ОТЗЫВ
научного руководителя о диссертационной работе
Ракитько Александра Сергеевича
“Идентификация значимых факторов
с помощью функционала ошибки”,
представленной на соискание ученой степени
кандидата физико-математических наук по специальности 1.1.4 –
теория вероятностей и математическая статистика

Данная диссертация посвящена выявлению значимых факторов, влияющих на изучаемый случайный отклик. Такого рода задачи возникают, например, при анализе медико-биологических данных. Важная проблема заключается в том, чтобы понять, какие «поломки» в геноме человека повышают риски определенных заболеваний таких, как сердечно-сосудистые, диабет и многие другие. Сложность исследований в этой области обусловлена тем, что отдельные повреждения генома могут не представлять опасности, а существенную роль играют их некоторые комбинации. В 21 веке работы в упомянутой области развернуты в ведущих научных центрах как в нашей стране, так и за рубежом. Разрабатывается целый ряд вероятностно-статистических методов, направленных на исследование различных сложных стохастических моделей. Достаточно назвать LASSO, MDR, SCAD, логистическую и логическую регрессии, случайные леса, а также методы, использующие взаимную информацию и различные дивергенции. Кроме того, публикуется множество статьей, в которых предлагаются разнообразные алгоритмы обработки массивов данных без должного теоретического обоснования. Целью диссертационной работы является развитие MDR (multifactor dimensionality reduction) метода и доказательство новых результатов, дающих математическое обоснование предлагаемых процедур. Установленные результаты иллюстрируются с помощью компьютерного моделирования. Таким образом, тематика исследования, выполненного А.С.Ракитько, несомненно, является актуальной как в теоретическом плане, так и для приложений.

Рассматриваемая диссертация, объемом 110 страниц, состоит из введения, трех глав, заключения и списка литературы, насчитывающего 100 наименований. Основные идеи и положения работы изложены в 10 научных работах, в том числе в 4 статьях, которые опубликованы в рецензируемых научных изданиях, рекомендованных для защиты в диссертационном совете МГУ по специальности 1.1.4. Все перечисленные публикации соответствуют теме диссертации и полностью отражают ее содержание. Установленные результаты докладывались автором на 10 международных конференциях.

Перейдем к подробному рассмотрению содержания диссертации. Во введении дается обзор работ по теме диссертации и обосновывается её актуальность. Напоминается, что метод MDR был предложен в 2001 году в статье M.D.Ritchie и ее соавторов для анализа генетических данных. С тех пор появился ряд модификаций этого непараметрического метода, направленного на исследование нелинейных стохастических моделей. В статье D.Gola et al. (2016) сказано, что за период с 2001 по 2015 опубликовано более 800 статей, посвященных MDR методу и его приложениям. В 2011-2012 году в МГУ имени М.В.Ломоносова под руководством академика РАН В.А.Садовничего и академика РАН В.А.Ткачука была проведена работа по идентификации значимых факторов, влияющих на сложные заболевания. Результаты этого исследования изложены в статье A.V.Bulinski et al. (2012), в которой был предложен новый вариант MDR метода, основанный на статистических оценках функционала ошибки предсказания случайного отклика. Этот метод получил название MDR-EFE (error function estimation). Далее он был развит в работах А.В.Булинского, а также автора диссертации.

В первой главе описывается MDR-EFE метод идентификации значимых факторов с помощью функционала ошибки (учитывающего отклонение прогноза отклика от его значений с учетом штрафной функции), и он обобщается на случай небинарной функции отклика. Это обобщение является нетривиальным, оно существенно для приложений. Если бинарный отклик позволяет фиксировать, например, болен или здоров пациент, то теперь появилась возможность более детального описания уровня его здоровья. Устанавливается критерий сильной состоятельности предлагаемых оценок функционала ошибки в случае небинарной функции отклика (теорема 1).

Доказывается теорема 3, обосновывающая стратегию выбора набора значимых факторов с помощью статистических оценок функционала ошибки, вовлекающих процедуру кросс-валидации. Впервые получен результат (теорема 4) о сильной состоятельности функционала ошибки в случае объясняющих факторов, имеющих абсолютно непрерывное распределение относительно меры Лебега. Для доказательств используются вероятностные и аналитические методы. В частности, применяется мартингальная техника и вероятностные неравенства.

Во второй главе исследуются асимптотические свойства построенных оценок функционала ошибки. Доказывается центральная предельная теорема (ЦПТ) для регуляризованных оценок функционала ошибки в случае небинарной функции отклика (теорема 7). Развивается теория перестановочных случайных величин для доказательства предельных теорем. Эта теория, инициированная работами B. de Finetti (см. также J.Blum, H.Chernoff, H.Teicher), оказалась полезной в статистических исследованиях. В диссертации устанавливается аналог теоремы Эрдеша (Erdős) и Каца (Kac) для перестановочных случайных величин (теорема 10). Замечательно, что автору удалось найти явный вид предельного распределения для нормированных максимумов (формула (2.36)). Доказан новый вариант ЦПТ для перестановочных случайных величин (лемма 6), с помощью которого получен новый вариант ЦПТ для оценок функционала ошибок (теорема 12). Эти результаты могут использоваться для получения асимптотических доверительных интервалов.

В третьей главе предлагается модификация MDR-EFE метода с последовательным выбором значимых переменных. Имеется ряд работ в этом направлении, например, H.Peng et al. (2005), F.Macedo et al. (2018), M.Chowdhury (2020). В случае модели наивного байесовского классификатора в диссертации получены оценки снизу для вероятности выбора значимого набора факторов MDR-EFE методом с последовательным отбором переменных (см. теорему 14). Оказалось, что рассматриваемую автором схему можно связать с определенной логистической регрессией. Установленный результат представляет несомненный интерес, поскольку дает теоретическое обоснование для быстрого алгоритма отбора факторов.

Применение MDR-EFE метода иллюстрируется в разделе 3.3 на данных компьютерного моделирования.

В заключении приводятся основные результаты работы. Отмечаются возможности их практического применения, а также указываются направления дальнейших исследований.

Подводя итог, можно сказать следующее. Диссертация А.С. Ракитько посвящена важному направлению математической статистики, связанному с идентификацией значимых факторов. Полученные результаты могут найти применение при анализе медико-биологических данных. А.С. Ракитько преодолел целый ряд технических трудностей, продемонстрировал большие творческие способности и искусное владение математическим аппаратом современной теории вероятностей.

Диссертация написана на высоком научном уровне. Она основана на 10 научных работах А.С. Ракитько. Четыре статьи автора опубликованы в рецензируемых журналах, входящих в базы данных Web of Science и Scopus. Из них три статьи написаны совместно с научным руководителем. В этих статьях научному руководителю принадлежит постановка задач и общий подход к их решению, а также доказательство двух лемм и двух теорем, которые отмечены как в диссертации, так и в автореферате. Четвертая статья написана в соавторстве с P.Alonso-Ruiz, которой принадлежит один результат (Proposition) и следствие к одной из теорем. Остальные результаты получены А.С. Ракитько самостоятельно. Кроме того автором диссертации опубликованы две статьи без соавторов. Им же произведены все эксперименты по компьютерному моделированию. Результаты диссертации прошли всестороннюю апробацию (А.С. Ракитько докладывал их на десяти международных конференциях). Кроме того, А.С. Ракитько делал доклады на Большом семинаре кафедры теории вероятностей и аспирантском коллоквиуме по теории вероятностей, руководимыми академиком РАН, профессором А.Н. Ширяевым, неоднократно на научном семинаре “Асимптотический анализ случайных процессов и полей” под руководством доктора физико-математических наук, профессора А.В. Булинского, а также других семинарах. Приятно отметить глубокую эрудицию автора (список

литературы включает 100 наименований). Автореферат правильно отражает содержание диссертации.

Таким образом, в диссертационной работе А.С. Ракитько «Идентификация значимых факторов с помощью функционала ошибки», представленной на соискание ученой степени кандидата физико-математических наук по специальности 1.1.4 – теория вероятностей и математическая статистика, решен ряд важных и трудных задач современной теории вероятностей и математической статистики. Работа удовлетворяет всем требованиям «Положения о порядке присуждения ученых степеней» и рекомендуется к защите в диссертационном совете МГУ.011.3(01.07).

Научный руководитель
профессор кафедры теории вероятностей
Механико-математического факультета МГУ имени М.В.Ломоносова
(119991 Москва, Ленинские горы 1, МГУ, Главное здание,
механико-математический факультет, тел. +74959391423, факс +74959392090,
сайт: https://www.math.msu.ru/),
доктор физико-математических наук, профессор
(тел. +74959391403, email: alexander.bulinski@math.msu.ru)

А.В. Булинский

7 апреля 2023

Подпись профессора А.В. Булинского удостоверяю.
Декан механико-математического факультета
МГУ имени М.В.Ломоносова
член-корр. РАН, доктор физико-математических наук, профессор

А.И. Шафаревич

7 апреля 2023