ОТЗЫВ официального оппонента
о диссертации на соискание ученой степени
кандидата физико-математических наук
Ракитько Александра Сергеевича на тему: «Идентификация значимых факторов с помощью функционала ошибки»
по специальности 1.1.4 – «Теория вероятностей и математическая статистика»

Диссертация А.С.Ракитько посвящена разработке методов выявления факторов, которые оказывают, в определенном смысле, основное влияние на изучаемый случайный отклик. Основная задача заключается в выборе набора компонент вектора факторов $X=(X_1, \ldots, X_n)$, оказывающих существенное влияние на случайный отклик Y. Задача важна в случаях, когда поведение Y зависит от сравнительно небольшого набора факторов, называемых значимыми, а n велико (точное определение содержит формула (1.70)). Такое сокращение переменных позволяет построить интерпретируемые модели и добиться сокращения времени на обработку данных наблюдений. Сложность задачи заключается в том, что совместное распределение X и Y, как правило, неизвестно. Поэтому статистические выводы делаются на основании оценок, построенных по независимым векторам $(X^1,Y^1),\ldots(X^N,Y^N)$, распределенным как (X,Y). Идентификация значимых факторов представляет не только теоретический интерес, но важна и для разнообразных приложений. Достаточно указать на монографии C.Giraud (2015), V.Bolon-Caledo et al. (2018), M.Kuhn, K.Johnson (2019). Подчеркнем, что в целом ряде публикаций авторы указывают различные алгоритмы отбора значимых факторов, однако при этом ограничиваются компьютерным моделированием для иллюстрации предложенного метода. Важность исследований А.С.Ракитько обусловлена тем, что они направлены на доказательство теорем, позволяющих выяснить условия применимости метода MDR-EFE для решения упомянутой выше задачи. Напомним, что метод MDR (multifactor dimensionality reduction) был предложен в статье M.D.Ritchie et al. (2001) и за прошедшие годы приобрел большую популярность. Авторы этой основополагающей статьи

использовали процедуру кросс-валидации, когда часть данных исключалась из рассмотрения, а по оставшимся строилась таблица на основе должного алгоритма. Затем процедура повторялась k раз, а после этого сравнивались полученные таблицы и осуществлялся выбор заранее назначенного количества факторов. Теоремы, оправдывающие применение этого метода, авторами не устанавливались. В статье А.В.Булинского и соавторов (2012) также на основе кросс-валидации строились статистические оценки функционала ошибки приближения бинарного отклика Y с помощью $f(X_S)$, где X_S – это часть компонент вектора X с индексами $i \in S$, S – подмножество $\{1,\ldots,n\}$, а f – неслучайная функция. При этом в функционал ошибки входила произвольная штрафная функция Ψ. Вектор X принимал значения в любом конечном множестве. В этой и последующих работах были доказаны теоремы, обосновывающие развитый метод MDR-EFE (error function estimation). Основное внимание в рассматриваемой диссертации уделяется изучению более общей ситуации, когда отклик Y может быть и не бинарным, что, например, дает возможность характеризовать состояние здоровья пациента не только как «здоров» или «болен». Таким образом, тематика диссертации А.С.Ракитько представляется важной и актуальной.

Диссертация А.С.Ракитько имеет объем 110 страниц и состоит из введения, трех глав, заключения и списка литературы из 100 наименований.

Во введении четко формулируются цели исследования и дается обзор предшествующих работ по теме диссертации.

Глава 1 содержит результаты, относящиеся к исследованию небинарного отклика Y в рамках MDR-EFE метода. Несомненный интерес представляет теорема 1 (стр. 20), дающая критерий сильной состоятельности построенных автором статистических оценок функционала ошибки (прогноза Y) с помощью процедуры кросс-валидации. Доказательство потребовало изобретательности, поскольку автору надо было учесть, что предсказательный алгоритм может давать хорошее приближение неизвестной функции не на всех подмножествах пространства значений вектора X. Кроме

того, существенную роль играет применение усиленного закона больших чисел в схеме серий. Следует отметить, что автор хорошо иллюстрирует доказанный результат примерами 1 и 2 на страницах 26-29. Кроме того, он объясняет трудности, возникающие при исследовании небинарного случайного отклика, а также возможности выбора штрафной функции.

Подчеркнем, что именно сильная состоятельность введенных оценок функционала ошибки позволила доказать важную для приложений теорему 3 (стр. 33), дающую обоснование выбора значимого набора факторов на основе статистических оценок функционала ошибки. Приоритетный характер носит теорема 4 (стр. 35), в которой сильная состоятельность оценок функционала ошибки доказана для модели бинарного отклика с вектором объясняющих факторов, имеющим плотность распределения по мере Лебега в R^n. Получение этого трудного результата (доказательство занимает 10 страниц) потребовало введения вспомогательного мартингала и применение неравенства Хефдинга – Азумы. Приятно отметить, что автор обсуждает условия теоремы 3 и показывает, в частности, их выполнение для важной модели логистической регрессии. Отметим также, что отдельных усилий потребовало исследование статистических оценок условных распределений, проведенное в разделе 1.3.2.

Вторая глава посвящена изучению предельного поведения построенных оценок функционала ошибки. Глубокий результат представляет теорема 7 (стр. 52), содержащая условия асимптотической нормальности должным образом регуляризованных оценок функционала ошибки. Доказательство занимает 10 страниц. Автору удалось выделить в сложных суммах случайных величин, возникающих в результате кросс-валидации, «основую часть», дающую главный вклад в предельное поведение. Отметим, что у предельного центрированного гауссовского закона явно указано нетривиальное выражение для его дисперсии (формула (2.4)). К большим достижениям автора относится доказательство новых предельных теорем для перестановочных случайных величин. Такие системы случайных величин были введены Б. де Финетти.

Поэтому самостоятельный интерес представляет теорема о сходимости распределений нормированных максимумов частных сумм перестановочных величин (теорема 10 на стр. 66), а также новый вариант ЦПТ (лемма 5 на стр. 73), позволяющий рассмотреть массив построчно перестановочных случайных величин и доказать асимптотическую нормальность регуляризованных статистических оценок функционала ошибки (теорема 13, стр. 80). В этом разделе диссертации А.С.Ракитько удалось обобщить некоторые результаты Д.Блюма, Г.Чернова, М.Розенблата и Г.Тэйлора.

В третьей главе исследуется важный для приложений последовательный вариант метода MDR-EFE. Рассматривается субоптимальный алгоритм выбора значимых факторов, заключающийся в последовательном добавлении к отобранным факторам одного нового. В общей ситуации такой последовательный метод не обязан приводить к идентификации набора значимых факторов. Поэтому заслуживает внимания теорема 14 (стр. 86), демонстрирующая, что в модели наивного байесовского классификатора удается установить нижнюю оценку вероятности правильной идентификации значимого набора факторов в рамках MDR-EFE метода. Точнее говоря, последовательная процедура отбора факторов позволяет строить статистические оценки лишь для функционалов ошибки, отвечающих факторам с индексами, принадлежащими части подмножеств множества {1,…,n}. Это существенно ускоряет производимый отбор. Отметим, что автор сумел связать изучаемую модель с логистической регрессией (теорема 15 на стр. 88). Раздел 3.3.1 содержит программный код MDR-EFE алгоритма. Теоретические результаты в диссертации проиллюстрированы с помощью компьютерного моделирования в разделе 3.3. Здесь автор рассматривает 4 модели эпистаза в генетике и сравнивает применение логистической регрессии, MDR и MDR-EFE методов. Результаты моделирования показали успешность использования MDR-EFE метода.

В заключении (стр. 99 и 100) автор дает краткий обзор проведенного исследования и намечает возможные перспективы дальнейших исследований.

Подводя итог, можно сказать, что диссертация А.С.Ракитько представляет собой глубокое и завершенное математическое исследование, выполненное на очень высоком научном уровне. Все установленные утверждения (теоремы и леммы) излагаются с верными, полными доказательствами. Автор диссертации решил ряд сложных и актуальных задач современной математической статистики. Следует отметить его эрудицию. Список литературы включает не только прежние работы, но и работы последних лет, в том числе публикации 2023 года. А.С.Ракитько продемонстрировал искусное владение разнообразными методами исследования. Приятно, что автор снабдил изложение рядом примеров. Доказанные результаты представляют не только теоретический интерес, но также применимы для анализа реальных данных. На наш взгляд, уровень данной диссертации по широте охвата материала и сложности установленных результатов превосходит обычный уровень кандидатских диссертаций.

Диссертация основана на 10 работах автора и прошла апробацию на 10 крупных конференциях в нашей стране и за рубежом. В каждой главе указан вклад автора в решение поставленных задач при наличии совместных статей. Автореферат подробно и правильно отражает содержание диссертации.

Диссертация тщательно написана. Опечаток и стилистических ошибок нами не обнаружено. Однако имело бы смысл сделать отдельный список использованных обозначений. Возможно, в главе 1 следовало бы подробнее изложить базовый MDR метод, а также остановиться на взаимно дополняющих методах идентификации значимых факторов (которые лишь упомянуты на страницах 5 и 6). В дальнейших исследованиях было бы целесообразно обратиться к анализу устойчивости метода MDR-EFE относительно малых случайных возмущений данных наблюдений.

Указанные замечания не умаляют значимости выполненного исследования. Диссертация отвечает требованиям, установленным Московским государственным университетом имени М.В.Ломоносова к работам подобного рода. Содержание диссертации соответствует

специальности 1.1.4. «Теория вероятностей и математическая статистика» (по физико-математическим наукам), а именно следующим ее направлениям: «Непараметрическая статистика» и «Анализ статистических данных». Диссертация удовлетворяет критериям, определенным пп. 2.1-2.5 Положения о присуждении ученых степеней в Московском государственном университете имени М.В.Ломоносова, а также оформлена согласно требованиям Положения о совете по защите диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук Московского государственного университета имени М.В.Ломоносова.

Таким образом, соискатель Александр Сергеевич Ракитько заслуживает присуждения ученой степени кандидата физико-математических наук по специальности 1.1.4. «Теория вероятностей и математическая статистика».

Официальный оппонент:

кандидат физико-математических наук,
эксперт по машинному обучению
управления продвинутой аналитики и машинного обучения дирекции
инновационного развития общества с ограниченной ответственностью
«Газпромнефть – Цифровые Решения»

КОЖЕВИН Алексей Александрович

9 июня 2023 года

Контактные данные:

тел.: 7 (812) 448-24-01, доб. 41021, e-mail: kozhevin.aa@gazprom-neft.ru
Специальность, по которой официальным оппонентом
защищена диссертация:
01.01.05 – «Теория вероятностей и математическая статистика»

Адрес места работы:

196084, г. Санкт-Петербург, ул. Киевская, д.5, к.4,
Общество с ограниченной ответственностью
«Газпромнефть – Цифровые Решения» (ООО «Газпромнефть – ЦР»)
Тел.: 7(812)448-24-01, доб. 41021; e-mail: kozhevin.aa@gazprom-neft.ru

Подпись сотрудника ООО «Газпромнефть – ЦР»
А.А.Кожевина удостоверяю:

09. 06. 2023

Белехова А. С.
Пов. № Д-1797
27.10.2022