О методах идентификации и оценки причинно-следственных связей в неэкспериментальных данных

А. М. Ченцов* , Н. И. Торопов‡

Начало 21-го века для наук о данных характеризуется появлением новых междисциплинарных областей, а также расширением общей теоретической базы и появлением статистических методов для решения новых задач. В данной статье описываются известные подходы к решению задачи статистической идентификации односторонних (причинно-следственных) связей между переменными для оценки с использованием неэкспериментальных данных и приводятся отличительные особенности статистических моделей, подходящих для этих целей.

Основной результат работы заключается в сравнении методов оценивания эффектов воздействия в случае, когда зависимости в данных являются существенно нелинейными. Для этого предложен алгоритм генерации данных при помощи свёрточных нейронных сетей, и исследуются два разных подхода к оценке — методом байесовских сетей и методом двойного машинного обучения. Показано, что оба таких подхода в рассмотренном случае дают неточные оценки индивидуальных эффектов, и приводятся рекомендации по оценке агрегированных эффектов воздействия.

Ключевые слова: статистическая идентификация, эффекты воздействия, DAG-модели, двойное машинное обучение, CATE.

Введение

Выявление и оценка причинно-следственных зависимостей является важнейшей составляющей большинства задач в области здравоохранения, общественных и поведенческих наук. Эта тема долгое время являлась периферийной для математической статистики [1] и развивалась, преимущественно, экономистами: так, например, в 2021 году Г. Имбенс и Дж. Ан-

* *Ченцов Александр Михайлович* — ассистент каф. дискретной математики ФПМИ МФТИ, e-mail: achentsov@nes.ru, ORCID: 0000-0003-0429-2664.

Chentsov Aleksandr Mikhailovich — assistant, Faculty of Applied Mathematics and Informatics of MIPT.

‡ *Торопов Никита Игоревич* — ассистент каф. дискретной математики ФПМИ МФТИ, e-mail: ntoropov@nes.ru, ORCID: 0000-0002-4274-2665.

Toropov Nikita Igorevich — assistant, Faculty of Applied Mathematics and Informatics of MIPT.

Опубликовано на условиях лицензии CC-BY 4.0
Published under CC-BY 4.0

грист получили Нобелевскую премию по экономике за развитие методов причинно-следственного анализа. Эта ситуация начала меняться в начале XXI-го века с появлением работ, исследующих теоретические основы причинно-следственной идентификации, в частности, с использованием графических моделей. В настоящее время происходит постепенный синтез подходов к оценке эффектов воздействия, разработанных в различных областях науки – эконометрики, машинного обучения и математической статистики [2]. Количество исследований, связанных с разработкой методологии и статистической оценкой эффектов воздействия с использованием неэкспериментальных данных существенно выросло за последние 10 лет, причем некоторые исследователи подчеркивают важность этой задачи для построения моделей сильного искусственного интеллекта [1].

Понятию причинно-следственной связи, несмотря на его распространённость и естественность, достаточно сложно дать формальное определение. Райхенбах [3] связывает это понятие с естественным направлением времени. Другие авторы определяют причинно-следственность как идею о том, что могло бы произойти или какие (контрфактуальные) значения могли бы быть у числовых характеристик в гипотетическом сценарии, отличном от наблюдаемого на практике. В работе Льюиса [4] такие сценарии называются возможными мирами (англ.: possible worlds), и используются для следующей формализации:

Утверждение вида "если бы A, то выполнялось бы B" верно в мире w только тогда, когда B верно в ближайшем к w мире, в котором A верно.

Вид метрики не специфицируется автором, хотя отмечается, что ближайший для некоторого заданного утверждения A мир к w можно получить, применяя соображения общего здравого смысла. Галлес и Перл [5] используют построение Льюиса для описания некоторых формальных ограничений на свойства контрфактуальных утверждений.

Большинство ранних работ по статистической оценке эффектов воздействия связаны с рандомизированными контролируемыми испытаниями, в которых объекты испытаний случайно разделяются на две или более группы, отличающиеся наличием или уровнем воздействия – что позволяет минимизировать влияние посторонних факторов и более точно оценить влияние воздействия на интересующие исследователя характеристики. В то же время, для многих практически важных задач проведение рандомизированных экспериментов недоступно, а любой набор неэкспериментальных данных (англ.: observational data) является неполным – поскольку для каждого исследуемого объекта наблюдаемым является лишь одно состояние, тогда как другие возможные лишь постулируются. Невозможность сравнения контрфактуальных характеристик напрямую

приводит к тому, что исследователям требуется дополнять данные посредством теоретических предположений, которые позволяли бы извлечь информацию из имеющихся наблюдений; в работе Холланда [6] это заключение было названо фундаментальной проблемой выводов (инференции) о причинно-следственных связях.

Далее в представленной работе приводится краткий обзор теоретических методов идентификации эффектов воздействия: модель потенциальных исходов Рубина и модели на ориентированных ациклических графах. Показан механизм, как от общих соображений о связях между переменными в исследуемой задаче во многих случаях можно прийти к статистической модели, позволяющей оценить эффекты воздействия без проведения экспериментов. В третьей главе приводится основной результат работы, заключающийся в сравнении разных методов оценивания эффектов воздействия с использованием симулированных данных с нелинейными зависимостями, и приводятся рекомендации по выбору оптимальной процедуры оценивания. Для этого предложен новый алгоритм генерации данных с помощью свёрточных нейронных сетей, а также алгоритм дискретизации таких данных с последующей оценкой условного распределения.

1. Модель потенциальных исходов Рубина

Исторически, первой из статистических моделей, применимых для идентификации и оценки эффектов воздействия при анализе неэкспериментальных данных, является модель потенциальных исходов Рубина – в честь автора, предложившего эту модель в публикации [7], – однако, некоторые элементы использованного подхода применительно к случаю рандомизированных контролируемых экспериментов были предложены ещё в 1923 году Нейманом [8].

В модели для каждого объекта рассматриваемой популяции постулируется существование спектра различных состояний, описываемых переменной воздействия. Это позволяет говорить о потенциальных исходах зависимой переменной и в этих терминах определять причинно-следственные зависимости. Когда состояний всего два, то их традиционно называют контрольным состоянием и воздействием. Если состояний больше двух, то одно из них назначается контрольным, а остальные расцениваются как различные уровни воздействия, либо как альтернативные варианты воздействий – если сравнивать их интенсивность некорректно.

На практике не всегда существует возможность специфицировать состояния объектов достаточно точно – в силу ограничений, непосред-

ственно связанных с аккуратной формулировкой самих состояний, либо из-за сложности последующего сбора данных о них. В то же время, от этого этапа построения модели зависит интерпретация оценок, а также то, насколько получаемые результаты могут быть распространены с выборки на всю моделируемую популяцию.

Под потенциальными исходами в модели Рубина понимаются числовые характеристики, возможно случайные – причём по одной величине приписывается каждой паре из объекта наблюдения и его состояния. В простейшем случае, когда состояний всего два, их можно обозначить за Y_{i1} и Y_{i0}. Иногда представляется целесообразным, особенно для рассмотрения общих понятий, сопоставить потенциальные исходы со случайным процессом $Y = Y(\omega, d)$, где $d \in \{0, 1\}$ соответствует отсутствию/наличию воздействия, а $\omega \in \Omega$ – состояние мира, отражающее случайность в выборе конкретного объекта, и случайность, связанную с любыми другими возможными причинами. Если потенциальных исходов всего два, то можно ввести индикатор воздействия $D_i : \Omega \to \{0, 1\}$, который определяет, в каком из состояний находится интересующий нас объект наблюдения. Тогда индивидуальный эффект воздействия можно определить, как разность между потенциальными исходами

$$\delta_i = Y_{i1} - Y_{i0}, \tag{1}$$

и, поскольку δ_i в такой постановке может быть случайной величиной, то практически всегда рассматривается лишь её усреднённое значение

$$\delta_i^E = \mathbb{E}Y_{i1} - \mathbb{E}Y_{i0}. \tag{2}$$

Заметим, что нет оснований определять эффект воздействия исключительно как разность двух величин. Например, иногда исследователя может интересовать процентное различие между двумя величинами, и тогда можно рассматривать эффект воздействия как частное $\left(\frac{Y_{i1}}{Y_{i0}} - 1\right) \times 100\%$; также возможны и другие функциональные формы.

Если ввести переменную Y_i, обозначающую фактически наблюдаемый потенциальный исход для i-го объекта, то для контрольной группы данная переменная будет принимать одно из возможных значений Y_{i0}, а для экспериментальной – одно из возможных значений Y_{i1}. Формально это можно записать в виде уравнения (3):

$$Y_i = D_i Y_{i1} + (1 - D_i) Y_{i0} \tag{3}$$

Одним из возможных решений фундаментальной проблемы инференции является агрегация эффектов воздействия для элементов популяции

в один средний эффект, при этом можно попытаться использовать информацию об объектах из экспериментальной группы для выводов о $\mathbb{E}Y_{i1}$ и, аналогично, объекты из контрольной группы можно использовать для оценки популяционного значения $\mathbb{E}Y_{i0}$.

Рубин в своей работе ввёл идентификационное предположение об устойчивости индивидуальной величины воздействия – набор требований, достаточных для того, чтобы считать среднее по подвыборкам в группе воздействия и в контрольной группе несмещёнными оценками соответствующих популяционных значений.

Устойчивость индивидуальной величины воздействия:

1. Эффект воздействия на один элемент популяции не зависит от значений воздействия у других элементов популяции;
2. Эффект воздействия на один элемент популяции не зависит от механизма получения этого воздействия, то есть, наблюдаемое значение случайной величины Y_i соответствует потенциальному исходу $Y(D_i)$.

Первую часть этого предположения можно интерпретировать как отсутствие экстерналий (внешних эффектов). Вторая компонента предположения близка по смыслу к свойству контрфактуальных утверждений, которое в [9] называется состоятельностью (англ.: consistency), а в работе Галлеса и Перла [5] – композицией; кроме того, в эконометрической литературе распространено близкое по смыслу понятие экзогенности. Оптимизационное поведение исследуемых объектов нередко ведёт к нарушению данного предположения в ситуациях, когда они самостоятельно отбираются для того, чтобы подвергнуться воздействию. Такие же проблемы можно наблюдать в случаях, когда воздействие осуществляется на основе тех или иных характеристик объектов изучения.

Предложенный подход можно проиллюстрировать, используя результаты исследования влияния госпитализации на здоровье людей, приведённые в [10]. В этом примере использовались данные опроса населения США The National Health Interview Survey, в котором включены вопросы, “Был ли госпитализирован опрашиваемый в течение последних 12 месяцев”, и “Как вы оцениваете своё здоровье по шкале от 1 (отличное) до 5 (плохое)”.

Из данных в таблице следует, что опрошенные после госпитализации имеют в среднем существенно худшие показатели здоровья. Однако, с другой стороны, наверняка такие люди изначально обладали существенно худшим здоровьем, чем те, кто не был госпитализирован.

Обозначая за Y_{i0} уровень здоровья человека, не подвергнутого госпитализации, а Y_{i1} – уровень здоровья того же человека после госпитализации,

Таблица 1: Описательная статистика показателя самооценки уровня здоровья, источник [10]

Группа	Размер выборки	Среднее значение	Стандартная ошибка
Воздействие	7774	2.79	0.014
Контрольная	90049	2.07	0.003

можно получить следующее представление для среднего эффекта воздействия:

$$\underbrace{\mathbb{E}\left(Y_i \mid D_i = 1\right) - \mathbb{E}\left(Y_i \mid D_i = 0\right)}_{\text{наблюдаемая разность}} = \underbrace{\mathbb{E}\left(Y_{i1} \mid D_i = 1\right) - \mathbb{E}\left(Y_{i0} \mid D_i = 1\right)}_{\text{ATT}} +$$

$$+ \underbrace{\mathbb{E}\left(Y_{i0} \mid D_i = 1\right) - \mathbb{E}\left(Y_{i0} \mid D_i = 0\right)}_{\text{смещение выборки}}$$

Нас интересует первый член в правой части равенства – ATT (от англ. average treatment effect on the treated – средний эффект воздействия в части популяции, подвергшейся воздействию) – поскольку он является наиболее подходящим доступным приближением среднего эффекта воздействия по всей популяции ATE (от англ. average treatment effect). При этом, наблюдаемая разность также содержит компоненту, интерпретируемую как смещение выборки, – то есть среднее различие здоровья между теми, кто был, и кто не был госпитализирован. В случаях, когда воздействие назначается случайно (включая рандомизированные контролируемые испытания), выполняется $\mathbb{E}\left(Y_{i0} \mid D_i = 1\right) = \mathbb{E}\left(Y_{i0} \mid D_i = 0\right)$, поэтому смещение выборки равно нулю, и разность средних значений имеет причинно-следственную интерпретацию.

Фундаментальные отличия между экспериментальной и контрольной группами в рассмотренном примере означают, что условия устойчивости индивидуальной величины воздействия нарушены, и без дополнительной информации средний эффект воздействия не идентифицирован. Заметим также, что если бы данные из таблицы 1 использовались для оценки регрессии Y_i на D_i и константу (с использованием метода наименьших квадратов), то оценка коэффициента при D_i также была бы равна разности средних показателей здоровья в двух группах. Таким образом в этом примере для того чтобы регрессионная оценка соответствовала причинно-следственному эффекту от госпитализации, требуется решить проблему смещённости выборки.

Подробный разбор модели Рубина для системы с конечным числом состояний можно найти в монографии [11].

2. Причинно-следственные диаграммы и структурные статистические модели

Другой подход к идентификации структурных параметров использует модели на орграфах (диаграммы причинно-следственных связей), в которых каждая из вершин соответствует одной из переменных, представляющих интерес в конкретной задаче; при этом никакая переменная не ставится в соответствие с более чем одной вершиной, а рёбра графа могут быть отождествлены с теми или иными зависимостями между переменными. Такие модели являются частью более общего класса графических моделей, в которых направленные связи используются для задания марковских свойств совместного распределения. В нашем случае, связанная пара вершин показывает наличие статистической связи между соответствующими переменными, а направленный путь отождествляется с причинно-следственной связью от вершин-родителей к детям. Ациклич-ность орграфов необходима, поскольку в обратном случае в переменных нельзя разделить причину и следствие.

Применение графических моделей для описания структуры зависимостей в данных впервые встречается в работах Сьюэла и Филипса Райта, [12], и было популяризовано в 2000-х годах работами Перла [13], некоторые идеи из которых представлены далее. Модели, основанные на таких графах, называются DAG-моделями (от англ. directed acyclic graph).

Для заданного ориентированного ациклического графа можно говорить о совместных распределениях, которые являются совместимыми с ним. Такие распределения можно определить, как все распределения, для которых существует такая нумерация переменных, что если вершина j является родителем вершины i в графе, то переменная j входит в число родителей переменной i в марковском смысле.

Содержательно понятие совместимости ориентированного ациклического графа с заданным совместным распределением играет ключевую роль, поскольку оно оказывается необходимым и достаточным условием для того, чтобы распределение могло быть порождено процессом, который последовательно генерирует переменные из соответствующих условных распределений, принимая во внимание значения переменных-родителей, сгенерированных на предыдущих шагах.

Одной из наиболее важных характеристик совместного распределения, которая представляет особый интерес в прикладных статистических исследованиях, является набор отношений условной независимости. Оказывается, что такую информацию можно получить, используя направленный ациклический граф, с которым заданное распределение совместимо. Для

этого требуется ввести понятие блокировки пути: ненаправленный путь S на графе DAG-модели блокирован набором вершин Z, если выполнено хотя бы одно из следующих условий:

- в S найдётся цепочка $X_1 \to X_2 \to X_3$ или вилка $X_1 \leftarrow X_2 \to X_3$, причем $X_2 \in Z$;
- в S найдётся обратная вилка (коллайдер) $X_1 \to X_2 \leftarrow X_3$, такая что $X_2 \notin Z$, и также никакой из потомков X_2 не содержится в Z.

Такие обозначения связаны со статистической зависимостью между переменными в модели. Действительно, если переменные X_1 и X_3 имеют общую причину X_2, то они статистически зависимы – в полном соответствии с принципом Райхенбаха [3]. В обратной ситуации, две безусловно независимые переменные X_1 и X_3, имеющие общий эффект X_2, становятся статистически зависимыми условно на значение X_2.

Будем также говорить, что наборы переменных X и Y разделены Z, если все пути от каждого из элементов X к каждому из элементов Y блокированы. Можно показать, что две группы случайных величин X и Y независимы условно на третью группу случайных величин Z при любом распределении, согласованном с заданным ориентированным ациклическим графом, если Z разделяет X и Y в этом графе в приведенном выше смысле. В обратную сторону, если Z не разделяет X и Y в этом графе, то существует согласованное с ним распределение, в котором X и Y не являются независимыми условно на Z.

Одна из причин удобства DAG-моделей заключается в предположении о том, что каждому ориентированному ребру в графе соответствует некий устойчивый и автономный механизм. Иными словами, можно представить ситуацию, где лишь одна из связей в графе меняется, оставляя все прочие неизменными. Организация знания в подобную модульную структуру позволяет использовать причинно-следственную диаграмму для предсказания эффекта внешних вмешательств с минимальной дополнительной информацией, поэтому подобная модель (предполагая, что она корректна) зачастую оказывается намного более информативной, чем набор уравнений.

Условия соответствия ациклического орграфа причинно-следственной модели.

Ориентированный ациклический граф G задаёт причинно-следственную модель, если выполнены следующие условия:

1. Все переменные, включенные в G, являются наблюдаемыми;
2. Все переменные, включенные в G, функционально зависят от своих родителей и идиосинкратических шоков – случайных величин,

которые независимы от других переменных в модели и друг от друга;

3. Каждая переменная X в графе гипотетически может быть подвергнута интервенции $do(X = x)$, которая
   - заменяет вероятностное распределение у X на фиксированное значение x,
   - удаляет из графа направленные рёбра, заканчивающиеся в X,
   - не производит никаких других эффектов по отношению к другим переменным в G (кроме тех эффектов, которые связаны с влиянием X на своих потомков в оставшемся графе).

DAG-модели и критерий обходных путей являются удобными инструментами для построения статистических моделей, называемых структурными, которые пригодны для оценки эффектов воздействия. Такие модели состоят из одного или нескольких уравнений, описывающих направленные связи между переменными. Они отличаются лишь набором предположений о свойствах ошибок (или, что эквивалентно, коэффициентов), которые для линейного случая можно записать в следующем виде:

$$Y := X^T\beta + \varepsilon. \tag{4}$$

Здесь Y – зависимая переменная, $X = (X_1, \ldots, X_\ell)^T$ – переменные воздействия/регрессоры, $\beta = (\beta_1, \ldots, \beta_\ell)^T$ – регрессионные коэффициенты. Знак присвоения в (4) подчеркивает направленную зависимость между переменными, и идентифицирует β, как такой параметр, при котором установка для X произвольного значения $x \in \mathbb{R}^\ell$, приводит к тому, что среднее значение Y становится равным $x^T\beta$.

Ненаблюдаемая переменная ε, называемая ошибкой модели, объединяет эффект всех переменных, являющихся причинами Y, но не включенных в модель. Параметр β из (4) также можно идентифицировать, задавая свойства ε с использованием *do*-нотации для интервенций:

$$\mathbb{E}\big(\varepsilon \mid do(X = x)\big) = 0,$$

что соответствует нулевому условному математическому ожиданию ошибки в модели, подвергнутой интервенции $do(X = x)$ и идентифицирует β через набор моментных соотношений, подробнее рассмотренных далее.

Примечание: авторы [10] рекомендуют при оценивании эффекта воздействия с использованием неэкспериментальных данных всегда представлять гипотетический рандомизированный эксперимент, соответствующий

рассматриваемой задаче. Такой подход достаточно полезен и для интерпретации результатов, и для проверки адекватности исследуемой модели поставленной задаче – хотя соответствующий рандомизированный эксперимент не всегда оказывается очевидным. Так, в работе [20] исследуется, насколько футбольные тренеры более склонны ставить автора победного гола в стартовый состав следующей игры, по сравнению с авторами голов, оказавшимися не последними. Поскольку в такой постановке важно отделить влияние именно того факта, что гол оказался победным – то есть, был забит при равном счете, и после него не было других голов, – то соответствующим рандомизированным испытанием можно считать эксперимент, в котором в матчах, выигранных командой с преимуществом в один мяч, и в которых было забито более одного гола, авторство забитых мячей меняется случайным образом среди игроков команды.

Определение. (Ациклическая) структурная модель, соответствующая графу $G = (V, E)$, – это набор случайных величин $\{X_j\}_{j \in V}$, задаваемых уравнениями

$$X_j := f_j(Pa_j, \varepsilon_j), \quad j \in V,$$

где Pa_j – родители по отношению к вершине j направленного графа, f_j – неизвестные (измеримые) функции, а ошибки ε_j являются в простейшем случае независимыми в совокупности.

Определение. Контрфактуальная структурная модель, получаемая вследствие интервенции $do(X_j = x_j)$ – это новая структурная модель, основанная на модифицированном графе $G(x_j) = (V, E^*)$ и наборе (контрфактуальных) переменных $(X_k^*)_{k \in V}$, где

- все рёбра, входящие в вершину j удалены, то есть $\forall i \in V$ если $e_{ij} \in E$, то $e_{ij}^* = 0$;
- все остальные рёбра сохранены, $\forall i, k \in V : \ k \neq j, \ e_{ik}^* = e_{ik}$;
- контрфактуальные случайные величины определяются следующим образом:

$$X_k^* := f_k(Pa_k^*, \varepsilon_k), \quad \text{при } k \neq j,$$
$$X_j^* := x_j,$$

  где Pa_k^* – родители X_k^* в E^*.

Для иллюстрации представленных понятий рассмотрим пример, основанный на так называемой “треугольной структурной модели” из [2].

Пусть структурная модель задана с помощью уравнений (5)-(7):

$$X := \varepsilon_x, \tag{5}$$

$$D := \alpha X + \varepsilon_d, \tag{6}$$

$$Y := \beta D + \gamma X + \varepsilon_y. \tag{7}$$

Такая структурная модель соответствует графу на рис. 1 слева:

Рис. 1: Слева: DAG-модель зависимости между рассматриваемыми переменными. Справа: результат интервенции $do(D = d)$

Для удобства предположим, что все переменные в модели совместно нормальны и имеют нулевое математическое ожидание, ошибки $\varepsilon_x, \varepsilon_y, \varepsilon_d$ независимы, а числовые коэффициенты α, β, γ таковы, что дисперсия у X, Y и D равна единице. Тогда легко видеть, что $\mathbb{E}(Y \mid D = d) = (\alpha\gamma+\beta)d$, но если установить для D значение d, то уравнение (6) теряет смысл, и заменяя в (7) значение D на константу, получаем, что функция регрессии примет вид $\mathbb{E}\big(Y \mid do(D = d)\big) = \beta d$, а средний эффект воздействия D на Y равен β. Полученной контрфактуальной модели соответствует граф на рис. 1 справа.

Сопоставляя свойства структурной линейной модели с моделью потенциальных исходов Рубина, легко видеть, что структурные модели являются моделями не для наблюдаемых, а для потенциальных исходов $Y(d)$, соответствующих различным возможным значениям переменной воздействия D. Тогда идентифицируемость β можно связать с условиями устойчивости индивидуальной величины воздействия. В рассмотренной модели (7) они выполнены благодаря предположению о независимости ошибок, и при оценке регрессии Y на D и X, оценка β методом наименьших квадратов (МНК) будет состоятельной (это можно проверить, применяя теорему Фриша-Ву-Ловелла). В то же время, в простой регрессии Y на D МНК-оценка будет иметь предел по вероятности, равный $\alpha\gamma+\beta$, что связано с нарушением предположения об экзогенности D: ошибка в такой модели содержит X и из-за этого имеет ненулевую корреляцию с D, что приводит к систематической зависимости между значением воздействия D и потенциальными исходами $Y(D)$.

В рассмотренном примере имела место типичная для прикладных исследований ситуация, когда для получения состоятельной оценки искомого эффекта D на Y требуется включить в модель дополнительную (контрольную) переменную X, не представляющую самостоятельного интереса. В более типичном случае совместная нормальность не предполагается и функции условного математического ожидания могут быть произвольными. Пусть, как обычно, D – переменная воздействия, Y – зависимая переменная, а $Y(d)$ – её потенциальные значения при $D = d$. Модель потенциальных исходов имеет вид

$$Y(d) := \beta_0 d + \varepsilon, \quad \mathbb{E}\varepsilon = 0,$$

где β_0 – значение параметра, интересующее исследователя, которое соответствует среднему изменению Y при увеличении d на единицу. Линейность этого эффекта может быть обоснована бинарностью переменной воздействия, или же в ситуации, когда нам важны лишь небольшие отклонения от некоторого значения d, что позволяет использовать линейное приближение. Предполагаем также, что переменная воздействия не является (безусловно) экзогенной, иначе β_0 можно было бы состоятельно оценить в простой регрессии Y на D, но у исследователя есть данные, частично объясняющие Y, то есть $\varepsilon = g(X) + u$, где X – вектор наблюдаемых случайных величин, $g : \mathbb{R}^k \to \mathbb{R}$ – неизвестная (измеримая) функция, вид которой не представляет самостоятельного интереса, а u – ошибка модели, обладающая свойством $\mathbb{E}(u \mid X) = 0$.

Предположение об условной экзогенности переменной воздействия в структурной причинно-следственной модели:

1. корректная спецификация модели (англ.: model consistency) – наблюдаемые данные Y_i сгенерированы в соответствии со структурной моделью $Y(d)$,
$$Y_i = Y(D_i);$$

2. условная некоррелированность:
$$\operatorname{cov}\big(D, Y(d) \mid X\big) = 0,$$
(наблюдаемая переменная воздействия D не коррелирует с ошибками модели условно на наблюдаемый набор контролей). В литературе на английском языке используются близкие по смыслу понятия conditional exogeneity и conditional ignorability, подразумевающие условную независимость D и $Y(d)$ при фиксированных X.

При таких предположениях в модели

$$Y = Y(D) := \beta_0 D + g(X) + u,$$

ошибка u ортогональна (в L^2-смысле) обоим регрессорам X и D, а значит параметр причинно-следственного эффекта β_0 является идентифицированным.

Оценка причинно-следственной связи на неэкспериментальных данных включает этап экспертного решения о выполнении условия экзогенности переменной воздействия в модели. Если такое условие может быть выполнено, то в оцениваемую модель следует включить контрольные переменные, обеспечивающие экзогенность переменной, то есть переменные, статистически зависимые с D и Y одновременно, причем не относящиеся к механизму воздействия D на Y. Пример использования такого подхода можно наблюдать в работе [14], в которой авторы исследовали гипотезу о влиянии открытости экономики на поведение валютных курсов.

В то же время, представленные условия дают возможность проверки полученных результатов следующим образом: в ходе дальнейшего исследования структуры зависимостей между переменными в рассматриваемой задаче могут быть обнаружены новые потенциально релевантные контрольные переменные. При этом, в случае, когда изначально предложенная структура зависимостей является верной, включение таких переменных в модель не должно приводить к существенному изменению оценок причинно-следственного эффекта. Подобная проверка может лишь увеличить убедительность полученного результата, но не окончательно доказать его (поскольку перебрать все потенциально возможные наборы контрольных переменных на практике невозможно, кроме гипотетической ситуации, когда удаётся подобрать модель, полностью объясняющую зависимую переменную), однако даёт достаточно эффективный механизм выявления ошибочных результатов; например, подобный подход использовался в работе [15].

3. Сравнение различных методов оценки эффектов воздействия с использованием симулированных данных

Для сравнения различных подходов к оценке индивидуальных эффектов воздействия использовались симулированные данные с существенно нелинейными зависимостями. Причинно-следственная диаграмма процесса порождения данных представлена на рис. 2 и при оценке зависимости

Y_i от D_i предполагалась известной. Выбор данной структуры объясняется тем, что это наиболее простой случай, требующий включения всех переменных в статистическую модель, что позволяет сфокусироваться на проблемах оценивания нелинейных зависимостей. Во многих практических задачах структура зависимостей будет сводиться к аналогичной после отбора переменных на основе критерия обходных путей. Кроме того, наличие возможных зависимостей между переменными $X_1, \ldots, X_k$ качественно не повлияет ни на одну из далее представленных процедур.

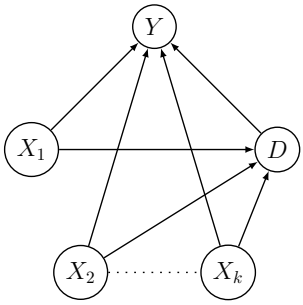

Рис. 2: Структура зависимостей между рассматриваемыми переменными

Генерация данных и оценка эффектов воздействия представлены в Алгоритмах 1-3.

Алгоритм 1. Генерация данных, соответствующих модели 2 с нелинейными зависимостями.

- Генерируется случайная выборка размером n нормальных векторов X_i размерностью $k \times 1$ с ковариационной матрицей $\sigma_x^2 I_k$. Полученная выборка подаётся на вход свёрточной нейронной сети с заданными параметрами глубины и числа слоёв, использующей активационные функции ReLU и наборы случайных центрированных весов w (использовались распределения с бесконечным вторым моментом, чтобы избежать сходимости в силу центральной предельной теоремы). На выходном слое к полученным переменным Z применялось преобразование $p = \frac{1}{1+e^{-Z^T w}}$
- Полученный вектор p используется для генерации переменных воздействия $D_i \sim \text{Bernoulli}(p_i)$.
- Переменные Y_i получались в результате применения к массиву Z, D еще одной аналогичной свёрточной нейросети, параметры весов которой сохранялись для вычисления контрфактуальных значений Y_{i0} и Y_{i1}.

Алгоритм 2. Вычисление эффектов воздействия с помощью дискретизации.

- Выбирается число узлов ℓ в сетке разбиения переменных, после чего значения X, Y округляются до ближайших выборочных квантилей уровня $\frac{1+2m}{2\ell}$, $m \in \{0, \ldots, \ell - 1\}$.
- Оценивается совместное распределение дискретизованных переменных (включая D), с использованием марковских свойств, соответствующих структуре зависимости в DAG-модели, рис. 3.
- Полученное совместное распределение используется для построения оценок функций регрессии $\mathbb{E}(Y_i \mid D_i = 1, X_i)$ и $\mathbb{E}(Y_i \mid D_i = 0, X_i)$, которые вычисляются в точках (дискретизованной) выборки
- $\hat{\mathbb{E}}(Y_i \mid D_i = 1, X_i) - \hat{\mathbb{E}}(Y_i \mid D_i = 0, X_i)$ используется в качестве оценок индивидуальных эффектов воздействия, а их усреднение по всем наблюдениям используется в качестве оценки среднего эффекта воздействия (ATE).

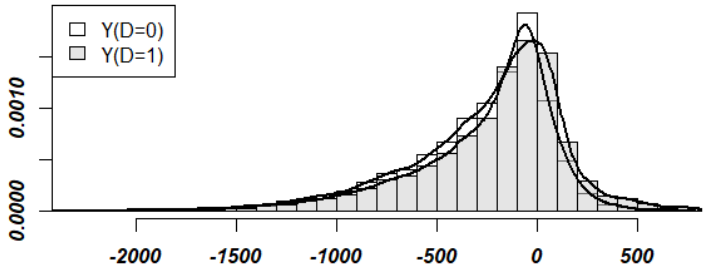

Рис. 3: Распределение Y при $D = 0$ и $D = 1$

Алгоритм 3. Вычисление эффектов воздействия с помощью метода двойного машинного обучения.

- Выборка используется для оценки структурной интерактивной регрессионной модели методом двойного машинного обучения [16]. Основное уравнение модели имеет вид:

$$Y_i := g(D_i, X_i) + \varepsilon_i, \quad \mathbb{E}(\varepsilon_i \mid X_i, D_i) = 0 \quad i \in \{1, \ldots, n\}, \tag{8}$$

где $g(\cdot)$ – неизвестная функция из $\mathbb{R}^k \times \mathbb{R}$ в $\mathbb{R}$, а ε_i – ошибки модели. Уравнение зависимости переменной воздействия D от контрольных переменных X:

$$D_i := m(X_i) + v_i, \quad \mathbb{E}(v_i \mid X_i) = 0. \tag{9}$$

Выбор метода оценки неизвестных функций осуществлялся с помощью алгоритма внутренней кросс-валидации, см. [14].

- Построение оценок индивидуальных эффектов воздействия осуществляется непараметрической оценкой $\mathbb{E}(Y_i \mid D_i = 1, X_i) - \mathbb{E}(Y_i \mid D_i = 0, X_i)$, после чего найденные значения ортогонально проецируются на базис из B-сплайнов для переменных X.
- Полученная функциональная форма позволяет вычислять оценки индивидуальных эффектов для любых наборов контрольных переменных, аналогично оценке, полученной в дискретном случае.

В таблице 2 приводятся результаты оценки среднего эффекта воздействия (ATE) с помощью линейных моделей. Из результатов видна нелинейная зависимость переменных от X: добавление контрольных переменных в модель регрессии Y на D лишь ухудшает оценку ATE (точное значение равно 70.0), но в присутствии полинома 5-й степени от X оценка становится достаточно точной, и близко совпадает с наилучшими оценками, полученными методом двойного машинного обучения.

Таблица 2: Оценка эффекта воздействия D на Y методом наименьших квадратов с различными наборами контрольных переменных

	(1)	(2)	(3)
	без контролей	линейные	полином 5й степ.
d	11.422	−20.555	53.224***
	(19.645)	(16.335)	(8.562)
x1		61.436***	...
		(3.115)	
x2		−20.779***	...
		(3.133)	
x3		200.987***	...
		(3.072)	
Число набл.	10,000	10,000	10,000
R^2	0.00003	0.320	0.822

Замечание: в скобках приведены стандартные отклонения оценок. Звёздочками помечены оценки, значимые на уровне 1%. В первом столбце оценивалась простая регрессия Y на D; во втором столбце оценивалась множественная регрессия Y на D и X_1, X_2, X_3; в третьем столбце в каче-

стве контрольных переменных использовались все компоненты полинома пятой степени от X_1, X_2, X_3.

В проведённых симуляционных экспериментах истинные значения индивидуальных эффектов сравнивались с предсказанными, и в большинстве случаев корреляция принимала значения от 0.1 до 0.2. Оценки усреднённых эффектов оказались неудовлетворительными при использовании дискретизации с числом точек p от 10 до 50. При этом достаточно хорошие оценки усреднённых эффектов можно было получить с использованием линейных моделей с полиномиальными контролями, а также при помощи двойного машинного обучения с использованием бустинга в качестве прогнозной модели. Из этого можно сделать вывод, что оценки индивидуальных эффектов в подобных процессах порождения данных могут быть существенно неточными. В частности, механизм настройки параметров моделей, предложенный в [19], заключающийся в использовании небольших выборок с заведомо экзогенными значениями переменной воздействия (получающиеся, например, путём проведения рандомизированного испытания в небольшом масштабе), в нашем случае, вероятнее всего, привёл бы к существенным ошибкам – поскольку маленькая выборка способствовала бы выбору модели с простой функциональной формой зависимости переменных от X, – тогда как в рассмотренном примере подобная оценка (столбец 2 в таблице 2) оказалась наихудшей из рассмотренных.

Заключение

В работе был проведен обзор теоретических подходов к моделированию и идентификации причинно-следственных связей и структурных параметров моделей. Рассмотрены типичные примеры, в которых оценка структурного параметра осложняется присутствием наблюдаемых и ненаблюдаемых переменных, статистически связанных одновременно с переменной воздействия и целевой переменной.

Основной результат работы заключается в исследовании оценок неоднородных эффектов воздействия в данных с нелинейными зависимостями. Для этого предложен алгоритм генерации данных при помощи свёрточных нейронных сетей, и исследуются два разных подхода к оценке – методом байесовских сетей и методом двойного машинного обучения. Показано, что оба таких подхода в рассмотренном случае дают неточные оценки индивидуальных эффектов, но при этом точность улучшается при переходе к агрегированным значениям. В частности, достаточно точными оказываются оценки средних эффектов воздействия, полученные с по-

мощью метода двойного машинного обучения, а также в множественной регрессии с полиномиальной зависимостью от контрольных переменных. Такой результат позволяет с осторожностью относиться к известному в литературе подходу к выбору модели для оценки эффектов воздействия, в котором предлагается использовать небольшую выборку, полученную с помощью рандомизированного эксперимента, для настройки модели, оцениваемой на (большой) неэкспериментальной выборке.

Список литературы

[1] Pearl J., Mackenzie D., *The Book of Why: The New Science of Cause and Effect*, Basic Books, 2018.

[2] Chernozhukov V., et. al., *Causal ML and AI*, 2025, `https://causalml-book.org/`.

[3] Reichenbach H., *The Direction of Time*, University of California Press, 1991.

[4] Lewis D., "Causation", *J. Philos.*, **70**:17 (1973), 556–567.

[5] Galles D., Pearl J., "An Axiomatic Characterization of Causal Counterfactuals", *Found. Sc.*, **3**:1 (1998), 151–182.

[6] Holland P., "Statistics and Causal Inference", *Journal of the American Statistical Association*, **81** (1986), 941–970.

[7] Rubin D., "Which Ifs Have Causal Answers?", *J. Am. Stat. Assoc.*, **81** (1986), 961–962.

[8] Neyman J., "On the application of probability theory to agricultural experiments. Essay on principles", *Statistical science*, **5** (1923), 465–480.

[9] Robins J. M., "Addendum to "a new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect"", *Computers & Mathematics with Applications.*, **14** (1987), 923–945.

[10] Angrist J., Pieschke J.-S., *Mostly Harmless Econometrics*, Princeton University Press, 2014.

[11] Imbens G., Rubin D., *Causal Inference for Statistics, Social, and Biomedical Sciences*, Cambridge University Press, 2015.

[12] Wright P., *The Tariff on Animal and Vegatable Oils*, The Macmillan company, New York, 1928.

[13] Pearl J., "A Probabilistic Calculus of Actions", *in: Uncertainty in Artificial Intelligence*, 1994, 454–462.

[14] Ченцов А. М., Торопов Н. И., "Применение подхода двойного машинного обучения для задачи анализа зависимости между отклонениями от непокрытого паритета процентных ставок и степенью открытости экономики", *Труды МФТИ*, **16**:3 (2024), 72–81.

[15] Campbell D., Chentsov, A., "Breaking Badly: The Currency Union Effect on Trade", *J. Int. Money Finance*, **136** (2023), 1–16.

[16] Chernozhukov V., et. al., "Double/debiased machine learning for treatment and structural parameters", *Econometric J.*, **21** (2018), 1–68.

[17] Frisch R., Waugh F. V., “Partial Time Regressions as Compared with Individual Trends”, *Econometrica*, **1**:4 (1933), 387–401.

[18] Imbens G. W., Angrist J. D., “Identification and Estimation of Local Average Treatment Effects”, *Econometrica*, **62**:2 (1994), 467–475.

[19] Facure, M., *Causal Inference in Python*, O’Reilly Online Learning, 2023.

[20] Smirnov, A., *Striking Success: How Goals Shape Coach Bias in Football*, Book of Abstracts European Conference on Sports Economics `2024-08-bookofabstracts`, 2024 .

Статья поступила 6 июля 2025 г.

Notes on identification and estimation of causal effects using observational data

A. M. Chentsov, N. I. Toropov

The beginning of the 21st century in data science is characterized by the emergence of new interdisciplinary fields as well as an expansion of the general theoretical framework and development of statistical methods for solving novel problems. This article describes known approaches to addressing the problem of statistical identification of unidirectional (causal) relationships between variables using non-experimental data, and highlights distinctive features of statistical models employed for this purpose. It also considers a case study on generating data with nonlinear dependencies among variables through convolutional neural networks, where two different estimation techniques are investigated — Bayesian networks and double machine learning. The results show that both these approaches yield inaccurate estimates of individual effects in the considered scenario, and recommendations are provided regarding aggregated effect evaluation.

Keywords: causal identification, treatment effects, DAG-models, double machine learning, CATE.

References

[1] Pearl J., Mackenzie D., *The Book of Why: The New Science of Cause and Effect*, Basic Books, 2018.

[2] Chernozhukov V., et al., *Causal ML and AI*, 2025, `https://causalml-book.org/`.

[3] Reichenbach H., *The Direction of Time*, University of California Press, 1991.

[4] Lewis D., “Causation”, *J. Philos.*, **70**:17 (1973), 556–567.

[5] Galles D., Pearl J., "An Axiomatic Characterization of Causal Counterfactuals", *Found. Sc.*, **3**:1 (1998), 151–182.

[6] Holland P., "Statistics and Causal Inference", *Journal of the American Statistical Association*, **81** (1986), 941–970.

[7] Rubin D., "Which Ifs Have Causal Answers?", *J. Am. Stat. Assoc.*, **81** (1986), 961–962.

[8] Neyman J., "On the application of probability theory to agricultural experiments. Essay on principles", *Statistical science*, **5** (1923), 465–480.

[9] Robins J. M., "Addendum to "a new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect"", *Computers & Mathematics with Applications.*, **14** (1987), 923–945.

[10] Angrist J., Pischke J.-S., *Mostly Harmless Econometrics*, Princeton University Press, 2014.

[11] Imbens G., Rubin D., *Causal Inference for Statistics, Social, and Biomedical Sciences*, Cambridge University Press, 2015.

[12] Wright P., *The Tariff on Animal and Vegetable Oils*, The Macmillan company, New York, 1928.

[13] Pearl J., "A Probabilistic Calculus of Actions", *in: Uncertainty in Artificial Intelligence*, 1994, 454–462.

[14] Chentsov A.M., Toropov N.I., "Application of double machine learning to estimation of the effect of country opennes of deviations from uncovered interest parit", *Trudy MIPT*, **16**:3 (2024), 72–81 (In Russian).

[15] Campbell D., Chentsov A., "Breaking Badly: The Currency Union Effect on Trade", *J. Int. Money Finance*, **136** (2023), 1–16.

[16] Chernozhukov V., et al., "Double/debiased machine learning for treatment and structural parameters", *Econometric J.*, **21** (2018), 1–68.

[17] Frisch R., Waugh F. V., "Partial Time Regressions as Compared with Individual Trends", *Econometrica*, **1**:4 (1933), 387–401.

[18] Imbens G. W., Angrist J. D., "Identification and Estimation of Local Average Treatment Effects", *Econometrica*, **62**:2 (1994), 467–475.

[19] Facure, M., *Causal Inference in Python*, O'Reilly Online Learning, 2023.

[20] Smirnov, A., *Striking Success: How Goals Shape Coach Bias in Football*, Book of Abstracts European Conference on Sports Economics `2024-08-bookofabstracts`, 2024.

Received on July 6, 2025