Пробл. передачи информ.*, 23:9–16, 1987.

[6] Р. Жан, М. Е. Жуковский, М. И. Исаев, И. В. Родионов. Стохастическая теория экстремумов для схемы серий зависимых случайных величин. *Успехи математических наук*, 75(5):193–194, 2020.

[7] А.Н. Ширяев. *Вероятность-1.* МЦНМО, 2017.

[8] Ю. Ю. Линке, И. С. Борисов. О построении явных оценок в задачах нелинейной регрессии. *Теория вероятн. и ее примен.*, 63(1):29—56, 2018.

[9] G. Adesso, Hall M. Datta, N., and T. Sagawa. Shannon's information theory 70 years on: applications in classical and quantum physics. *Journal of Physics A: Mathematical and Theoretical*, 52(32):1–4, 2019.

[10] S.E. Ahmed. *Penalty, Shrinkage and Pretest Strategies. Variable Selection and Estimation.* Springer, Cham, 2014.

[11] P. Alonso-Ruiz and E. Spodarev. Entropy-based inhomogeneity detection in porous media. *arXiv:1611.02241*, 2016.

[12] E. Archer, I.M. Park, and J.W. Pillow. Bayesian entropy estimation for countable discrete distributions. *Journal of Machine Learning Research*, 15:2833–2868, 2014.

[13] B. Arras and C. Houdré. Applications to sums of independent random variables. In *On Stein's Method for Infinitely Divisible Laws with Finite First Moment*, 77–88. Springer, Cham, 2019.

[14] K. Balagani and V. Phoha. On the feature selection criterion based on an approximation of multidimensional mutual information. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(7):1342–1343, 2010.

[15] J. Beirlant, E.J. Dudewicz, L. Györfi, and E.C. van der Meulen. Nonparametric entropy estimation: An overview. *International Journal of Mathematical and Statistical Sciences*, 6:17–39, 1997.

[16] L. Benguigui. The different paths to entropy. *European Journal of Physics*, 34:303–321, 2013.

[17] M. Bennasar, Y. Hicks, and R. Setchi. Feature selection using joint mutual information maximisation. *Expert Systems with Applications*, 42:8520–8532, 2014.

[18] T.B. Berrett, R.J. Samworth, and M. Yuan. Efficient multivariate entropy estimation via k-nearest neighbour distances. *arXiv:1606.00304v2*, 2016.

[19] G. Biau and L. Devroye. *Lectures of the Nearest Neighbor Method*. Springer, Cham, 2015.

[20] Bolón-Candedo, V. and Alonso-Betanzos, A. *Recent Advances in Ensembles for Feature Selection*. Springer, 2018.

[21] V. Bolón-Canedo, N. Sanchez-Maroño, and A. Alonso-Betanzos. *Feature Selection for High-Dimensional Data*. Springer, Cham, 2015.

[22] V.S. Borkar. Springer-Verlag, New York, 1995.

[23] P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data. Methods, Theory and Applications*. Springer, Heidelberg, 2011.

[24] A. Bulinski. Some statistical methods in genetics. In V. Schmidt, editor, *Stochastic Geometry, Spatial Statistics and Random Fields. Lecture Notes in Mathematics*, volume 2120, 293–320. Springer-Verlag, Berlin, 2014.

[25] A. Bulinski. Central limit theorem related to MDR-method. in: Asymptotic laws and methods in stochastics. *A volume in Hounor of Miklos Csorgo. Fields Institute Communications*, 76:113–128, 2015.

[26] A. Bulinski, O. Butkovsky, V. Sadovnichy, A. Shashkin, P. Yaskov, A. Balatskiy, Samokhodskaya L., and V. Tkachuk. Statistical methods of snp data analysis and applications. *Open Journal of Statistics*, 2(1):73–87, 2012.

[27] A. Bulinski and D. Dimitrov. Statistical estimation of the Shannon entropy. *Acta Mathematica Sinica. English series*, 35:17–46, 2019.

[28] A. Bulinski and D. Dimitrov. Statistical estimation of the Kullback - Leibler divergence. *Mathematics*, 544(9):1–36, 2021.

[29] A. Bulinski and A. Rakitko. MDR method for nonbinary response variable. *J. of Multivariate Analysis*, 135:25–42, 2015.

[30] A. Bulinski and A. Rakitko. Simulation and analytical approach to the identification of significant factors. *Commun. in Statistics. Part B: Simulation and Computation*, 44:1–23, 2015.

[31] A. Charzyńska and A. Gambin. Improvement of of the k-nn entropy estimator with applications in systems biology. *Entropy*, 18:1–19, 2015.

[32] F. Coelho, A.P. Braga, and M. Verleysen. A mutual information estimator for continuous and discrete variables applied to feature selection and classification problems. *International Journal of Computational Intelligence Systems*, 9(4):726–733, 2016.

[33] P. L. Conti and Ya. Yu. Nikitin. Rates of convergence for a class of rank tests for independence. *Zap. Nauchn. Sem. POMI*, 260:155–163, 1999.

[34] T.M. Cover and J.A. Thomas. *Elements of Information Theory, 2-nd ed.* Wiley - Interscience, New York, 1991.

[35] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods.* Cambridge University Press, 2000.

[36] A. Dehman, C. Ambroise, and P. Neuvialm. Performance of a blockwise approach in variable selection using linkage disequilibrium information. *BMC Bioinformatics*, 16:1–14, 2015.

[37] S. Delattre and N. Fournier. On the Kozachenko-Leonenko entropy estimator. *Journal of Statistical Planning and Inference*, 185:69–93, 2017.

[38] G. Doquire and M. Verleysen. A comparison of mutual information estimators for feature selection. In *Proc. of the 1st International Conference on Pattern Recognition Applications and Methods*, 176–185, 2012.

[39] D. Evans. A computationally efficient estimator for mutual information. *Proceedings of the of the Royal Society of London, Series A*, 464:1203–1215, 2008.

[40] P. Favatti, G. Lotti, and F. Romani. Algorithm 691: Improving quadpack automatic integration routines. *ACM Transactions on Mathematical Software*, 17(2):218–232, 1991.

[41] F. Fleuret. Fast binary feature selection with conditional mutual information. *Journal of Machine Learning Research*, 4:1531–1555, 2004.

[42] J. Friedman. Stochastic gradient boosting. *Computational Statistics and Data Analysis*, 38:367–378, 2002.

[43] M. Gabrié, A. Manoel, C. Luneau, J. Barbier, N. Macris, F. Krzakala, and L. Zdeborová. Entropy and mutual information in models of deep neural networks. *arXiv:1805.09785*, 2018.

[44] W. Gao, S. Kannan, S. Oh, and P. Viswanath. Estimating mutual information for discrete-continuous mixtures. In *31st Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, USA*, 1–12, 2017.

[45] A. Gola, J.M.M. John, K. van Steen, and R. König. A roadmap to multifactor dimensionality reduction methods. *Briefings in Bioinformatics*, 1–16, 2015.

[46] M. N. Goria, N. N. Leonenko, V. V. Mergel, and P. L. Novi. A new class of random vector entropy estimators and its applications in testing statistical hypotheses. *Journal of Nonparametric Statistics*, 17(3):277–297, 2005.

[47] D. Granziol, B. Ru, S. Zohren, X. Doing, M. Osborne, and S. Roberts. Meme: An accurate maximum entropy method for efficient approximations in large-scale machine learning. *Entropy*, 21(6):1–18, 2019.

[48] R. M. Gray. *Entropy and Information Theory*. Springer US, 2011.

[49] M. E. Grigoryeva, V. Yu. Korolev, and A.I. Zeifman. Limit theorems for statistics with random sample sizes. *Stability Problems for Stochastic Models*, 260:100–120, 1999.

[50] J. Gronsbell, J. Minnier, S. Yu, K. Liao, and T. Cai. Automated feature selection of predictors in electronic medical records data. *Biometrics*, 75(1):268–277, 2019.

[51] P. Hall and S.C. Morton. On the estimation of entropy. *Annals of the Institute of Statistical Mathematics*, 45:69–88, 1993.

[52] Y. Han, J. Jiao, T. Wissman, and Y. Wu. Optimal rates of entropy estimation over lipschitz balls. *arXiv:1709.06212v2*, 2017.

[53] J.M. Hilbe. *Practical Guide to Logistic Regression.* CRC Press, Boca Raton, 2015.

[54] W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58(301):13–30, 1963.

[55] K. Ishibashi, T. Iwasaki, S. Otomasa, and K. Yada. Model selection for financial statement analysis: Variable selection with data mining technique. *Procedia Computer Science*, 96:1681–1690, 2016.

[56] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning with Applications in R.* Springer Science + Business Media, New York, 2013.

[57] J. Jiao, W. Gao, and Y. Han. The nearest neighbor information estimator is adaptively near minimax rate-optimal. *arXiv:1711.08824v3*, 2017.

[58] D.G. Kleinbaum and M. Klein. *Logistic Regression. A Self-Learning Text, 3-rd ed. with contributions by E.R.Pryor.* Springer, New York, 2010.

[59] V.Yu. Korolev and Kruglov V. M. Limit theorems for random sums of independent random variables. In V.V. Kalashnikov and V.M. Zolatarev, editors, *Stability Problems for Stochastic Models. Lecture Notes in Mathematics*, volume 1546, 100–120. Springer, Berlin, Heidelberg, 1991.

[60] A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Physical review E*, 69:066138, 2004.

[61] Kuhn, M. and Johnson, K. *A Practical Approach for Predictive Models.* CRC Press, Boca Raton, 2020.

[62] Kuhn, M. and Johnson, K. *A Practical Approach for Predictive Models*. CRC Press, Boca Raton, 2020.

[63] S. Lessmann and S. Voß. Feature selection in marketing applications. In Huang. R., Q. Yang, J. Pei, J. Gama, X. Meng, and X. Li, editors, *Lecture Notes in Computer Science*. Springer, Berlin, Heidelberg, 2009.

[64] L. Lin, D. Liang, C.-C. Yeh, and Huang J.C. Novel feature selection methods to financial distress prediction. *Expert Systems with Applications*, 41:2472–2483, 2014.

[65] F. Macedo, R. Oliveira, A. Pacheco, and R. Valadas. Theoretical foundations of forward feature selection methods based on mutual information. *Neurocomputing*, 325:67–89, 2019.

[66] J-M. Marin and C. Robert. *Bayessian Essentials with R*. Springer Science + Business Media, New York, 2014.

[67] L. Massaron and A. Boschetti. *Regression Analysis with Python*. Packt Publishing Ltd., Birmingham, 2016.

[68] P. Massart. Concentration inequalities and model selection. In *Ecole d'Eté de Probbilités de Saint-Flour XXXIII - 2003*. Springer - Verlag, Berlin, 2007.

[69] E.G. Miller. A new class of entropy estimators for multidimensional densities. In *Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP03)*, 297–300. Hong Kong, China, 2003.

[70] J. Montalvão, R. Attux, and D. Silva. A pragmatic entropy and differential entropy estimator for small datasets. *Journal of Communications and Information Systems*, 29:29–36, 2014.

[71] J.H. Moore and S.M. Williams, editors. *Epistasis: Methods and Protocols. Methods in Molecular Biology*, volume 1253. Springer Science + Business Media, New York, 2015.

[72] I. Muqattash and M. Yahdi. Infinite family of approximations of the digamma function. *Mathematical and Computer Modelling*, 43:1329–1336, 2006.

[73] C. Nair, B. Prabhakar, and D. Shah. Entropy for mixtures of discrete and continuous variables. *arXiv:cs/0607075v2*, 2007.

[74] I. Nemenman, F. Shafee, and W. Bialek. Entropy and inference, revisited. In *Advances in Neural Information Processing Systems*, 471–478. MIT Press, Cambridge, MA, 2002.

[75] J. Novovičová, P. Somol, M. Haind, and P. Pudin. Conditional mutual information based feature selection for classification task. In Mery D. Ruedz, L. and J. Kittler, editors, *CIARP 2007, LNCS 4756*, 417–426. Springer-Verlag, Berlin, 2007.

[76] D. Pál, B. Póczos, and C. Szepesvári. Estimation of rényi entropy and mutual information based on generalized nearest-neighbor graphs. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems*, 1849–1857. Vancouver, British Columbia, Canada, 2010.

[77] L. Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15:1191–1253, 2003.

[78] J. Park. Independent rule in classification of multivariate binary data. *J. of Multivariate Analysis*, 100(10):2270–2286, 2009.

[79] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information criteria of maxdependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005.

[80] M.D. Penrose and J.E. Yukich. Limit theory for point processes in manifolds. *The Annals of Applied Probability*, 23:2161–2211, 2013.

[81] M. D. Ritchie, L. W. Hahn, N. Roodi, R. Bailey, W. D. Dupont, F. F. Parl, and J.H. Moore. Multifactor dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Amer. J. Human Genetics*, 69:139–147, 2001.

[82] G. Ritter. *Robust Claster Analysis and Variable Selection.* CRC Press, Boca Raton, 2015.

[83] I. Sason and S. Verdú. F-divergence inequalities. *IEEE Transactions on Information Theory*, 62:5973–6006, 2016.

[84] J. Shang, J. Zhang, Y. Sun, D. Liu, Ye D., and Y. Yin. Performance analysis of novel methods for detecting epistasis. *BMC Bioinformatics*, 12, 2011.

[85] C.E. Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, 27:379–423 and 623–656, 1948.

[86] S. Shilaskar and A. Ghatol. Feature selection for medical diagnosis : Evaluation for cardiovascular diseases. *Expert Systems with Applications*, 40(10):4146–4153, 2012.

[87] G.F. Simiqueli and M.D.V. de Resende. Entropy and mutual information in genome-wide selection: the splitting of k-fold cross-validation sets and implications for tree breeding. *Tree Genetics and Genomes*, 16(37):1–14, 2020.

[88] S. Singh and B. Pószoc. Analysis of k-nearest neighbor distances with application to entropy estimation. *arXiv:1603.08578v2*, 2016.

[89] S. Solorio-Fernández, J.A. Carrasco-Ochoa, and J.F. Martínez-Trinidad. A review of unsupervised feature selection methods. *Artificial Intelligence Review*, 53:907—948, 2020.

[90] K. Sricharan, D. Wei, and A.O. Hero. Ensemble estimators for multivariate entropy estimation. *IEEE Transactions on Information Theory*, 59:4374–4388, 2013.

[91] D. Stowell and M.D. Plumbley. Fast multidimensional entropy estimation by k-d partitioning. *IEEE Signal Processing Letters*, 16, 2009.

[92] R. L. Taylor and T.-C. Hu. Strong laws of large numbers for arrays of row-wise independent random elements. *Int. J. Math. Math. Sci.*, 10:805–814, 1987.

[93] A.B. Tsybakov and E.C. van der Meulen. Root-n consistent estimators of entropy for densities with unbounded support. *Scandinavian Journal of Statistics*.

[94] D. Velez, B. White, A. Motsinger, W. Bush, M. Ritchie, S. Williams, and J. Moore. Balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. *Genetic Epidemiology*, 31:306–315, 2007.

[95] J.R. Vergara and P.A. Estévez. A review of feature selection methods based on mutual information. *Neural Comput and Applications*, 24:175–186, 2014.

[96] M. Verleysen, F. Rossi, and D. Franëcois. Advances in feature selection with mutual information. *arXiv:0909.0635v1*, 2009.

[97] Q. Wang, S.R. Kulkarni, and S. Verdú. Divergence estimation for multidimensional densities via k-nearest-neighbor distances. *IEEE transactions on information theory*, 55(5):2392–2405, 2009.

[98] S.J. Winham, A.J. Slater, and A.A. Motsinger-Reif. A comparison of internal validation techniques for multifactor dimensionality reduction. *BMC Bioinformatics*, 11:1–16, 2010.

[99] J. Yeh. *Real Analysis: Theory of Measure and Integration (3rd edition).* World Scientific, Singapore, 2014.

Работы автора по теме диссертации:

Статьи в научных журналах Web of Science, SCOPUS, RSCI

[100] A. Bulinski and A. Kozhevin. New version of the MDR method for stratified samples. *Statistics, Optimization and Information Computing*, 5:1–18, 2017.

ИФ SJR - 0.3 / 1.09 п.л. / вклад соискателя 0.9 п.л.

А.В.Булинскому принадлежит постановка задач и общий подход к их решению, им также доказана лемма 1. Все остальные результаты доказаны А.А.Кожевиным, им проведены все компьютерные симуляции.

[101] A. Bulinski and A. Kozhevin. Statistical estimation of conditional Shannon entropy. *ESAIM: Probability and Statistics*, 23:350-386, 2019.

ИФ WoS (JIF) - 0.745 / 2.28 п.л. / вклад соискателя 1.47 п.л.

А.В.Булинскому принадлежит постановка задач и общий подход к их решению, им также доказаны леммы 3.1 и 3.3, все остальные результаты доказаны А.А.Кожевиным.

[102] A. Bulinski and A. Kozhevin. Statistical Estimation of Mutual Information for Mixed Model. *Methodology and Computing in Applied Probability*, 23:123–142, 2021.

ИФ WoS (JIF) - 1.147 / 1.25 п.л. / вклад соискателя 1 п.л.

А.В.Булинскому принадлежит постановка задач и общий подход к их решению, им также доказана лемма 1. Все остальные результаты доказаны А.А.Кожевиным, им проведены все компьютерные симуляции.

[103] A. Kozhevin. Feature selection based on estimation of mutual information *Siberian Electronic Mathematical Reports*, 18(1):720-728, 2021.

ИФ WoS (JCI) - 0.45 / 0.56 п.л. / вклад соискателя 0.56 п.л.

Тезисы конференций

[104] A. Kozhevin. Variable selection method with group overlapping. In *3rd International Workshop "Analysis, Geometry and Probability Ulm, Germany*, p.47, 2015.

[105] A. Kozhevin. Feature selection and the mutual information estimation. In *Abstracts of the 9-th International Workshop on Applied Probability 18-21 June 2018, Budapest, Hungary*, p.137–138., 2018.

[106] А. Кожевин. Информационный подход к отбору значимых признаков. *Теория вероятностей и ее применения*, 65(1):171-172, 2020 // A. Kozhevin. An information theory-based approach to feature selection. *Theory of Probability and its Applications*, 65(1):138-139, 2020.