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

Рассматриваемая диссертация посвящена установлению новых результатов, относящихся к идентификации значимых (в определенном смысле) факторов, оказывающих существенное влияние на исследуемую случайную переменную. Это обширное направление стало развиваться в недрах классического регрессионного анализа. Однако важная особенность заключается в том, что изучается стохастическая модель, в которой переменная (отклик) описывается с помощью некоторого неизвестного поднабора факторов, который требуется идентифицировать по имеющимся наблюдениям. Такая задача представляет несомненный интерес не только в теоретическом плане, но и для разнообразных приложений. А.С. Ракитько интересуют, главным образом, применения к анализу генетических данных, но нетрудно видеть, что такого рода задачи возникают, например, в стохастической финансовой математике и других областях, где из большого числа факторов требуется выделить (сравнительно небольшой) набор, позволяющий создать интерпретируемую модель. При этом большую роль играет сокращение объема данных и времени на их обработку. Основное внимание в диссертации уделяется развитию метода MDR (multifactor dimensionality reduction), введенного М.Д. Ритчи и соавторами (M.D.Ritchieetal. (2001)). Точнее говоря, в этой основополагающей статье был предложен определенный алгоритм выявления значимых факторов, использующий процедуру кросс-валидации (когда массив данных разбивается на попарно непересекающиеся подмножества и

для них в отдельности осуществляется определенная процедура отбора значимых факторов, а на заключительном этапе проводится окончательный выбор при сравнении полученных наборов). В упомянутой работе и многих других предлагаются алгоритмы, но не даются условия, обеспечивающие их применение. Диссертация А.С. Ракитько содержит ряд важных теорем, которые позволяют обосновать развиваемые методы. Следовательно, тематика данной диссертации является весьма актуальной.

Диссертация А.С. Ракитько, имеющая объем 110 страниц, состоит из введения, трех глав, заключения и списка цитированной литературы, содержащего 100 источников. Во введении автор дает качественный обзор предшествующих работ по теме диссертации и обосновывает актуальность выполненного им исследования. Как уже отмечено выше, основное внимание в диссертации уделяется развитию непараметрического метода MDR. Этот метод является весьма популярным. В частности, в обзорной статье D. Golaetal. (2016) написано, что с 2001 года по 2015 год опубликовано свыше 800 работ, связанных с построением различных модификаций MDR метода и их применениями. А.С. Ракитько четко объясняет целесообразность MDR-EFE метода, основанного на статистических оценках функционала ошибки прогноза отклика с помощью различных наборов факторов (название EFE отражает аббревиатуру error function estimation). При этом важной особенностью данного метода является возможность изучать не только бинарные отклики (хотя даже бинарные отклики представляют большой интерес, поскольку позволяют адекватно описывать, например, состояния пациента «болен» или «здоров»).

Глава 1 посвящена, главным образом, развитию метода MDR-EFE для анализа небинарного случайного отклика. Вводится функционал, который характеризует отклонение изучаемого отклика

от его прогноза. В определении этого функционала используется произвольная неотрицательная штрафная функция. Нетривиальная проблема заключается в том, что совместное распределение вектора факторов ($X_1, \ldots X_n$) и вектора отклика Y, как правило, неизвестно. Поэтому статистические выводы делаются на основании оценок упомянутого функционала. Вначале автор рассматривает случайные векторы, принимающие значения в дискретных множествах. Важной является теорема 1, дающая критерий сильной состоятельности введенных оценок функционала ошибки. Существенную роль в доказательстве играет применение усиленного закона больших чисел для серий независимых случайных величин, установленного T.-C. Hu, F.Moricz, R.Taylor. Отметим также, что А.С.Ракитько исследовал различные формы справедливости основного условия теоремы 1 (следствия 1 и 2). В замечании 6 хорошо объясняется различие результатов для бинарного и небинарного случайного отклика. Теорема 3 дает обоснование подходу выявления значимого набора факторов с помощью статистических оценок функционала ошибки. Теорема 4 демонстрирует состоятельность MDR-EFE метода в случае, когда отклик бинарный, а факторы не являются дискретными (вектор факторов имеет распределение, абсолютно непрерывное относительно меры Лебега. Доказательство этого результата потребовало изобретательности. Автор использовал введение вспомогательных мартингалов, неравенство Хёфдинга – Азума, аппарат условных математических ожиданий. Приятно отметить, что условия доказанной теоремы подробно обсуждаются. В частности, они проверяются для распространенной модели логистической регрессии.

В главе 2 исследуются предельные свойства введенных оценок функционала ошибки. Эта задача является трудной, поскольку изучаются суммы зависимых величин, возникающие в результате процедуры кросс-валидации. Теорема 7 показывает, что при весьма

широких условиях должным образом нормализованные и регуляризованные оценки функционала ошибки даже в случае небинарного отклика удовлетворяют центральной предельной теореме. При этом явно указана дисперсия предельного (центрированного) гауссовского закона. Это сложная теорема, ее доказательство занимает 10 страниц. К достижениям А.С.Ракитько также следует отнести вклад в теорию перестановочных величин, создание которой связано с трудами Б. де Финетти (B. DeFinetti) и его последователей. Теорема 10 представляет собой аналог классической теоремы Эрдеша – Каца (P.Erdos – M.Kac) о распределении максимума частных сумм последовательности независимых одинаково распределенных величин с нулевым средним и единичной дисперсией. Для нормированных максимумов первых n частных сумм последовательности перестановочных случайных величин теорема 10 явно описывает предельный закон, возникающий, когда n стремится к бесконечности. Пример 4 на странице 71 показывает, что установленный результат представляет интерес при моделировании рисков финансовых активов. А.С.Ракитько в разделе 2.3 диссертации доказал новый вариант центральной предельной теоремы для перестановочных величин (см., например, лемму 5). Этот результат является весьма содержательным, поскольку в отличие от предшествующих работ удалось рассмотреть случай, когда в каждой строке из k_n перестановочных величин формируется определенная нормированная сумма из m_n слагаемых, причем m_n/k_n при $n\to\infty$ сходится к числу $\alpha \in [0,1)$. Распределения упомянутых сумм (слабо) сходятся к центрированному нормальному закону с явно указанной дисперсией (в формулу для которой в качестве множителя входит $1-\alpha$). Теорема 12 представляет новый вариант центральной предельной теоремы для регуляризованных статистических оценок функционала ошибки (приближения отклика с помощью прогноза), который

установлен с помощью развитой автором техники перестановочных величин. Доказанные результаты об асимптотической нормальности используемых статистик позволяют строить приближенные доверительные интервалы для функционала ошибки.

В главе 3 установлен важный вариант MDR-EFE метода, реализуемого с помощью последовательного выбора значимых факторов. Идея последовательного отбора факторов в рамках различных методов применялась и ранее. Достаточно указать, например, на статьи H.Pengetal. (2005) и F.Macedoetal. (2018). Теорема 14 впервые для модели наивного байесовского классификатора дает нижнюю оценку вероятности последовательного отбора значимого набора факторов с помощью MDR-EFE метода. А.С. Ракитько удалось связать исследуемую задачу с рассмотрением логистической регрессии. Этот результат заслуживает внимания, поскольку дает возможность быстрой реализации указанного метода отбора факторов. Приятно отметить, что теоретические результаты иллюстрируются в разделе 3.3 диссертации данными компьютерного моделирования.

В кратком заключении отмечены основные результаты диссертации. Подчеркнуто, что выполненное исследование допускает приложения к практическому анализу данных, а также приведены некоторые направления дальнейших исследований.

Таким образом, диссертация А.С. Ракитько относится к важному и перспективному направлению современной математической статистики. Доказанные результаты могут применяться при анализе медико-биологических данных. Автор доказал важные результаты, часть из которых носит неулучшаемый характер. Полученные результаты снабжены полными корректными доказательствами. А.С.Ракитько продемонстрировал владение разнообразными математическими средствами при решении сложных и актуальных

задач. Диссертация написана на высоком научном уровне, превосходящем стандартный уровень кандидатской диссертации. Она основана на 10 научных работах автора. Следует отметить большую эрудицию автора (список цитированной литературы включает 100 источников). Подчеркнем, что у автора имеется еще целый ряд опубликованных статей, которые выходят за рамки данной диссертации и посвящены разнообразным аспектам анализа генетических данных. Результаты диссертации А.С. Ракитько докладывал на 10 международных конференциях, список которых приведен на страницах 8 и 9. Автореферат правильно отражает содержание диссертации.

Диссертация тщательно написана, поэтому не требуется делать замечаний по ее оформлению. Возможно, имело бы смысл включить в текст несколько рисунков, поясняющих процедуру кросс-валидации и иллюстрирующих классический MDR метод. В теореме 14 было бы полезно для величин вида $o(1)$ и $o(N^{-1/2})$ иметь явные оценки их модулей, позволяющие проводить расчеты для каждого N.

Вместе с тем, 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), направления исследований: «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».

С.Я. Шоргин

05 июня 2023 года

Официальный оппонент: Шоргин Сергей Яковлевич, доктор физико-математических наук по специальности 08.00.13 – «Математические и инструментальные методы экономики», профессор, главный научный сотрудник отдела «Информационные технологии управления и моделирования информационных систем» Федерального государственного учреждения «Федеральный исследовательский центр «Информатика и управление» Российской академии наук»
Тел.: +7 (916) 162-79-89
E-mail: sshorgin@ipiran.ru

Федеральное государственное учреждение «Федеральный исследовательский цент «Информатика и управление» Российской академии наук» (ФИЦ ИУ РАН)
Адрес: 119333, Москва, Вавилова, д. 44, кор. 2
http://www.frccsc.ru/
Тел: +7 (499) 135-62-60
E-mail: ipiran@ipiran.ru

Подпись Сергея Яковлевича Шоргина заверяю.

Ученый секретарь ФИЦ ИУ РАН В.Н. Захаров

06 июня 2023 года