London School of Hygiene and Tropical Medicine
Faculty of Epidemiology and Population Health
Department of Medical Statistics

Assumption-Lean Inference for Causal and Statistical Questions in the Era of Machine Learning

Oliver Hines

Supervised by
Karla Diaz-Ordaz
&
Stijn Vansteelandt

December 2022

Thesis submitted in accordance with the requirements for the degree of Doctor of Philosophy
of the
University of London

Funded by the Medical Research Council

Declaration

I Oliver Hines, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Summary

Owing to the advent of sophisticated machine learning methods that excel at prediction modelling tasks, the field of statistics finds itself at a crossroads. Rather than pure prediction, the goal of statistics is usually more fundamental: to answer scientifically motivated questions of interest e.g. in the fields of epidemiology, sociology, psychology and economics. Traditionally, parametric statistical models have been used to frame and answer such questions, since model parameters often act as convenient and interpretable summaries of the aspects of the data which are of interest. This has lead to an uneasy tension between choosing complicated models that more accurately reflect the relationships between the variables of interest versus choosing simpler models that provide greater scientific interpretability. To overcome this tension, a so-called 'roadmap' was developed in which analysis is centred around target 'estimands' rather than model parameters. In this context, estimands are nonparametrically defined mappings of the true data generating distribution, which quantitatively answer scientific questions of interest. According to the roadmap, estimand inference is carried out using machine learning based estimators for requisite statistical functionals, or else more rarely, under limited semi-parametric assumptions.

These developments are quite revolutionary and have heralded new directions in how data is analysed. It is my view that for the roadmap to be successful it is necessary to enrich the space of available estimands, which at present is relatively unexplored. More often than not, estimands are proposed and interpreted within the framework of causal inference, with the average treatment effect of a binary exposure on an outcome being a canonical example. Extensions related to treatment effect heterogeneity and continuous exposures, however, are limited and this thesis makes contributions in both of these settings. Moreover, when considering potential estimands, it remains unclear the extent to which efficiency and model extrapolation concerns should be prioritised against scientific relevance of the estimand. This thesis studies questions of this type e.g. by considering optimal estimands that minimise nonparametric efficiency bounds, and by considering score based inference approaches that perform well when normality of the estimator breaks down. I argue that, in many cases, greater scientific insight can be gained by focussing on estimands that are less ambitious, in the sense that they pose questions about counterfactual worlds which are more similar to our own. These estimands can often be estimated with greater efficiency and with a lesser reliance on correct modelling of statistical functionals.

Acknowledgements

I must of course begin by acknowledging my two insightful and patient supervisors, Karla Diaz-Ordaz and Stijn Vansteelandt. Their continued support has been invaluable to my academic development and, whether chatting about statistics or nothing in particular, I will miss our regular meetings, which I have thoroughly enjoyed over the years.

I was also fortunate to have worked with some fantastic collaborators, and in particular I must thank Yalda Jamshidi and her research group at St. George's for helping me navigate genetic epidemiology; Mark van der Laan for supervising my work at the University of California Berkeley; the team at Novo Nordisk in Copenhagen for insightful conversations on methodological applications; and Maddalena Ardissino at Imperial College London for an energetic collaboration and for being a dear friend.

I must also acknowledge all those who made the LSHTM, and the Centre for Statistical Methodology in particular, a fantastic place to work and I would especially like to thank Ruth Keogh, Nick Jewel, Tom Godec, Schadrac Agbla, Kleio Kipourou, and Darren Scott in this regard. Additionally I must mention Lara Crawford and Lauren Dalton who worked behind the scenes at the MRC and the LSHTM to make my project possible.

There are many family and friends who have provided a great deal of support to me over the years. Though a thoroughly non-exhaustive list, I would especially like to thank my parents James & Sian, Gareth, Christian, Lily, Ruby, Aisha, Alex & Maddy, Tristan & Ella, Toby & Ophelia, Maddy & Ben, Elettra & Jack, Elisa, Paul & Fabi, Jow & Family, Fulham Jack, and California Toby.

Finally, I consider myself immensely privileged to have been able to dedicate several years to researching something that I am passionate about, and I thank the MRC London Inter-Doctoral Training Partnership for affording me this opportunity.

List of publications

[1] Hines, O., Diaz-Ordaz, K., Vansteelandt, S., & Jamshidi, Y.
Causal graphs for the analysis of genetic cohort data.
Physiological Genomics (2020).

[2] Hines, O., Vansteelandt, S., & Diaz-Ordaz, K.
Robust inference for mediated effects in partially linear models.
Psychometrika (2021).

[3] Ardissino, M., Vincent, M., Hines, O., Amin, R., Eichhorn, C., Tang, A. R., Collins, P., Moussa, O., Purkayastha, S.
Long-term cardiovascular outcomes after orlistat therapy in patients with obesity: a nation-wide, propensity-score matched cohort study.
European Heart Journal - Cardiovascular Pharmacotherapy (2021).

[4] Hines, O., Dukes, O., Diaz-Ordaz, K., & Vansteelandt, S.
Demystifying statistical learning based on efficient influence functions.
The American Statistician (2022).

[5] Hines, O., & Diaz-Ordaz, K.
Oliver Hines and Karla Diaz-Ordaz's contribution to the discussion of 'Assumption-lean inference for generalised linear model parameters' by Vansteelandt and Dukes.
Journal of the Royal Statistical Society: Series B (Statistical Methodology) (2022).

[6] Young, W.J., Lahrouchi, N., Isaacs, A. et al.
Genetic analyses of the electrocardiographic QT interval and its components identify additional loci and pathways.
Nature Communications (2022).

[7] Hines, O., Diaz-Ordaz, K., & Vansteelandt, S.
Parameterising the effect of a continuous exposure using average derivative effects.
arXiv pre-print (2020).

[8] Hines, O., Diaz-Ordaz, K., & Vansteelandt, S.
Variable importance measures for heterogeneous causal effects.
arXiv pre-print (2022).

The work presented in this thesis contains results from the publications [1-2], [4-5] and pre-print articles [7-8].

List of conference presentations

[1] *Causal machine learning workshop.*
European Causal Inference Meeting, virtual (May 2021).

[2] *Parameterising and inferring the effect of a continuous exposure using average derivative effects.*
European Causal Inference Meeting, virtual (May 2021).

[3] *Assumption-lean causal inference for direct and indirect effects.*
Joint Statistical Meeting, virtual (August 2021).

[4] *Variable importance measures for heterogeneous causal effects.*
International Biometric Conference, Riga (July 2022).

[5] *Demystifying statistical learning based on efficient influence functions.*
CMStatistics, London (December 2022).

The work presented in this thesis contains results from all of these presentations. The "Young Statistician Prize" at the International Biometric Conference 2022 was awarded for [4].

Contents

Chapter 1

Introduction

1.1 Background

Modern statistical theory is built on a framework of model based inference, where the targets of inference are the parameters indexing assumed semi-parametric statistical models[1]. It is difficult to understate the impact that this theory has had on society (epidemiology, psychology, economics etc.), especially since the latter half of the 20th century when advances in computational technology have meant that data is more routinely collected, and the cost of computing increasingly intricate analyses has been significantly reduced. As human beings, we find parametric models relatively straightforward to reason about, with model parameters encoding different aspects of the data generating mechanism. For instance, the model parameters indexing generalised linear models can inform investigators about the main effect of an exposure on an outcome, modification of this main effect by other variables, or mediation of the main effect through other variables. Parameter interpretations of this type, however, are inherently "causal" in nature, but despite this, little regard was given to the causal nature of the statistical model for much of the development of modern statistics. Instead, causal reasoning was understood to be simply outside of the remit of statisticians, with deliberately non-causal language used to describe statistical results.

Over recent decades, a theory of causal inference has developed[2] whereby a second "causal model" is specified alongside the statistical model. Formally, the causal model is a mathematical structure which encodes the assumed conditional independence relationships between the random variables of interest that is required to interpret model parameters "causally". Contrary to its name, the methods of "causal inference" are not able to infer whether one variable causes another or vice-versa, rather one might say that the goal of causal inference is to interpret the objects of ordinary statistical inference, in view of the assumed causal model. Indeed the algorithms and statistical machinery used in causal inference analyses are often identical to those used to make non-causal statements regarding association and correlation. Moreover, causal modelling relies on untestable causal assumptions, with domain-specific expert knowledge required to elicit and defend causal assumptions.

Philosophically speaking, the separation of the causal model and the statistical model is appealing since conditional independence assumptions that are made for the purposes of interpretation (the causal model) are distinct from those which encode a priori known parametric structure about the data-generating mechanism (the statistical model). Oftentimes, however, assumptions regarding the statistical model do not represent a priori known parametric structure, but instead are made either to facilitate inference or simplify model interpretability[3]. For example, time-to-event analyses in medical research routinely assume Cox proportional hazards models for convenience and because hazard ratio parameters are (arguably) easy to interpret. This results in two main issues which arise when the statistical model is misspecified:

[1] Likelihood based inference developed in the late 19th and early 20th century with pioneering work by Galton, K. Pearson, Fisher, E. Pearson, Neyman, Cramer, Rao etc..

[2] Causal developments are discussed in Chapter 2 with early work by Rubens, Pearl and others see e.g. Pearl (1986); Rubin (2005); Glymour (2006); Hernán and Robins (2020).

[3] Criticisms of this type can be found in Breiman (2001b); van der Laan (2015); Vansteelandt and Dukes (2022).

firstly, different estimators of the same model parameter may converge to different results; and secondly, it is not so clear how the resulting (estimator dependent) estimates should be interpreted, even in the limit as sample size grows to infinity. Worse still, these problems persist even when model/variable selection strategies are used to mitigate the risk of model misspecification, not least because uncertainty due to model selection is rarely acknowledged.

To address these issues, it has become increasingly common to centre analyses around nonparametrically defined targets of inference, called 'estimands', instead of focussing on statistical model parameters[4]. Like statistical model parameters, estimands can often be ascribed a causal interpretation under an assumed causal model, though the study of estimands remains interesting even in settings where this is not the case. The advantage of targeting a model-free estimand is that analysts can be more flexible in the modelling strategies used to estimate requisite statistical functionals, since the interpretation of the estimand does not rely on any particular form of the statistical model. In effect, this means that statistical models can be replaced with more flexible "algorithmic machine learning models" (e.g. lasso, neural networks, gradient boosting, random forests, ensemble learning etc.), which are routinely used for prediction tasks in the computer sciences.

Moreover, these developments are significant for the machine learning community, since complicated machine learning models, which may perform well in prediction tasks, are sometimes criticised as being 'black box' due to their lack of model interpretability[5]. The nonparametric theory surrounding estimands therefore provides a valuable tool for explaining the broad trends which are encoded in machine learning prediction models.

The current PhD project sits at the intersection of the four aforementioned topics: statistical modelling, causal modelling, estimand based inference, and algorithmic machine learning, and makes several contributions as outlined below.

1.2 Contributions

This PhD thesis consists of several self-contained chapters, intended to read like a series of thematically linked journal articles. This structure was chosen principally because this is the way that the field of statistics usually develops, by considering specific limited problems, with novel results communicated through standalone articles. Additionally, several of the chapters are in fact published (or pre-print) journal articles. Where this is the case, the associated publication is referenced according to the list of publications in the front matter of this thesis.

Chapter 2, which is published in [1], gives an introduction to causal modelling and outlines several common causal model structures which occur in the field of genetics and genetic epidemiology. The application area of genetic data is interesting, since it represents a field where parametric statistical modelling techniques are routinely used e.g. to parameterise the effect of a particular genetic variant on a physical trait, and to account for genetic cohorts with heterogeneous ancestry. We use causal directed acyclic graphs (DAGs) as a tool for representing causal assumptions and deriving implied independencies. Whilst the use of DAGs is common when discussing some genetic applications, such as Mendelian randomisation, we have not seen elsewhere similar discussions regarding genome wide association studies and ancestral confounding, with only limited DAG based discussions of selection biases in genetic cohorts. The main contribution of this chapter is therefore to consolidate these causal model structures and explain how they may be used to ascribe a causal interpretation to statistical model parameters.

Chapter 3, which is published in [2], focusses on the problem of inferring natural direct and natural indirect effects, under standard causal assumptions, and assuming certain semi-parametric partially linear models. Natural direct and indirect effects are nonparametric estimands that arise in mediation analyses in epidemiology, psychometrics, and economics. They quantify the amount by which a 'mediating variable' transmits the main effect of an exposure on an outcome, and under common partially linear statistical

[4]These ideas are codified in the 'Roadmap' by van der Laan and Rose (2011); Petersen and van der Laan (2014) and rely on results from nonparametric statistics, which we discuss in Chapter 4, see e.g. Pfanzagl and Wefelmeyer (1985); Pfanzagl (1990); Bickel et al. (1993).

[5]See e.g. Ribeiro et al. (2016); Lundberg and Lee (2017) for proposals related to interpreting black-box predictions.

model assumptions, respectively reduce to a single model coefficient and a product of two model coefficients. The latter product of coefficients makes inference in this context particularly challenging, and we use a so-called "G-estimation strategy" to address this inference problem in a new way. Our estimators demonstrate appealing robustness properties when parts of the model are misspecified and we make use of recent score-type testing results to test null effect hypotheses.

Whilst Chapter 3 evokes semi-parametric statistical models to infer nonparametric estimands, Chapter 4, which is published in [4], demonstrates how estimand inference can be carried out in a model-free way, so long as the estimand is 'pathwise differentiable'. Such estimands usually permit efficient estimators that are amenable to data-adaptive/ machine learning estimation of requisite statistical functionals, representing a fundamental and revolutionary departure from parametric statistical modelling in terms of how data is analysed and results are interpreted. One of the main challenges in deriving efficient estimators is first to derive the estimand's 'pathwise derivative', also called its 'efficient influence function/curve'. The goal of Chapter 4 is to demystify estimand inference, with a particular focus on influence curve derivations, often regarded as somewhat of a dark art. We advocate a 'point mass contamination' method for influence curve derivation and rederive several literature influence curves using this approach. In later Chapters, we use this same method to derive efficient influence curves for new estimands.

Chapter 5, which is submitted as a pre-print in [8], contains a proposal for a new estimand to quantify the importance of covariates in explaining heterogeneity in the effect of a binary treatment on an outcome. The proposed estimands are a novel contribution of the current thesis and relate analogous 'variable importance estimands' in nonparametric regression analysis to recent 'variance of treatment effect' estimands, which act as global measures of treatment effect heterogeneity. We assume a canonical causal model as found in the literature on (conditional) average treatment effects, making the proposed methods immediately applicable to e.g. both clinical trials data and observational 'real world' data.

One common feature of the proposed variable importance estimands, the regression variable importance estimands, and the variance of treatment effect estimand, is that they are all defined on a bounded support e.g. $[0, \infty)$ or $[0, 1]$. This makes subsequent inference challenging since the asymptotic normality of the estimator breaks down in finite samples when the true estimand value is close to the boundary of the support. This issue is addressed in Chapter 6, which contains a generic proposal for score based inference of nonparametric estimands, as opposed to the typical Wald type methods described in Chapter 4. Our proposal builds on ideas from 'targeted learning' (TMLE) and the score testing procedures considered in Chapter 3, and is shown to perform well in simulation studies in terms of confidence interval coverage.

Although the theory of model-free estimand based inference is most often applied to settings where the estimand is causally interpreted under a causal model, it is not necessary that the estimand is causally motivated. In Chapter 7 we present results for weighted derivative effect estimands, which have classically been studied in the econometrics literature in the context of single index models, though they remain equally applicable to epidemiological problems. These estimands consider how the conditional response surface of an outcome varies, on average, for small changes in the exposure. Traditional estimators are based on nonparametric kernel density estimators, however these introduce complicated biases as the number of the predictors grows. By considering nonparametric efficiency bounds, we derive an optimally weighted average derivative estimand and connect it to literature on so-called projection estimands in partially linear models. We propose a class of 'least squares estimands' containing the optimal one and derive efficient estimators under the model-free estimand inference framework, reviewed in Chapter 4. In Chapter 8, least-squares estimands, and other weighted derivative effect estimands, are ascribed a causal interpretation in terms of so-called stochastic interventions.

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT

T: +44 (0)20 7299 4646
F: +44 (0)20 7299 4656
www.lshtm.ac.uk

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	1804954	**Title**	Mr.
First Name(s)	Oliver		
Surname/Family Name	Hines		
Thesis Title	Assumption-Lean Inference for Causal and Statistical Questions in the Era of Machine Learning		
Primary Supervisor	Karla Diaz-Ordaz		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?	Physiological Genomics		
When was the work published?	2020		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	**Yes**	Was the work subject to academic peer review?	**Yes**

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	
Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	This study was conceived by all authors. I carried out the research into existing literature and writing of the manuscript under the supervision of the other authors.

SECTION E

Student Signature	Ohines
Date	14 December 2022

Supervisor Signature	Karla Diaz-Ordaz
Date	14 December 2022

Chapter 2

Causality in genetics

2.1 Introduction

Genetic cohort data is increasingly used to look for associations between candidate genes or genome regions and specific outcome measures, or else between modifiable risk factors and disease outcomes. Genome Wide Association Studies (GWAS), for example, are a popular and effective approach to analysing Single Nucleotide Polymorphism (SNP) data, which identifies reproducible regions of the genome associated with common traits. Observed GWAS associations, however, are not necessarily indicative of causal relationship, unless one is willing to make additional assumptions on the causal structure of the cohort data.

Mendelian Randomisation (MR) is another popular method, which uses genetic cohort data (or GWAS summary statistics) to establish causal effects between two phenotypes. MR seeks to exploit random genotype allocation, which occurs naturally due to Mendelian inheritance. The requisite MR assumptions are strong, and the causal structure underlying the data must be carefully considered so that biases are not unwittingly introduced. Since both GWAS and MR rely on genetic cohort data, it is more important than ever to understand, and communicate the causal structures found in these datasets, so that findings remain clinically relevant.

Universal frameworks to study causal structures have emerged in the past few decades, based on potential outcomes modelling (Rubin, 2005) or causal graphs (Pearl, 1986), contributing towards a modern causal understanding of several existing techniques, such as, randomised controlled trials, instrumental variable, and observational data techniques (propensity score methods and sample matching). Causal graphs may inform both the design and analysis of observational studies, and have successfully been applied to problems in epidemiology (Glymour, 2006; Glymour and Spiegelman, 2017), social science (Brady, 2013) and economics (Imbens, 2019) to represent causal assumptions, and derive causal quantities from observed data.

Eliciting and defending causal assumptions requires an expert understanding of the problem at hand. Here we review methods from genomics and genetic epidemiology, highlighting common causal structures which can bias observed associations. We advocate the use of causal graphs, firstly as a formal tool for representing and communicating the causal assumptions regarding data collection and study design, which underly analytical methods, and secondly, for deriving testable implications based on those assumptions. Causal graphs have several attractive properties in this regard. As a communication tool they are inherently diagrammatic and equation-free, aiding interpretability, whilst as a derivation tool one may apply powerful and rigorous mathematical rules, which link causal relations to statistical associations. These rules are summarised in Section 2.2.1.

We will initially introduce causal concepts which form the basis of our discussion. These are then applied to an example of pleiotropy in Section 2.1.2. Section 2.2 discusses causal methods for analysing selection biases, using, as an example, the analysis of case-control data for secondary trait association. Here we see the utility of causal graphs in deriving associations between variables which occur under selection. Section 2.3 then reviews GWAS assumptions, addressing issues related to population structure,

while Section 2.4 reviews MR causal assumptions, highlighting several ways in which they may be violated.

2.1.1 Introduction to statistical causal inference

There exists rich philosophical debate on what it means for one thing to *cause* another (Vandenbroucke et al., 2016), however, in the study of causal inference an interventionalist definition is used (Pearl, 1986; Glymour, 2006; Hernán and Robins, 2020). In this way, questions of causality are reduced to questions of the type: *what would happen if...?*

For example, for two variables A and B, we say that A **causes** B if the value that B takes would be different (or different in probability) if we had intervened by setting A to some other value. In this context we might also say that A **causally influences** B or that B is **causally dependent** on A. Two variables are said to be **statistically dependent** (or associated) if knowing the value of A in some way provides some information about the value of B (or vice-versa). Statistical dependence may arise due to a causal dependence between A and B, but also as a result of a causal dependence of both A and B on a third variable C, as we will see in the example in Section 2.1.2. Conversely, two variables are **statistically independent** if knowing the value of A does not provide any information about the value of B (and vice-versa).

This notion of causality may also be graphically represented using an arrow (Glymour, 2006; Hernán and Robins, 2020; Pearl, 1995, 2000), for example, $A \rightarrow B$ reads as "A causes B, but B could not possibly cause A". This arrow says nothing about the magnitude or direction of the effect that A has on B, just that if we were to intervene on A, then something would happen to B. Using these arrows one can form **paths**, which are any sequence of variables linked by arrows. For example, if A and B shared a common cause, C, then one may write the path, $A \leftarrow C \rightarrow B$. All possible paths containing three variables are given in Table 2.1. A path is *causal* if all the arrows point in the same direction. The path $A \rightarrow C \rightarrow B$, for example, is causal since A causes C which causes B, therefore if we were to intervene on A, the value of B could be different. Depending on the directions of the arrows, we also have additional terminology for the intermediate variable, also given in the table.

Path	Description	Terminology
$A \rightarrow C \rightarrow B$	A causes B (through C)	Mediator
$A \leftarrow C \leftarrow B$	B causes A (through C)	Mediator
$A \leftarrow C \rightarrow B$	A and B share a common cause C	Confounder
$A \rightarrow C \leftarrow B$	A and B both cause C	Collider

Table 2.1: All possible paths between three variables (A, B, C), with a brief description and additional terminology for the intermediate variable C

On its own, a single path is of limited use, motivating a network structure to represent several paths at once. The causal Directed Acyclic Graph (DAG) is such a structure, which for a set of variables, contains *all possible* paths between them. Causal graphs are said to be **acyclic** if there are no causal paths from one variable back to itself. It may seem obvious to say that any two variables, A and B, on a causal graph could either be linked by the arrow $A \rightarrow B$, the arrow $B \rightarrow A$, or no arrow at all. Each configuration makes different assertions about the impossible causal relationship between A and B. Respectively these are that B is not a direct cause of A, A is not a direct cause of B, or that A and B could not possibly be direct causes of each other. In this sense the arrows which are absent, and those which are present are equally important. Similarly, one must be careful to include common causes of A and B, even if they are unmeasured, since to not do so is to assert that it is impossible for such variables to exist.

At this stage it is also useful to introduce some terminology, which will become important later on. Firstly, a **collider** is any variable on a path which is causally dependent on the two variables adjacent to it, as in the final example in Table 2.1. Secondly, the **ancestors** of a variable are those which causally influence it (i.e. there is a causal path from each ancestor to the variable), and finally the **descendants** of a variable are those which are caused by it (i.e. there is a causal path from the variable to its descendants).

2.1.2 Example using pleiotropy

Our first example is inspired by a recent discussion of pleiotropy of the fat mass and obesity-related gene *(FTO)*(Ganeff et al., 2019). Consider a Single Nucleotide Polymorphism (SNP) in the *FTO* gene, such as rs1421085, which has been found to be associated with adiposity and brain function (Chuang et al., 2015). Suppose that a genetic cohort study has been conducted where, for each individual in the study population, an investigator measures body mass index (BMI), B, cerebral blood flow, C, and genotype rs1421085 in the *FTO* gene, denoted by F and coded as 0,1 or 2.

The original authors suggested that reduced cerebral blood flow in the medial prefrontal cortex may effect impulse control and hence BMI (Ganeff et al., 2019). As an illustration, we will attempt to refute the null hypothesis, that there is no causal relationship between cerebral blood flow and BMI by (1) positing the causal relationships that we believe hold amongst the variables involved; (2) representing these causal relationships using a causal graph; and (3) examining the graph, using formal operations, to derive testable assumptions.

Since a person's genome is assigned before their BMI or cerebral blood flow is determined, we argue that it is safe to assume that B and C could not possibly cause F. This assumption, however, says nothing about whether F causes B or C. Since it is possible that F causes B and C we must include the arrows $F \rightarrow B$ and $F \rightarrow C$ in our causal graph. For the purposes of illustration, we will additionally make the strong assumption that no other measured or unmeasured variables causally influence both B and C.

The causal graph in Fig.2.1 represents the causal assumptions posited between F, B and C under the null hypothesis that there is no causal relationship B and C. These assumptions are unnecessarily strong for the purpose of illustration, since additional variables might be included such as age or physical activity level, which are common causes of both B and C. Other violations of our assumption, which could arise due to population structure, are discussed in Section 2.3. We remark that while the causal graph in this example is perhaps oversimplified, such assumptions are not uncommon, and by using a causal graph representation we are required to be transparent about them.

Figure 2.1: Causal graph representing the causal assumptions between a patients *FTO* gene variant, F, body mass index, B, and cerebral blood flow, C.

In the graph in Fig.2.1, there is no causal path between B and C, but that does not mean that they are statistically independent. In fact one might expect a negative correlation between BMI and cerebral blood flow since those who inherit the *FTO* variant are likely to have a higher BMI and also a lower cerebral blood flow. This statistical dependency can be read off the graph in the form of the possible path: $B \leftarrow F \rightarrow C$. It is a general rule that two variables will be statistically independent if all paths between them that contain colliders. For this reason, we can refer to paths that do not contain a collider as *open paths* and those that do as *closed paths*.

Using our causal graph, we may derive testable assumptions in an attempt to falsify our null hypothesis. Imagine, for example, that we are told the value of B for a particular patient, and are asked to predict their value of C. The value of B may inform our prediction since B and C may be statistically dependent (due to confounding by F). If, however, we are subsequently told the patient's *FTO* variant then, under our causal assumptions, a new prediction based on F and B is no better than a prediction based on F alone, since B only informed our prediction in so much as it may have conferred some information about F.

This important observation is an example of how one may *block* open paths, such as $B \leftarrow F \rightarrow C$, by *conditioning* on an intermediate variable (F). Conditioning on a variable can be done either by stratifying by that variable or by including it as an independent variable in a regression model for B or C. These conditional independences are essential as they allow us to falsify our causal assumptions.

In practice, this means that if one were to stratify our imaginary study population by their *FTO* gene variant, then, under our causal assumptions, no association between B and C should be observed within

strata. An association between B and C within strata is, therefore, evidence that our assumptions are invalid. This could be because our null hypothesis does not hold, and B and C are causally related, or else because the relationship between them is confounded by some other variables, which we have not accounted for.

2.2 Selection bias

Due to the considerable cost of obtaining original genetic cohort data, it is common for case-control data to be repurposed for analysis of a secondary trait, such as human height (Gudbjartsson et al., 2008; Weedon et al., 2011), obesity (Loos et al., 2008), or plasma lipid concentration (Willer et al., 2008). Methods that fail to account for the case-control study design, are known to result in inflated error rates when testing for null association using GWAS (Lin and Zeng, 2009). Indeed it has been argued that epidemiological data analysis depends as much on study design and background information, as on the data itself (Robins, 2001).

Gene-phenotype associations, induced as a consequence of study design, are problematic in GWAS analyses because they are indistinguishable from underlying causal associations in GWAS results. Using causal graphs we may gain some insight into how the non-random selection of individuals to the study cohort propagates to non-randomness in our variables of interest. We will consider an illustrative example, inspired by a real study on the effect of Sex Hormone Binding Globulin (SHBG) on Type 2 diabetes in women (Ding et al., 2009). Consider that the study cohort was recruited on a case-control basis and consists of women with a recent Type 2 diabetes diagnosis ($D = 1$) and controls ($D = 0$), with genotyping carried out for all women. We shall examine the issues which arise when this cohort is used to conduct a GWAS analysis, with SHBG as the outcome of interest.

SHBG is a glycoprotein, produced in the liver, and the level of SHBG in an individual's blood plasma will be denoted by H. The original authors found that high levels of SHBG were associated with a lower risk of Type 2 diabetes and for this example we shall assume that diabetes status does not causally influence SHBG level. Imagine also a specific SNP, G, which does not causally influence SHBG, but does causally influence diabetes diagnosis by some other mechanism. As with the example in Section 2.1.2 we shall make the "no unobserved confounding" assumption, i.e. that there are no common causes of H, G, or D that we have not accounted for.

Due to the case-control design, diabetes status D causally influences selection to cohort, S. By definition $S = 1$ for all women in the cohort and $S = 0$ for all other women in the population as a whole. Our causal assumptions are represented by the causal graph in Fig.2.2a.

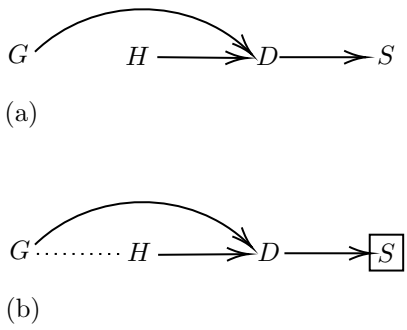


Figure 2.2: (a) Causal graph representing the causal assumptions between a specific gene of interest, G, Type 2 diabetes status, D, SHBG level, H, and selection to the cohort, S. (b) Causal graph when considering only individuals in the cohort ($S = 1$). The selection variable has been conditioned on, indicated by the box around it. The induced association between G and H is represented by the dashed line.

Under these assumptions, G and H are statistically independent as there are no open paths between them. One would expect, therefore, to observe no association between G and H for women sampled

from the population. Our cohort, however, is not randomly sampled from the population, but instead we observe only those for whom $S = 1$. This is equivalent to an unavoidable stratification by S, which allows us to observe only the $S = 1$ stratum. In this stratum, a "spurious" association between G and H may be induced, which we demonstrate by first examining the $D = 1$ and $D = 0$ strata separately.

In the cases group ($D = 1$) an association between G and H would be observed, since, if an individual's genotype suggests they are unlikely to have diabetes, then their diabetes status is more likely due to a low level of SHBG, and vice-versa. For women in the control group ($D = 0$) an association between G and H would be observed, since women in this group are less likely to carry the genotype associated with diabetes and are also more likely to have high SHBG.

We see, therefore, that G and H are associated in both the $D = 0$ and $D = 1$ strata and that this association must be induced by the stratification process, since G and H are not associated in the population. Worse than this, however, is that stratifying by S also induces associations between G and H because the proportions of each D strata in our cohort are not representative of the population as a whole. For selection problems such as these we have no choice but to consider only the strata $S = 1$.

In this simple example we were able to reason that selection bias may influence our results, however, in other examples it may not be so clear. Causal graphs may go some way to elucidate selection biases. It is a general rule that conditioning on a collider, or the descendants of a collider, induces statistical dependencies between the ancestors of the collider. In our case-control example D was a collider on the path: $G \rightarrow D \leftarrow H$ and we were forced to condition on S, which is a descendant of D. This conditioning resulted in a statistical dependency between G and H (the ancestors of D). This induced dependency is represented by the dashed line on the causal graph in Fig.2.2b.

In Section 2.1.2 we saw how open paths on causal graphs could be blocked by conditioning on intermediate variables. In this example, however, conditioning has the opposite effect. By unintentionally conditioning on colliders, we are effectively unblocking a path that was otherwise closed, thereby inducing associations. Several solutions have been proposed, which allow case-control data to be used for secondary trait analysis in association studies. Example analysis strategies include analysing the cases and controls separately, re-weighting the data using additional models, or including case-control status as a covariate (Tchetgen Tchetgen and Shpitser, 2014; Song et al., 2016).

Biases introduced by conditioning on colliders are generally referred to as *collider stratification biases* (Bareinboim et al., 2014). The inclusion of selection variables in causal graphs, like the variable S in the case-control example, can also be useful for expressing selection and retention assumptions which suffer from similar collider stratification biases (Munafò et al., 2018). The UK Biobank is an example of a cross-sectional cohort study ($n \approx 500,000$) self-selected from a population of 9 million individuals invited to participate. The resultant cohort contains a lower proportion of current smokers (11% in the UK Biobank, vs approximately 19% in the general population), with a similar discrepancy observed in educational qualification attainment. For a highly self-selected cohort, such as the UK Biobank, causal graphs may be useful in exposing subtle biases induced by this self-selection.

2.2.1 D-separation

The rules discussed in Sections 2.1.1 and 2.2 are collectively known as the rules of d-separation (statistical dependence separation). These rules describe statistical dependencies implied by causal graphs before and after conditioning on variables. Table 2.2 gives a summary of these rules for all possible paths of three variables. To consider longer, more complex paths one must 'chain together' these triplets, and to consider the statistical dependence between variables on the whole causal graph, one must consider all possible paths.

For complex, multivariate causal graphs this could result in a laborious manual analysis. Fortunately, however, the tool `www.dagitty.net` may be used to examine statistical dependence on causal graphs using an online web tool or R package (Holland, 1986).

Path	Before conditioning on C	After conditioning on C
$A \to C \to B$	open	closed
$A \leftarrow C \leftarrow B$	open	closed
$A \leftarrow C \to B$	open	closed
$A \to C \leftarrow B$	closed	open

Table 2.2: Summary of the rules of d-separation for all possible paths containing three variables. The two additional columns describe the statistical dependence of A and B before and after conditioning on the intermediate variable C.

2.3 Causal graphs for genome wide association studies

GWAS studies are a popular and effective approach to analysing SNP data, which identifies reproducible regions of the genome associated with common traits. As of February 2020, the GWAS Catalogue contains 4439 publications and 175870 associations(Buniello et al., 2019). Despite their popularity, it is important to remember that the associations discovered by GWAS are not necessarily causal unless one is willing to make additional assumptions. In this section, we use causal graphs to make these assumptions explicit. Genetic relatedness between individuals in the study population poses an additional, well-known challenge that results in individuals with shared ancestry inheriting similar common variants. Heterogeneous study populations, therefore, complicate the task of separating the contributions of individual genetic variants toward phenotypes of interest. We refer to the problem of heterogeneous ancestry as confounding by ancestry, since this more closely aligns with the language of causal inference. It is also referred to as population structure or population stratification, when at the population level, and kinship, at the familial level.

As an illustrative example, we will use Carotid Intima-Media Thickness (CIMT) as a phenotype of interest Y. In its most basic form, one assumes that the study population is in Hardy-Weinberg Equilibrium (HWE), that is, for each individual, the value of their value of a particular SNP of interest, G, is drawn from a binomial distribution with some fixed minor allele frequency for the population.

Common practice is to model a continuous phenotype, Y, using a model which is linear in G, and other relevant variables, such as age and sex, denoted by the 'Environmental' vector, E. When Y is a binary outcome, generalised linear models such as the logistic model, are often used. The linear model for a continuous phenotype, Y, may be written as

$$Y = \alpha G + \sum_{j=1}^{p} \beta_j E_j + \epsilon \tag{2.1}$$

where ϵ is a noise term, with constant mean given G and E, and β is a vector of parameters associated with the p environmental variables contained in the vector E. The unknown model parameters, α and β, may be estimated by Ordinary Least Squares (OLS). Ideally we would like to interpret the α parameter as *a parameter which quantifies the influence that the gene of interest has on the phenotype*, however, to do so is to make a causal assertion, requiring an examination of causal assumptions. We note that for a discussion of causal assumptions, the exact form of the regression model is not important. Instead, from a causal perspective, we are concerned with the variables which are and are not included in the regression model.

One possible causal graph for the basic GWAS analysis, which gives the α parameter the desired causal interpretation is given in Fig.2.3a. This graph is not unique since it is not strictly required that G and E are independent. Using the running example, the key features of this graph required to interpret α causally are

1. CIMT does not influence the gene of interest, but the reverse may be true.
2. CIMT does not influence age or sex, but the reverse may be true.
3. There are no variables (observed or otherwise), which are common causes of CIMT and the gene of interest, or of CIMT and age or sex.

The first of these assumptions is justified through the biological understanding that G is assigned before phenotypes are determined, hence reverse causation is not possible. Likewise, the second assumption is reasonable from a biological perspective. Assumption 3, however, is where the basic model breaks down. Under modern theories of Mendelian inheritance, the gene of interest depends on an individual's parental genotypes, or more generally on their ancestry. Along with the gene of interest, each individual inherits many other genetic variants, G^*, each of which could also have a causal influence over Y. The ancestry of an individual is therefore a confounder as it may be a common cause of both G and Y.

This effect is, however, negated if one assumes that Y is monogenic, so is causally affected by only one single SNP. Conversely the effect is amplified for polygenic traits, such as CIMT, which are thought to be affected by multiple genetic variants.

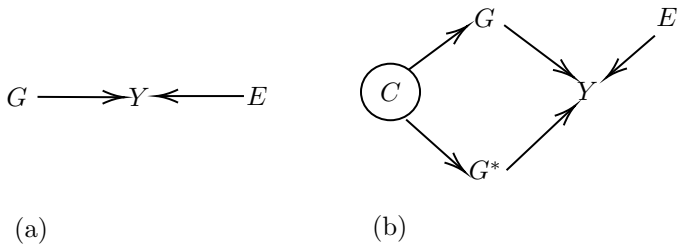


Figure 2.3: Causal graphs for GWAS analysis. Graph (a) shows the basic causal GWAS model, where the phenotype of interest, Y, is dependent on the gene of interest, G, and some other environmental factors, E. Graph (b) accounts for confounding by the ancestry of the individual, C, which affects the gene of interest, and the remaining genes, G^*. This modified graph assumes that a polygenic trait, Y, depends on both the gene of interest, and the remaining genes. By convention, unobserved (or latent) variables, such as the ancestry variable, C, are circled.

To adequately adjust for confounding by ancestry, the basic GWAS graph Fig.2.3a must be updated to reflect Mendelian inheritance assumptions. Fig.2.3b shows a causal graph, modified to include an unmeasured ancestry variable, C, which affects the phenotype of interest through both the gene of interest, G, and other inherited variants, G^*. In this updated causal graph, we see that there are two open paths by which the gene of interest is associated with CIMT, specifically the $G \to Y$ causal path and the $G \leftarrow C \to G^* \to Y$ non-causal path. If one were able to block the non-causal path, then, the remaining association between G and Y must be due to the causal path.

One strategy for blocking the path is to condition on ancestry by stratification. Since C is unmeasured, one must assume that the population consists of one strata, which is homogeneous in ancestry with a random mating scheme and no natural selection. Under these assumptions, the HWE model is recovered, whereby G is drawn from the same distribution for all individuals, hence G and Y are not confounded by ancestry.

The causal graph in Fig.2.3b made several additional assumptions regarding the ancestry variable, C. The first is that there is no direct path $C \to Y$. Modern epigenetic theory, however, does permit such paths through 'imprinting' mechanisms, whereby an individual inherits DNA of the same sequence, whose function is altered by the presence of additional methyl groups.

Furthermore, Fig.2.3b assumes that C and E are independent. This may not be true, however, for a global study, where individuals from different ethnic groups, may have been brought up in different geographical locations, and hence, different meteorological and socio-economic conditions. It is reasonable, therefore, to posit a $C \to Y$ path through some unobserved environmental variables. We emphasise again that the arrows absent from a causal graph are important as they represent causal relationships which are assumed not to exist, whilst the arrows represent causal relationships which may exist.

2.3.1 Using principal components to adjust for ancestral confounding

Examining the causal graph in Fig.2.3b, we discussed how the non-causal path: $G \leftarrow C \to G^* \to Y$ may be blocked by conditioning on C when one assumes the study population is homogeneous. For

heterogeneous populations, however, stratification by C is not possible because it is unmeasured. Instead, the non-causal path can be blocked by conditioning on the remaining observed SNPs, G^*. This involves using G^* in a regression model for Y, or using G^* for stratification.

Intuitively, conditioning on G and G^* removes any dependency between C and Y since, if the full genotype of an individual is used to predict their phenotype, then knowledge of their ancestral genotypes provides no new information to improve our prediction. Using the full genotype in a regression model for Y requires careful consideration, since the number of covariates (SNPs), p, may exceed the number of individuals in the study, $n < p$. Such 'high-dimensional' problems require alternative models and estimation techniques.

Due to the high-dimensionality, modifying the linear model in Eq.2.1 to include the remaining genes as covariates would result in a model which is impossible to fit by OLS. One very common solution is to drastically reduce the dimensionality of the genetic information, using Principal Components (PCs).

PCs are used in several ways within genomic analysis: (i) PCs can be used to cluster individuals, either by excluding anomalous individuals from the dataset (Anderson et al., 2010), or else clustering the data for use in a Structured Association analysis, (ii) some PC values may be included as fixed effects in a GWAS analysis, thereby accounting for some of the phenotype variation, which can be explained by the remaining SNPs, and (iii) PCs may be included as random effects in the GWAS analysis, an approach which is equivalent to using a Linear Mixed Model (LMM) (Hoffman, 2013).

Method (i) may be causally interpreted as stratifying the population into one or more sub-populations, for which we believe that HWE holds. Analysis of each sub-population may be conducted using a basic GWAS analysis. Limitations of this method are that confounding by ancestry is not accounted for within strata and it is not clear how to tune the stratification process.

The linear model for methods (ii) and (iii) may be written as

$$Y = \alpha G + \sum_{j=1}^{p} \beta_j E_j + \sum_{j=1}^{q} \gamma_j P_j + \epsilon \tag{2.2}$$

where P is the vector of q principal components, summarising the genetic data of a particular individual, each component of which has a coefficient given by the γ parameter vector, and where ϵ has constant mean given G, E and P. In the fixed effect model (method ii), the q-dimensional parameter vector, γ is treated as a fixed covariate, which may be estimated using conventional methods such as by OLS.

Alternatively, one may treat the parameters γ_j as random effects (method iii), by assuming a normally distributed prior for γ, resulting in a LMM. The use of LMMs in genomic data is not restricted to GWAS analyses. They are frequently applied to phenotype prediction, heritability estimation, and rare-variant analysis (Lippert et al., 2013). One key feature of LMMs is that the random effect (given by $\sum_{j=1}^{q} \gamma_j P_j$ above) may be written in terms of a 'genetic similarity matrix', which is used to model the covariance between any pair of individuals in the cohort. A more detailed discussion of LMMs and methods for measuring genetic similarity can be found in Appendix A.1.

2.4 Causal graphs for Mendelian randomisation

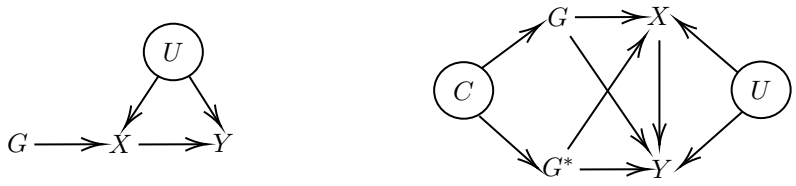
Mendelian Randomisation (MR) studies also make use of genetic SNP data, or GWAS summary statistics, with the aim of inferring the effect of a genetically modified exposure (e.g. alcohol consumption) on another phenotype (e.g. cardiovascular disease). GWAS results from multiple cohorts may be used to conduct Two- Sample MR analysis. MR base which is a database of GWAS statistics for conducting Two-Sample MR, contained associations from 1673 GWAS, as of May 2018 (Hemani et al., 2018). Another systematic review estimates a 10-fold increase in published MR studies between 2004 and 2015, with the majority (51%) in the fields of cardiovascular disease and diabetes (Swerdlow et al., 2016). MR is therefore increasing in popularity, most likely due to the increasing availability of GWAS summary statistics and large cohorts with genetic and phenotypic data.

This section provides an overview of the technique, from the statistical causal inference framework. We refer the interested reader to Didelez and Sheehan (2007); Burgess and Thompson (2015); Sheehan and Didelez (2018).

2.4.1 Instrumental variable methods

MR exploits the idea that a particular genotype affects the phenotype of interest only indirectly, through the exposure of interest, and that this genotype is assigned randomly (given the parents' genes) at meiosis, independently of the possible confounding factors. This is essentially using the genotype as a so-called *instrumental variable* (IV) for the effect of the exposure on the outcome (Didelez et al., 2010). This is appealing, as it allows to estimate causal effects event in the presence of exposure-outcome unobserved confounding. Nevertheless, MR makes a number of causal assumptions, known as IV assumptions, which are not always carefully stated and evaluated in applications and are separate from any parametric modelling assumptions, which may also be required.

For illustration, we consider a specific example where the interest is to investigate the causal effect of the level of C-reactive Protein (CRP) on CIMT by exploiting random assignment of a genetic variant, G, associated with CRP (Kivimáki et al., 2008). Here CRP is referred to as the exposure, X, CIMT as the outcome, Y, and G as the instrumental variable (or instrumental gene).



(a) (b)

Figure 2.4: Causal graphs for MR analysis. Graph (a) shows the traditional IV causal graph, where the gene, G, acts as an IV for the $X \to Y$ relationship of interest, itself confounded by the unmeasured variable, U. Graph (b) shows modifications to graph (a) which relax assumptions by allowing for confounding by ancestry, and some pleiotropic effects.

Note that the IV causal graph permits unmeasured variables that may influence both the exposure CRP and the outcome CIMT, here denoted by U. The IV assumptions encoded by the causal graph in Fig.2.4a can be written formally as follows

1. CIMT does not influence CRP, but the reverse may be true.
2. Relevance: The instrumental gene is associated with the level of CRP.
3. Exclusion restriction: The instrumental gene may affect CIMT only through its effect on CRP.
4. Unconfoundedness: There is no variable, observed or otherwise, which is a common cause of the instrumental gene and CIMT.

For assumption 1, domain specific knowledge is generally required to defend the $X \to Y$ causal relationship over the alternative, $Y \to X$. For this example, it is usually assumed that proteins causally influence disease outcomes, rather than the other way round. Collectively, assumptions 2 to 4 are known as the IV assumptions as they describe the relationship between the IV and the variables U, X and Y. In a randomised control trial (RCT), where the IV is the randomly assigned treatment group, these assumptions are more simple to justify, since the randomisation process is known, and we can engineer the randomised treatment so that it is (a) associated with the exposure, and (b) does not influence the outcome except through the exposure, although in some settings justification of the exclusion restriction remains challenging.

In the MR setting, we justify the relevance condition (assumption 2) by choosing instrumental genes following a GWAS analysis. In practice, several candidate instrumental genes are often used to support or discredit the evidence of a single one. The exclusion restriction (assumption 3) is, however,

more problematic as genetic variants may have independent pleiotropic effects on multiple phenotypes. Pleiotropic effects violate the exclusion restriction by introducing alternative paths of the type $G \to Y$.

Recent developments in MR do allow for some limited pleiotropy, such as MR-Egger (Bowden et al., 2015), which permits a direct path from $G \to Y$ in Two-Sample studies (under specific assumptions), and the MRGxE method (Spiller et al., 2018), which allows for pleiotropic 'Gene-by-environment' interactions provided they reside on the $G \to X$ path. Selection of instrumental genes in MR is, however, an open topic of debate, both in terms of statistical and biological considerations (Swerdlow et al., 2016). Recent statistical work considers variable selection methods, such as the Lasso, to select IVs (Windmeijer et al., 2018). Whilst the exclusion restriction cannot be proven, it may sometimes be possible to show that they are inconsistent with prior evidence. Methods for doing so include leveraging prior causal assumptions, identifying modifying subgroups, or by use of instrument inequality tests (Glymour et al., 2012).

Unconfoundedness (assumption 4) prohibits edges of the type $U \to G$, which is reasonably well justified on the basis of Mendelian inheritance. As in Section 2.3, however confounding by ancestry violates this assumption, since unobserved ancestry variables, C, may causally influence the outcome through their effect on other genetic variants as well as causally influencing the instrumental gene itself. Ancestrally heterogeneous populations are therefore known to violate the unconfoundedness in MR, and practitioners are recommended where possible to use homogeneous cohorts, thought to be in HWE.

A modified causal graph, which relaxes the IV assumptions to allow for confounding by ancestry, and limited pleiotropic effects, can be seen in Fig.2.4b. This graph represents a more general set of causal assumptions, to emphasise the assumptions of the IV graph. The standard IV graph may be recovered by removing arrows from the modified causal graph, or in other words, by assuming certain null causal relationships.

If only the $G \to Y$ arrow is removed from the causal graph in Fig.2.4b (i.e. G has no pleiotropic effect on Y) then G may be used as a *conditional instrumental variable*, assuming one collects adequate data on the other genetic variants G^*. In a *conditional instrumental variable* analysis, the gene G acts as an instrumental variable after conditioning on G^* in the models for X and for Y. This conditioning has the effect of blocking the open paths: $G \leftarrow C \to G^* \to X$ and $G \leftarrow C \to G^* \to Y$. Once blocked, unconfoundedness is no longer violated so G again acts as an instrument, allowing for valid MR analysis with ancestrally heterogeneous cohorts. Conditioning on G^* may be achieved using the methods in Section 2.3.1.

Violation of any of the IV assumptions would result in invalid causal estimates. We refer the interested reader to VanderWeele et al. (2014) for a comprehensive discussion of the challenges faced by MR studies when justifying the IV assumptions and on how to conduct sensitivity analyses.

2.4.2 Survivor bias in Mendelian randomisation

One setting where causal graphs are especially useful for evaluating MR assumptions is in the use of genetic instruments to asses survival biases. Here we consider the example given in Vansteelandt et al. (2018), namely where an MR analysis of the effect of vitamin D levels on mortality is performed using a cohort of ancestrally homogenous, genotyped individuals between the ages of 40 and 71 years old. Using causal graphs, we show how survivor bias may be introduced because recruitment to the cohort depends on an individual having survived long enough to be eligible for recruitment.

Selection to the cohort depends on T, the lifetime of an individual, being larger than some index time, T_0. By definition, an index time is actually assigned only to individuals in the cohort (who are indexed at some point between the ages of 40 and 71), however, we could imagine that individuals outside the cohort could also be given an index time, for example by sampling from the birth register. As before, we will denote selection to the cohort by the variable S, with $S = 1$ for all individuals in the cohort.

Let D be the level of vitamin D at index and assume that it captures the effect on lifetime of an individual's entire exposure to vitamin D since birth. This assumption is implicit in all MR studies, since to not assume it would generally violate the exclusion assumption, in the sense that we could imagine an additional variable (e.g. adolescent vitamin D level) which causally influences the vitamin D level recorded at index, as well as the lifetime of the individual directly.

Finally we shall assume that an appropriate genetic instrument (e.g. filaggrin genotype) has been recorded, which we shall denote, G, and assume is randomised by Mendelian inheritance, since the cohort is homogenous. As with the standard MR causal graph, we shall permit unmeasured confounding variables which might causally influence both vitamin D level and lifetime. Our causal assumptions for this example are represented by the causal graph in Fig.2.5a. In this example, S, is a variable which we have no choice but to condition on, hence we must be very careful to consider collider stratification biases, as discussed in Section2.2.

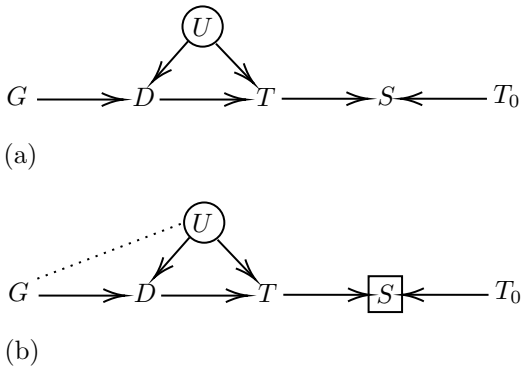


Figure 2.5: Causal graphs for MR analysis of a survival outcome. Graph (a) shows the instrumental gene, G, acts as an IV for the $D \to T$ relationship of interest where D is vitamin D level and T is lifetime. Graph (b), however, shows that conditioning on selection to the cohort, S, which depends on an individual surviving to index time T_0, introduces associations between G and U which violate the IV exclusion assumption.

We see that S is a descendent of D, due to the $D \to T \to S$ path, and that D is also a collider on the path $G \to D \leftarrow U$. Hence, by selecting only individuals who have survived, the ancestors of D (namely G and U) become associated. This violates the exclusion assumption, since association between G and T may arise from either the causal path $G \to D \to T$ or from the path $U \to T$, where U is associated with G.

The association induced by conditioning on selection is illustrated by the dashed line in Fig.2.5b. Recent work proposes various strategies for MR estimation under survivor bias, using a semi-parametric additive hazard model, similar to the canonical Cox proportional hazards model (Vansteelandt et al., 2018). This relates to similar work on MR for censored survival outcomes (Tchetgen Tchetgen et al., 2015).

Interestingly, however, this problem of survivor bias disappears when testing the null hypothesis that D has no causal influence on T. Under this null hypothesis, there is, by definition, no $D \to T$ arrow, hence G is not an ancestor of T and no association between G and U is induced.

2.5 Conclusion

We have demonstrated, through examples of the most common analytical techniques employed in genetic studies, that a causal inference framework, and in particular the use of causal graphs, allows the analyst to (i) to represent their knowledge of the causal relationships involved in the question at hand, and (ii) use the rules of d-separation, to query the assumptions under which popular genetic analysis methods lead to causal interpretations.

Causal graphs may also inform intuition regarding the advantages and limitations of different analytical techniques from the outset and are useful in deciding which variables should (and should not) be conditioned on to avoid subtle confounding and selection biases, arising from study design or data collection methods. Recognising these biases is necessary so that unbiased estimates of causal effects may be obtained.

Despite their utility, causal inference methods, and in particular causal graphs, do have limitations.

Unavoidably, expert knowledge is still required to elicit and defend causal assumptions, and it is recommended that sensitivity analyses be conducted to explore the consequences that departures from causal assumptions have on estimates of interest. Moreover, even in situation where causal assumptions may be well justified, correct specification of regression models remains an issue. These regression models may be required to adequately block open paths. In Section 2.3.1, we saw that specification of regression models is especially difficult in genomic applications, where dimensionality reduction strategies are required to condition on high-dimensional genetic information. These strategies come with their own model validity assumptions, separate from the causal ones we have discussed.

We reiterate that causal graphs are not the only framework for representing causal assumptions and deriving statistical dependencies, and that this can be done within other causal frameworks, for example Rubin (2005). We hope this review may, however, contribute to the discourse of GWAS and MR analyses by allowing causal assumptions to be explicitly acknowledged and communicated in a transparent and intuitive manner. Finally, since causal graphs are common in the communication and development of novel analytical methods, we hope to have contributed to a better understanding of them, thus helping the adoption of new analytical methods in the future.

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT

T: +44 (0)20 7299 4646
F: +44 (0)20 7299 4656
www.lshtm.ac.uk

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	1804954	**Title**	Mr.
First Name(s)	Oliver		
Surname/Family Name	Hines		
Thesis Title	Assumption-Lean Inference for Causal and Statistical Questions in the Era of Machine Learning		
Primary Supervisor	Karla Diaz-Ordaz		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?	Psychometrika		
When was the work published?	2021		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	**Yes**	Was the work subject to academic peer review?	**Yes**

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	
Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	This study was conceived by all authors. I carried out the mathematical research, computational simulations and writing of the manuscript under the supervision of the other authors.

SECTION E

Student Signature	Ohines
Date	14 December 2022

Supervisor Signature	Karla Diaz-Ordaz
Date	14 December 2022

Chapter 3

Partially linear mediation

3.1 Introduction

Testing and estimation of mediated effects is important in psychology, sociology, epidemiology, and econometrics, typically as a secondary analysis to understand the mechanism by which an exposure (X) effects an outcome (Y) through a mediating variable (M) (MacKinnon, 2008; Hayes, 2018). When the exposure is binary, one often considers the natural decomposition of the average treatment effect, into a natural indirect effect (NIDE), and natural direct effect (NDE) (Robins and Greenland, 1992; Pearl, 2001), which, under standard identifiability assumptions (sequential ignorability and consistency), may be written as functionals of the observed distribution (Imai et al., 2010a). These assumptions primarily require observation of a set of variables, (Z), that are sufficient to adjust for confounding of the association between X and M and between (X, M) and Y.

Assuming fully parametric models, maximum likelihood inference for the NIDE and NDE can be based on the so-called mediation formula (Pearl, 2001; VanderWeele and Vansteelandt, 2009; Imai et al., 2010a). Use of this formula gives rise to the popular difference and product-of-coefficient methods (Alwin and Hauser, 1975; MacKinnon et al., 2002) when simple linear models for the mediator and outcome hold (VanderWeele and Vansteelandt, 2009), but it readily allows extension to non-linear models (Imai et al., 2010a). A key concern about this approach is that misspecification of models for the mediator or outcome can lead to NIDE and NDE estimators with large bias; such misspecification can be difficult to diagnose when some confounders are strongly associated with either the exposure or the mediator (Vansteelandt, 2012).

In contrast, nonparametric inference gives rise to so-called triple robust estimators (Tchetgen Tchetgen and Shpitser, 2012) of the NIDE and NDE. These are less model-dependent, though still necessitate some form of modelling in view of the curse of dimensionality. In particular, these demand correct specification of an appropriate subset of: (i) the conditional expectation, $E(Y|M, X, Z)$, (ii) the conditional density of M given X, Z, and (iii) the conditional density of X given Z. These estimators are called triple robust due to their similarity with 'double robust' methods. Double robust methods of the average treatment effect, for example, are Consistent Asymptotically Normal (CAN) provided that either a mean outcome model, or propensity score model is correctly specified (Kang and Schafer, 2007). The triple robust estimator of the marginal NDE and NIDE is CAN provided any pair of (i), (ii), (iii) are correctly specified. These methods are also efficient (under the nonparametric model) provided that (i), (ii), and (iii) are all correctly specified. Additionally, Tchetgen Tchetgen and Shpitser (2014) provide CAN estimators for the parameters indexing a correctly specified parametric model for the conditional NDE given Z, provided either (i), or (ii) and (iii) are correctly specified.

Considering the common use of continuous measurements of mediator and outcome in psychology, which lend themselves to linear modelling, we will consider a different approach in the current work. In particular, we will continue to rely on linear modelling, but in view of the aforementioned concerns about model misspecification, will consider estimation and inference of the NIDE in a semi-parametric partially

linear model indexed by (β_1, β_2) which obeys

$$E(M|X,Z) = \beta_1 X + f(Z) \tag{3.1}$$
$$E(Y|M,X,Z) = \beta_2 M + g(X,Z) \tag{3.2}$$

where $g(x,z)$ and $f(x)$ are arbitrary functions. For the NDE, we consider the partially linear model indexed by (β_2, β_3) which obeys

$$E(Y|M,X,Z) = \beta_2 M + \beta_3 X + g(Z) \tag{3.3}$$

where $g(z)$ is an arbitrary function. The intersection of these models (i.e. when (3.1) and (3.3) both hold) is indexed by $\beta = (\beta_1, \beta_2, \beta_3)$. Early work by Baron and Kenny (1986) on the intersection model defined indirect and direct effects as the product $\beta_1\beta_2$ and coefficient β_3 respectively, with the total effect given by the sum of the two effects: $\beta_1\beta_2 + \beta_3$. When the exposure is binary and the intersection model holds, then the NIDE, NDE, and average treatment effect reduce to the effect definitions of Baron and Kenny (1986) (See Section 3.2 for details).

G-estimation is a method of parameter estimation in structural nested models, such as those in (3.1) and (3.3) developed by James Robins (and collaborators) over a number of years (Robins, 1994; Vansteelandt and Joffe, 2014; Naimi et al., 2017). In the current work, G-estimators for the NIDE and NDE are constructed, assuming that the mediator and outcome mean models are partially linear. Denoting, $h(Z) = E(X|Z)$, we show that our G-estimator for the NIDE is CAN when either

(a) (3.1) and (3.2) hold and a parametric model for $f(z)$ is correctly specified

(b) (3.1) and (3.3) hold and parametric models for both $g(z)$ and $h(z)$ are correctly specified

and for the NDE, our G-estimator is CAN when either

(c) (3.3) holds and a parametric model for $g(z)$ is correctly specified

(d) (3.1) and (3.3) hold and parametric models for both $f(z)$ and $h(z)$ are correctly specified

Compared with the triple robust estimators of Tchetgen Tchetgen and Shpitser (2012); Tchetgen Tchetgen and Shpitser (2014) (which do not require partial linearity), the proposed G-estimation methods (which do require some partial linearity) have the advantage that conditional densities for M and X do not need to specified, and conditional mean models for Y and M are sufficient to estimate the NIDE and NDE respectively. Extensions to the G-estimation methods where partial linearity is violated are discussed in Section 3.8.

We also consider testing of the no-mediation hypothesis ($H_0 : \beta_1\beta_2 = 0$) and the no-direct-effect hypothesis ($H_1 : \beta_3 = 0$). Testing of the no-mediation hypothesis, H_0, is problematic since the function used to constrain the hypothesized parameter space ($\psi_0(\beta) = \beta_1\beta_2 = 0$) has a Jacobian which is full rank almost everywhere, except for a singular point at $\beta_1 = \beta_2 = 0$. This generally gives rise to test statistics with different asymptotic behaviour at this singular point, and in finite samples, tests for H_0 which are underpowered in its neighbourhood. We refer the interested reader to Dufour et al. (2013); Drton and Xiao (2016) for further details.

Ordinary Least Squares (OLS) is routinely used for estimation of the target parameter β and nuisance parameter vector $\gamma = (\gamma_{m0}, \gamma_m, \gamma_{y0}, \gamma_y)$ in the intersection model when $f(z)$ and $g(z)$ are parametrically defined by $f(z) = \gamma_{m0} + \gamma_m^\top z$ and $g(z) = \gamma_{y0} + \gamma_y^\top z$. Here γ_{m0} and γ_{y0} represent scalar intercept terms and γ_m and γ_y are parameter vectors. Classical tests of the no-mediation hypothesis in this setting are constructed from the squared t-test statistics, $T_j^{(OLS)} = (\hat{\beta}_j^{(OLS)}/\hat{\sigma}_j^{(OLS)})^2$, where for $j = 1, 2, 3$, $\hat{\beta}_j^{(OLS)}$ denotes the OLS estimator, with estimated standard error, $\hat{\sigma}_j^{(OLS)}$.

Using these squared t-statistics, a Wald test for H_0, also known as the Sobel test (Sobel, 1982) can be constructed, based on the test statistic $W^{(OLS)}$. Alternatively, a joint significance test (also known as a

Likelihood Ratio (LR) test)(MacKinnon et al., 2002; Giersbergen, 2014), has been constructed, based on the statistic $LR^{(OLS)}$. These test statistics are

$$W^{(OLS)} = \frac{T_1 T_2}{T_1 + T_2} = \frac{\hat{\beta}_1^2 \hat{\beta}_2^2}{\hat{\beta}_1^2 \hat{\sigma}_2^2 + \hat{\beta}_2^2 \hat{\sigma}_1^2} \tag{3.4}$$

$$LR^{(OLS)} = \min(T_1, T_2) \tag{3.5}$$

where, for readability, the superscript (OLS) has been dropped from all terms on the right hand side. These statistics have received considerable attention, especially due to unexpected properties regarding the relative power of total, direct, and indirect effect tests under different true parameter values (Wang, 2018; Kenny and Judd, 2014; Fritz et al., 2012).

We propose two alternative tests based on moment conditions of the G-estimator: a Wald type approach, and an approach analogous to a classical score (or Lagrange Multiplier) test, but derived using a Generalized Methods of Moments (GMM) hypothesis testing framework (Newey and West, 1987; Dufour et al., 2017). The relative merits of the new tests against the OLS based tests above are discussed in Section 3.5.3. From a robustness perspective, tests based on OLS estimating equations require that (3.1) and (3.3) hold and $f(z)$ and $g(z)$ are correctly specified, whereas those based on the G-estimation equations inherit the same robustness to model misspecification as the G-estimator itself, provided that nuisance parameters are estimated orthogonally (in a sense defined in Section 3.4). A simulation study is carried out in Section 3.6 to assess the behaviour of the new robust tests in finite samples, followed by an illustration on clinical data in Section 3.7. All methods are made available through an R package, which can be found at `github.com/ohines/plmed`.

3.2 Identifiability

Suppose iid data on (Y, M, X, Z) is collected for n individuals. We assume that there exists a potential outcome variable $Y(x, m)$, which expresses the outcome that would have been observed if the exposure and mediator had taken the values (x, m). Similarly, we assume a potential outcome, $M(x)$, corresponding to the mediator if the exposure had taken the value x. We define the expected potential outcome

$$\eta(x, x^*, z) = E[Y(x, M(x^*))|Z = z]$$

for arbitrary (x, x^*) on the support of X. We define the NIDE and NDE, conditional on $Z = z$, respectively by

$$E[Y(x_0, M(x_1)) - Y(x_0, M(x_0))|Z = z]$$
$$E[Y(x_1, M(x_1)) - Y(x_0, M(x_1))|Z = z]$$

for two pre-specified levels of the exposure, (x_0, x_1). For a binary exposure coded 0 or 1, $(x_0, x_1) = (0, 1)$, however our definition also permits continuous exposures. Letting $P(m|x, z)$ denote the probability measure of M conditional on $X = x$ and $Z = z$, then

$$\eta(x, x^*, z) = \int E(Y|X = x, M = m, Z = z) dP(m|x^*, z)$$

under standard identifiability assumptions (Pearl, 2001; Imai et al., 2010a,b; VanderWeele and Vansteelandt, 2009). These assumptions require consistency,

$$X = x \implies M(x) = M \text{ almost surely}$$
$$X = x \text{ and } M = m \implies Y(x, m) = Y \text{ almost surely}$$

and sequential ignorability, which states that for all m on the support of M,

$$Y(x, m) \perp\!\!\!\perp M|X = x^*, Z$$
$$(Y(x, m), M(x^*)) \perp\!\!\!\perp X|Z$$

Here $A \perp\!\!\!\perp B|C$ denotes independence of A and B conditional on C. Under these identifiability assumptions and the partial linearity in (3.2), then

$$\eta(x, x^*, z) = \beta_2 f(x^*, z) + g(x, z)$$

where $f(X, Z) = E(M|X, Z)$. We see, therefore, that one obtains the following two expressions for the conditional NIDE and NDE when (3.1) and (3.3) hold respectively,

$$\eta(x_0, x_1, z) - \eta(x_0, x_0, z) = \beta_1\beta_2(x_1 - x_0)$$
$$\eta(x_1, x_1, z) - \eta(x_0, x_1, z) = \beta_3(x_1 - x_0)$$

for all z. It follows that when (3.1) and (3.2) hold then the product of coefficients $\beta_1\beta_2$, represents the conditional NIDE per unit change in X, and when (3.3) holds then β_3 represents the conditional NDE per unit change in X. Since these effects are constant then the marginal effects are equal to the conditional effects. By way of comparison, Tchetgen Tchetgen and Shpitser (2014) consider estimation of a parameter ψ which indexes parametric models for the NDE conditional on $Z = z$, when the exposure is binary,

$$\eta(1, 1, z) - \eta(0, 1, z) = \delta(z, \psi)$$

where $\delta(z, \psi)$ is a known function. Our methods, in effect, consider the case $\delta(z, \psi) = \psi$, i.e. that the NDE is constant in subgroups of Z, however, we additionally require that partially linear models hold. These assumptions are relaxed in Section 3.8.

3.3 The G-estimator for mediation

Our objective is to derive direct and indirect effect estimators which are asymptotically linear and hence CAN in the sense that they asymptotically follow normal distributions centred at the true value, with variances of order n^{-1}. We refer readers to Kennedy (2015) for an introduction to asymptotically linear estimators and influence function theory in causal inference.

We will consider the target parameter, $\beta = (\beta_1, \beta_2, \beta_3)$ in the intersection model (i.e. when (3.1) and (3.3) both hold), and begin by introducing parametric working models for $h(z), f(z), g(z)$ which we denote $h(z; \gamma_x), f(z; \gamma_m), g(z; \gamma_y)$ where h, f, g are known differentiable functions parametrized by the nuisance parameter vector $\gamma = (\gamma_x, \gamma_m, \gamma_y)$. This nuisance parameter and the target parameter itself will be estimated jointly. Specifically, we consider an iterative estimation procedure by which the nuisance parameter estimate is obtained from a previous target parameter estimate using a CAN estimator $\hat{\gamma} = \hat{\gamma}(\hat{\beta})$ which is consistent in the sense described by assumptions A1 to A3 below. The target parameter estimate may then be updated using the updated nuisance parameter estimate. We assume that each component of the nuisance parameter estimator is consistent when the associated part of the model is correctly specified, that is, denoting the correct parameter values by superscript 0, we assume,

A1. If $h(z)$ is correctly specified then $\operatorname{plim} \hat{\gamma}_x(\beta) = \gamma_x^0$ for all β

A2. If $f(z)$ is correctly specified then $\operatorname{plim} \hat{\gamma}_m(\beta) = \gamma_m^0$ for all β such that $\beta_1 = \beta_1^0$

A3. If $g(z)$ is correctly specified then $\operatorname{plim} \hat{\gamma}_y(\beta) = \gamma_y^0$ for all β such that $(\beta_2, \beta_3) = (\beta_2^0, \beta_3^0)$

where assumptions A2 and A3 are only well defined when (3.1) and (3.3) hold respectively. For the target parameter, we propose estimation based on the product of residuals in the intersection model given by the vector $U(\beta, \gamma)$, which we refer to as the G-moment conditions, with components,

$$U_1(\beta, \gamma) = \{X - h(Z; \gamma_x)\}\{M - \beta_1 X - f(Z; \gamma_m)\} \tag{3.6}$$
$$U_2(\beta, \gamma) = \{M - \beta_1 X - f(Z; \gamma_m)\}\{Y - \beta_2 M - \beta_3 X - g(Z; \gamma_y)\} \tag{3.7}$$
$$U_3(\beta, \gamma) = \{X - h(Z; \gamma_x)\}\{Y - \beta_2 M - \beta_3 X - g(Z; \gamma_y)\} \tag{3.8}$$

When all models are correctly specified, these residual products are zero in expectation, i.e. $E\{U(\beta^0, \gamma^0)\} = 0$. The G-estimator for β, denoted by $\hat{\beta}$ is the value which sets the sample average of moment conditions to zero, that is it solves the system of three equations

$$E_n[U\{\hat{\beta}, \hat{\gamma}\}] = 0 \tag{3.9}$$

where $E_n[.] = n^{-1} \sum_{i=1}^{n} [.]_i$ is the expectation with respect to the empirical distribution of the data. To examine the behaviour of this estimator under model misspecification, we introduce notation for the probability limit of this G-estimator, $\beta^* = \operatorname{plim} \hat{\beta}$, with associated nuisance parameter estimator limit, $\gamma^* = \operatorname{plim} \hat{\gamma}(\beta^*)$. The probability limit of the G-estimator is the solution to

$$E\{U(\beta^*, \gamma^*)\} = 0$$

We additionally assume (A4) that the 3 by 3 matrix,

$$E\left(\frac{\partial U(\beta^*, \gamma^*)}{\partial \beta}\right)$$

is non-singular. Finally we assume (A5) that β^* is unique. This assumption can be partly justified by the linearity of the G-moment conditions, which implies that β^* is unique when we disregard the dependence of γ^* on β^*, treating γ^* as constant. It is sufficient, therefore, to assume that there is no pathological way by which the nuisance parameter estimator might introduce extra solutions.

Under assumptions A1 to A5 one can derive the conditions for which the G-estimators of (β_1, β_2) and of (β_2, β_3) are consistent, by examining the conditions under which all three moment conditions are zero in expectation. These results are given in Lemmas 1 and 2 respectively. For completeness, the intersection of these two cases, under which $(\beta_1, \beta_2, \beta_3)$ is consistent, is given by Lemma 3. See Appendix B for proofs.

Lemma 1 (Consistency of the G-estimator of (β_1, β_2)) *Provided the models for M and Y are partially linear in X and M respectively, such that (3.1) and (3.2) both hold, and either*

(i) The model for $f(Z)$ is correctly specified

(ii) $g(X, Z) = \beta_3 X + g(Z)$ and the models for $g(Z)$ and $h(Z)$ are both correctly specified

then $\beta^* = (\beta_1^0, \beta_2^0, \beta_3^*)$, hence the G-estimator is consistent for (β_1, β_2).

Lemma 2 (Consistency of the G-estimator of (β_2, β_3)) *Provided the model for Y is partially linear in (M, X) such that (3.3) holds and either*

(i) The model for $g(Z)$ is correctly specified

(ii) $f(X, Z) = \beta_1 X + f(Z)$ and the models for $f(Z)$ and $h(Z)$ are both correctly specified

then $\beta^* = (\beta_1^*, \beta_2^0, \beta_3^0)$, hence the G-estimator is consistent for (β_2, β_3).

Lemma 3 (Robustness of G moment conditions) *Provided that the models for M and Y are partially linear in X and (M, X) respectively, such that (3.1) and (3.3) both hold and any pair of*

(i) The model for $h(Z)$

(ii) The model for $f(Z)$

(iii) The model for $g(Z)$

are correctly specified, then $\beta^* = (\beta_1^0, \beta_2^0, \beta_3^0)$, hence the G-estimator is consistent for the full target parameter. For proof observe that the conditions in Lemmas 1 and 2 are satisfied.

Theorem 1 describes the conditions under which the G-estimator will be asymptotically linear.

Theorem 1 (Asymptotically linearity of $\hat{\beta}$) *Let β^* denote the probability limit of the G-estimator, as set out in Lemmas 1 to 3, and assume the nuisance parameter estimator, $\hat{\gamma}(\beta)$ is CAN and obeys assumptions A1 to A3 such that*

$$\hat{\gamma} - \gamma^* = E_n\{\phi(\beta^*, \gamma^*)\} + o_p\left(n^{-1/2}\right) \tag{3.10}$$

where o_p denotes stochastic order notation so that $A_n = o_p(r_n^{-1})$ means that $A_n r_n \xrightarrow{p} 0$ and $\xrightarrow{p}$ denotes convergence in probability. Then, subject to regularity conditions, the estimator $\hat{\beta}$ is CAN

$$\hat{\beta} - \beta^* = E_n\{\varphi(\beta^*, \gamma^*)\} + o_p\left(n^{-1/2}\right)$$

with influence function $\varphi(.)$ given by

$$\varphi(\beta^*, \gamma^*) = E\left\{-\frac{\partial U(\beta^*, \gamma^*)}{\partial \beta}\right\}^{-1}\left(U(\beta^*, \gamma^*) + E\left\{\frac{\partial U(\beta^*, \gamma^*)}{\partial \gamma}\right\}\phi(\beta^*, \gamma^*)\right). \tag{3.11}$$

See Appendix B for proof.

The G-estimator of the NIDE under the model in (3.1)–(3.2) is the product $\hat{\beta}_1\hat{\beta}_2$ with influence function,

$$\omega(\beta^*, \gamma^*) = \beta_1\varphi_2(\beta^*, \gamma^*) + \beta_2\varphi_1(\beta^*, \gamma^*) \tag{3.12}$$

where $\beta^* = (\beta_1, \beta_2, \beta_3^*)$ and for $j = 1, 2, 3$, $\varphi_j(\beta^*, \gamma^*)$ is the jth component of the influence function in (3.11). Derivation of this influence function can be found in Appendix B. Similarly, the G-estimator of the NDE under model (3.3) is $\hat{\beta}_3$ with influence function $\varphi_3(\beta^*, \gamma^*)$ where $\beta^* = (\beta_1^*, \beta_2, \beta_3)$ and with consistency guaranteed under the conditions in Lemma 2.

When, in truth, $(\beta_1, \beta_2) = (0, 0)$ then the (first-order) influence function in (3.12) is exactly zero. In this case the NIDE estimator $\hat{\beta}_1\hat{\beta}_2$ is asymptotically linear and CAN in the sense that $n^{1/2}\hat{\beta}_1\hat{\beta}_2$ asymptotically follows a normal distribution with zero variance. Multiplying $\hat{\beta}_1\hat{\beta}_2$ by higher powers of n yields more interesting behaviour. When all models are correctly specified, $n\hat{\beta}_1\hat{\beta}_2$ asymptotically follows a 'product normal' distribution (the distribution of two mean zero normal variables with known variances) (Aroian, 1947).

3.4 Nuisance parameter estimation

Theoretical results show that the choice of nuisance parameter estimators does not impact the asymptotic variance of double robust estimators when both working models are correctly specified (Tsiatis, 2006). Similarly, in our case, it is straightforward to show that, when the models for X, M and Y are correctly specified then

$$E\left\{\frac{\partial U(\beta^*, \gamma^*)}{\partial \gamma}\right\} = 0 \tag{3.13}$$

So the influence function in (3.11) does not depend on the nuisance influence function $\phi(\beta^*, \gamma^*)$. This property is sometimes referred to as (Neyman) orthogonality (Neyman, 1959; Chernozhukov et al., 2017), with the intuition that the G-moment conditions are locally insensitive to nuisance parameters when all models are correctly specified. Orthogonal estimators are particularly useful for the construction of score tests, which we describe in Section 3.5.1. Moreover, they ensure that our asymptotic results continue to be valid when consistent variable selection procedures (e.g. lasso) are employed for selecting confounders in each of the three working models.

When (3.13) is not satisfied, as may happen under model misspecification, then the influence function in (3.11) does depend on the nuisance influence function $\phi(\beta^*, \gamma^*)$. Under model misspecification, (3.11)

therefore represents a class of G-estimators, indexed by the choice of nuisance parameter estimation method. Choosing the nuisance parameter estimator under misspecification is non-trivial as it may greatly affect the asymptotic variance of the estimator. Various proposals have been suggested for conventional double robust estimators which aim to minimize either the variance under misspecification (Rotnitzky and Vansteelandt, 2014) or the bias when models are misspecified (Vermeulen and Vansteelandt, 2015; Avagyan and Vansteelandt, 2017). This second approach, referred to as a bias-reduction strategy, involves constructing a nuisance parameter estimator, which is pseudo-orthogonal to the target parameter estimator so that (3.13) is approximately satisfied. In effect, (3.13) is used as a set of moment conditions by which the nuisance parameters are estimated.

To implement the bias-reduction strategy for our G-estimator, we must first augment the G-moment functions (3.6)–(3.8), such that each estimating equation has a unique nuisance parameter. In practice, this means, for example, that different estimators of γ_x may be used in (3.6) and (3.7), which are both consistent when $h(z)$ is correctly specified, with the same being true of γ_m and γ_y. Denoting the augmented nuisance parameters with superscript (1) and (2), the augmented moment functions are given by

$$U_1(\beta,\gamma) = \left\{X - h\left(Z;\gamma_x^{(1)}\right)\right\}\left\{M - \beta_1 X - f\left(Z;\gamma_m^{(1)}\right)\right\} \tag{3.14}$$

$$U_2(\beta,\gamma) = \left\{M - \beta_1 X - f\left(Z;\gamma_m^{(2)}\right)\right\}\left\{Y - \beta_2 M - \beta_3 X - g\left(Z;\gamma_y^{(1)}\right)\right\} \tag{3.15}$$

$$U_3(\beta,\gamma) = \left\{X - h\left(Z;\gamma_x^{(2)}\right)\right\}\left\{Y - \beta_2 M - \beta_3 X - g\left(Z;\gamma_y^{(2)}\right)\right\} \tag{3.16}$$

with full nuisance parameter, $\gamma = (\gamma_x^{(1)}, \gamma_x^{(2)}, \gamma_m^{(1)}, \gamma_m^{(2)}, \gamma_y^{(1)}, \gamma_y^{(2)})$. The bias-reduced nuisance parameter estimator is that which solves

$$E_n\left\{\frac{\partial U(\hat{\beta},\hat{\gamma})}{\partial \gamma}\right\} = 0$$

This estimator is asymptotically linear and obeys the consistency assumptions A1 to A3. For identifiability, we require models where $\mathrm{Dim}(\gamma_x) = \mathrm{Dim}(\gamma_m) = \mathrm{Dim}(\gamma_y)$. Such restrictions are not uncommon e.g. Rotnitzky et al. (2012) and may be satisfied by enlarging the working models. The accompanying `plmed` package implements G-estimation methods with bias-reduced parameter estimation in the setting where $f(z,\gamma_m)$ and $g(z,\gamma_y)$ are linear predictors and $h(z,\gamma_x)$ is modelled by a Generalized Linear Model (GLM).

3.5 Hypothesis testing

We now consider tests of the null hypothesis, $H_\alpha : (\alpha-1)\beta_1\beta_2 + \alpha\beta_3 = 0$, with $\alpha \in [0,1]$ known. This hypothesis includes the no-mediation hypothesis ($\alpha = 0$) and the no-direct effect hypothesis ($\alpha = 1$) as special cases. We begin by constructing a score test, for general α, based on the G-moment conditions. Provided the nuisance parameters are orthogonally estimated, the score test is robust to certain model misspecification. Also, in the specific case where $\alpha = 0$ and the true parameter takes the value $(\beta_1^0, \beta_2^0) = (0,0)$, the score test is conservative, in the sense that the Type I error rate is below the nominal test size.

The score test is compared with Wald tests for the special cases of the no-mediation hypothesis and the no-direct-effect hypothesis. These Wald tests are constructed using the influence function of the G-estimator, and inherit the robustness properties of the G-estimator, without requiring orthogonal nuisance parameter estimation.

Sobel (1982) proposed a Wald test of the no-mediation hypothesis, based on the OLS moment conditions. Our Wald test for the no-mediation hypothesis is similar enough to Sobel's work that we shall refer to it as the Robust Sobel test (or Robust Wald test).

3.5.1 The Score Test

The score test is based on the observation that, since $E\{U(\beta^*, \gamma^*)\} = 0$, the classical central limit theorem implies

$$n^{1/2} E_n\{U(\beta^*, \gamma^*)\} \xrightarrow{d} \mathcal{N}\left(0, E\left\{U(\beta^*, \gamma^*) U(\beta^*, \gamma^*)^{\mathrm{T}}\right\}\right)$$
$$n E_n\{U(\beta^*, \gamma^*)\}^{\mathrm{T}} E\{U(\beta^*, \gamma^*) U(\beta^*, \gamma^*)^{\mathrm{T}}\}^{-1} E_n\{U(\beta^*, \gamma^*)\} \xrightarrow{d} \chi^2_3 \tag{3.17}$$

where $\xrightarrow{d}$ denotes convergence in distribution (as $n \to \infty$) and $\mathcal{N}(\mu, \Sigma)$ and χ^2_r respectively denote a normal distribution with mean μ and covariance Σ, and a chi-squared distribution with r degrees of freedom. The left hand side of (3.17) is similar in form to a GMM estimator, based on the objective function

$$M_n(\beta, \gamma, I) = E_n\{U(\beta, \gamma)\}^{\mathrm{T}} I^{-1} E_n\{U(\beta, \gamma)\} \quad \geq 0 \quad \forall (\beta, \gamma)$$

where I is a positive semi-definite 3 by 3 matrix. The GMM estimator of β is the minimizer $\arg\min_\beta M_n(\beta, \hat{\gamma}(\beta), I)$. In our case the GMM estimator is said to be exactly specified since $\mathrm{Dim}(\beta) = \mathrm{Dim}(U(\beta, \gamma))$. In this exactly specified setting, minimization of the GMM objective function is equivalent to solving (3.9) and the estimator is independent of the choice of weighting matrix, I.

The minimization of $M_n(\beta, \hat{\gamma}(\beta), I)$ over a constrained parameter space, however, may be exploited for hypothesis testing, using results by Newey and West (1987) for the GMM Two-Step estimator, later extended by Dufour et al. (2017) to the GMM-Continuous Updating Estimator (CUE), both discussed below.

Work by Hansen (1982) in the over-specified setting, (i.e. when $\mathrm{Dim}(\beta) < \mathrm{Dim}\{U(\beta, \gamma)\}$), showed that the optimal GMM estimator is constructed using weights proportional to the variance matrix, $I \propto E\{U(\beta^*, \gamma^*) U(\beta^*, \gamma^*)^{\mathrm{T}}\}$. This is optimal in the sense that the asymptotic covariance matrix of the resulting estimator is as small as possible (in the positive definite sense) among the class of GMM estimators.

This optimal choice also lends itself to hypothesis testing, as suggested by (3.17). In this work we consider minimization of the objective function, M_n, under two proposals. The first (Two-Step) proposal first estimates nuisance parameters and the variance matrix $E\{U(\beta^*, \gamma^*) U(\beta^*, \gamma^*)^{\mathrm{T}}\}$. Then, using these initial estimates, constrained estimates of β are obtained by a subsequent minimization of the GMM estimator. The second, (CUE) proposal allows the estimates of nuisance parameters and the variance matrix I to be updated continuously. Writing,

$$\hat{I}_n(\beta, \gamma) = E_n\{U(\beta, \gamma) U(\beta, \gamma)^{\mathrm{T}}\}$$

then the proposed Two-Step and CUE objective functions are respectively given by

$$S(\beta) = n M_n\Big(\beta, \hat{\gamma}(\hat{\beta}), \hat{I}_n(\hat{\beta}, \hat{\gamma}(\hat{\beta}))\Big) \tag{3.18}$$
$$\tilde{S}(\beta) = n M_n\Big(\beta, \hat{\gamma}(\beta), \hat{I}_n(\beta, \hat{\gamma}(\beta))\Big) \tag{3.19}$$

where $\hat{\beta}$ is the unconstrained G-estimate of β. Defining the null parameter space as $B_\alpha = \{\beta | (\alpha - 1)\beta_1\beta_2 + \alpha\beta_3 = 0\}$, the Two-step and CUE score type test statistics may be written as

$$S_\alpha = \min_{\beta \in B_\alpha} S(\beta)$$
$$\tilde{S}_\alpha = \min_{\beta \in B_\alpha} \tilde{S}(\beta)$$

In practice, computation of the Two-Step score statistic may be achieved using the method of Lagrange Multipliers to construct estimating equations for the constrained minimization problem. These may then be solved with a Newton-Raphson scheme. Similarly, computation of the CUE score statistic can be achieved using Lagrange Multipliers to construct estimating equations for β, however the Newton-Raphson

procedure should additionally include the nuisance parameter estimating equations. When the bias-reduced nuisance estimation strategy is used, computation by Newton-Raphson requires that $f(.)$, $g(.)$ and $h(.)$ are twice continuously differentiable.

To derive the asymptotic distributions of S_α and $\tilde{S}_\alpha$, we consider the general problem of minimizing over some hypothesis set $B_\psi = \{\beta|\psi(\beta) = 0\}$, where $\psi(.)$ is a differentiable function. For example, when $\psi(\beta) = \beta - \beta^*$ then B_ψ represents a single point, and the asymptotic distribution in (3.17) is recovered. Theorem 2 gives the general result for the asymptotic distribution of the objective functions, which follows by extending results for the test statistic in Section 5.1 of Dufour et al. (2017), with related work by Newey and West (1987). Our result accommodates nuisance parameter estimation and relies on the orthogonality of the nuisance parameter estimator, see Appendix B for details.

Theorem 2 (Constrained GMM) *Consider a null hypothesis $H_0 : \psi(\beta^*) = 0$, where ψ is a vector of dimension $r \in \{1, 2, 3\}$ and is continuously differentiable in some non-empty, open neighbourhood, N, of the true limiting value β^*. Provided that for all $\beta \in N$*

$$\text{Rank}\left(\frac{\partial\psi(\beta)}{\partial\beta}\right) = r \tag{3.20}$$

and $\hat{\gamma}$ is estimated orthogonally, in the sense that (3.13) *holds, then for $B_\psi = \{\beta|\psi(\beta) = 0\}$,*

$$\min_{\beta\in B_\psi} S(\beta) \xrightarrow{d} \chi^2_r$$
$$\min_{\beta\in B_\psi} \tilde{S}(\beta) \xrightarrow{d} \chi^2_r.$$

Applying Theorem 2 to the target hypothesis, H_α, we see that the rank condition in (3.20) is not necessarily satisfied for the no-mediation hypothesis ($\alpha = 0$). Letting $\psi_\alpha(\beta) = (\alpha - 1)\beta_1\beta_2 + \alpha\beta_3$ then

$$\text{Rank}\left(\frac{\partial\psi_\alpha(\beta)}{\partial\beta}\right) = \text{Rank}\begin{pmatrix}(\alpha-1)\beta_2\\(\alpha-1)\beta_1\\\alpha\end{pmatrix} = \begin{cases}0 & \text{for } \alpha = \beta_1 = \beta_2 = 0\\1 & \text{otherwise}\end{cases}$$

Therefore for $\alpha \neq 0$ one may apply the result in Theorem 2 directly to construct a test which rejects H_α when $S_\alpha > c$ for some critical value, c. This test size has size $1 - F_{\chi^2_1}(c)$ where $F_{\chi^2_1}(x)$ is the distribution function of a χ^2_1 variable. One can show that this test is also a valid test of the no-mediation hypothesis, H_0. To do so, we define the null parameter space $B_0 = \{\beta|\beta_1\beta_2 = 0\}$ and let $C_j = \{\beta|\beta_j = 0\}$ for $j = 1, 2$ so that $B_0 = C_1 \cup C_2$. Hence

$$S_0 = \min\left\{\min_{\beta\in C_1} S(\beta), \min_{\beta\in C_2} S(\beta)\right\}$$
$$\tilde{S}_0 = \min\left\{\min_{\beta\in C_1} \tilde{S}(\beta), \min_{\beta\in C_2} \tilde{S}(\beta)\right\}.$$

Under H_0 we know that either $\beta_1 = 0$ or $\beta_2 = 0$ and since the constraint function $\psi_j(\beta) = \beta_j$ does satisfy the rank condition in (3.20), one can show, that under H_0,

$$\sup_{\beta^*\in B_0} P_{\beta^*}(S_0 > x) \to 1 - F_{\chi^2_1}(x) \tag{3.21}$$
$$\sup_{\beta^*\in B_0} P_{\beta^*}\left(\tilde{S}_0 > x\right) \to 1 - F_{\chi^2_1}(x) \tag{3.22}$$

where P_{β^*} denotes the probability measure with a true limiting parameter value of β^* and $\to$ denotes convergence as n tends to infinity. See Appendix B for details.

3.5.2 Wald Tests

Using the asymptotic linearity of $\hat{\beta}_1\hat{\beta}_2$ in (3.12) and under the conditions of Lemma 1, one can demonstrate that $n^{1/2}\hat{\beta}_1\hat{\beta}_2$ asymptotically follows a normal distribution when $(\beta_1, \beta_2) \neq (0, 0)$. Estimating the variance of $n^{1/2}\hat{\beta}_1\hat{\beta}_2$ by $E_n\{\omega^2(\hat{\beta}, \hat{\gamma}(\hat{\beta}))\}$ one arrives at a Wald test statistic, W, for the no-mediation hypothesis: $\beta_1\beta_2 = 0$

$$
\begin{aligned}
n^{1/2}(\hat{\beta}_1\hat{\beta}_2 - \beta_1\beta_2) &\xrightarrow{d} \mathcal{N}\left(0, E\{\omega^2(\beta_0, \gamma^*)\}\right) \\
W &= \frac{n\hat{\beta}_1^2\hat{\beta}_2^2}{E_n\{\omega^2(\hat{\beta}, \hat{\gamma}(\hat{\beta}))\}} \\
&= \frac{T_1T_2}{T_1 + T_2 + 2\rho\sqrt{T_1T_2}} \\
&= \frac{\hat{\beta}_1^2\hat{\beta}_2^2}{\hat{\beta}_1^2\hat{\sigma}_2^2 + \hat{\beta}_2^2\hat{\sigma}_1^2 + 2\hat{\beta}_1\hat{\beta}_2\Delta}
\end{aligned}
$$

where for $j = 1, 2, 3$, the squared t-statistic is represented by $T_j = \hat{\beta}_j^2/\hat{\sigma}_j^2$, and $\rho = \Delta/\hat{\sigma}_1\hat{\sigma}_2$ with $\hat{\sigma}_j^2$ and Δ given by $n^{-1}E_n\{\varphi_j^2(\hat{\beta}, \hat{\gamma})\}$ and $n^{-1}E_n\{\varphi_1(\hat{\beta}, \hat{\gamma})\varphi_2(\hat{\beta}, \hat{\gamma})\}$ respectively.

The distribution of W is problematic since at the true parameter value $(\beta_1, \beta_2) = (0, 0)$, then $\text{var}\left(n^{1/2}\hat{\beta}_1\hat{\beta}_2\right) \to 0$ as $n \to \infty$. A characterisation of Wald-type statistics for testing polynomial constraints with singular points is given by Dufour et al. (2013). For the constraint $\beta_1\beta_2 = 0$, Glonek (1993) demonstrated that

$$
W \xrightarrow{d} \begin{cases} \frac{1}{4}\chi_1^2 & \text{for } \beta_1 = \beta_2 = 0 \\ \chi_1^2 & \text{otherwise} \end{cases}
$$

This result suggests that one may reject the no-mediation hypothesis when the Wald statistic exceeds some critical value, c, chosen with reference to the χ_1^2 distribution. Such a test will have size $1 - F_{\chi_1^2}(c)$ when the null is satisfied but one of β_1 or β_2 is non-zero, and will be conservative when $\beta_1 = \beta_2 = 0$. The fact that W behaves differently at a singular point is known to greatly restrict the power of the Wald test to detect small indirect effects in finite samples (MacKinnon et al., 2002).

Construction of a Wald based test for the NDE is fairly trivial. Under the conditions of Lemma 2, the squared t-statistic, $T_3 \xrightarrow{d} \chi_1^2$ when $\beta_3^0 = 0$.

3.5.3 Comparison of methods

Revisiting the classical tests for the no-mediation hypothesis, as given in (3.4) and (3.5) we see that $0 \leq W^{(OLS)} \leq LR^{(OLS)}$ with equality as $T_j^{(OLS)}$ approaches infinity for either $j = 1$ or $j = 2$, which occurs in the asymptotic limit when $\beta_j \neq 0$. In fact, away from the singularity at $\beta_1^0 = \beta_2^0 = 0$, both statistics have the same χ_1^2 asymptotic distribution and (including the singular point) a test which rejects when $W^{(OLS)} > c$ has equal size to that which rejects when $LR^{(OLS)} > c$ for some critical value c. Hence, the test based on $LR^{(OLS)}$ is uniformly more powerful (van Garderen and van Giersbergen, 2019).

We highlight this comparison between the two classical tests because it gives some intuition as to why the G-estimation score test, which we argue is analogous to $LR^{(OLS)}$, might be more powerful than the G-estimation Wald test, analogous to $W^{(OLS)}$. The analogy is made clearer by rewriting $LR^{(OLS)}$ as a minimization over an objective function.

$$
\begin{aligned}
S^{(OLS)}(\beta) &= \sum_{j=1}^{2}\left(\frac{\hat{\beta}_j^{(OLS)} - \beta_j}{\hat{\sigma}_j^{(OLS)}}\right)^2 \\
LR^{(OLS)} &= \min_{\{\beta|\beta_1\beta_2=0\}} S^{(OLS)}(\beta)
\end{aligned}
\tag{3.23}
$$

This objective function resembles a sum of OLS squared t-statistics, minimization of which (under the constraint $\beta_1\beta_2 = 0$), either sets $(\beta_1, \beta_2) = (0, \hat{\beta}_2^{(OLS)})$ or $(\hat{\beta}_1^{(OLS)}, 0)$, thus removing the contribution of a single term from the sum. To demonstrate the analogy between $S^{(OLS)}(\beta)$ and our G-estimation score objective functions in (3.18) and (3.19), we consider the case where (3.1) and (3.3) hold and $f(z), g(z)$ and $h(z)$ are correctly specified.

In this setting, the G-estimating equations are always orthogonal to the nuisance parameter estimates. For illustration we additionally assume that $\text{var}\,(Y|M, X, Z) = \text{var}\,(Y|X, Z)$, so that the true covariance matrix, $I = E\{U(\beta,\gamma)U(\beta,\gamma)^{\text{T}}\}$ is diagonal. Under these assumptions, the squared t-statistics for the null hypothesis $\beta_j = 0$, are given by

$$T_j = \frac{\hat{\beta}_j^2}{\hat{\sigma}_j^2} = \frac{nE_n\{\varphi_j(\hat{\beta}_{-j}, \hat{\gamma})\}^2}{E_n\{\varphi_j^2(\hat{\beta}, \hat{\gamma})\}} \text{ for j = 1,2,3}$$

which reduces to

$$T_j = \frac{nE_n\{U_j(\hat{\beta}_{-j}, \hat{\gamma})\}^2}{E_n\{U_j^2(\hat{\beta}, \hat{\gamma})\}} \text{ for j = 1,2}$$

$$T_3 = \frac{nE_n\{U_3(\hat{\beta}_{-3}, \hat{\gamma})\}^2}{E_n\{U_3^2(\hat{\beta}, \hat{\gamma})\} + \hat{\beta}_1^2 E_n\{\widehat{\text{var}}\,(X|Z)\}E_n\{\widehat{\text{var}}\,(M|X, Z)\}^{-1}E_n\{U_2^2(\hat{\beta}, \hat{\gamma})\}}$$

where $\hat{\beta}_{-j}$ denotes the G-estimate of β with jth parameter set to zero and $\widehat{\text{var}}\,(.)$ denotes conditional variance estimated using the parameter G-estimates. Note that the denominator of T_3 contains an additional term due to the non-zero value of $E\{\partial U_3(\beta,\gamma)/\partial\beta_2\}$, which happens to be the only non-zero off-diagonal term of the matrix $E\{\partial U(\beta,\gamma)/\partial\beta\}$.

Since the covariance matrix, I is diagonal in this setting, the two-step and CUE objective functions may be written as

$$S(\beta) = \sum_{j=1}^{3} \frac{nE_n\{U_j(\beta, \hat{\gamma})\}^2}{E_n\{U_j^2(\hat{\beta}, \hat{\gamma})\}}$$

$$\tilde{S}(\beta) = \sum_{j=1}^{3} \frac{nE_n\{U_j(\beta, \hat{\gamma}(\beta))\}^2}{E_n\{U_j^2(\beta, \hat{\gamma}(\beta))\}}$$

As in (3.23), these score test objective functions resemble sums of squared t-statistics, making the G-estimation score test analogous to one based on $LR^{(OLS)}$. Theorem 2 may consequently be given the interpretation that the minimization procedure under the null 'minimizes out' independent χ_1^2 terms from this score test objective function, leaving a sum of independent χ_1^2 terms equal in number to the dimensions of the constraint.

3.6 Simulation study

3.6.1 Simulation study for estimation

A simulation study was carried out to examine the bias and variance of NIDE and NDE estimators in finite samples and under model misspecification. G-estimation methods (using bias-reduced nuisance parameter estimation) were compared against the triply robust methods of Tchetgen Tchetgen and Shpitser (2012) (using maximum likelihood methods to fit nuisance parameters). Both the G-estimation methods and triply robust methods (referred to as TTS methods) are available in the `plmed` package. The performance of the proposed score and Wald tests was also compared with classical and TTS derived methods for the no-mediation hypothesis (H_0) and the no-direct effect hypothesis (H_1). Datasets of size n were generated for different $(\beta_1, \beta_2, \beta_3)$ values using several hierarchical data generating processes, the first of which

(Process A) was given by

$$\begin{aligned} Z &\sim \mathcal{N}(0,1) \\ X &\sim \text{Bernoulli}\left(\text{expit}(Z + s_x Z^2)\right) \\ M &\sim \mathcal{N}\left(\beta_1 X + Z + s_m Z^2, 1\right) \\ Y &\sim \mathcal{N}\left(\beta_2 M + \beta_3 X + Z + s_y Z^2, 1\right) \end{aligned}$$

with $s_x, s_m, s_y \in \{0,1\}$ used to indicate model misspecification and where expit is the inverse-logit function. Additional data generating processes (B and C) used the same models for Z, X, Y with the mediator models instead respectively generated by

$$\begin{aligned} M &= \beta_1 X + Z + s_m Z^2 + \epsilon \\ M &\sim \text{Bernoulli}\left(\text{expit}(\beta_1 X + Z + s_m Z^2)\right) \end{aligned}$$

where ϵ follows a Student's t-distribution with 5 degrees of freedom. This Student's t-distribution was chosen as a scenario where an investigator using the TTS methods might fail to correctly model the fat tails of the mediator density. For the G-estimation methods, analysis was conducted under the assumed model

$$\begin{aligned} E(X|Z) &= \text{expit}(\gamma_{x1} Z + \gamma_{x2}) \\ E(M|X,Z) &= \beta_1 X + \gamma_{m1} Z + \gamma_{m2} \\ E(Y|M,X,Z) &= \beta_2 M + \beta_3 X + \gamma_{y1} Z + \gamma_{y2} \end{aligned}$$

whereas, for processes A and B, the TTS methods additionally assumed that the mediator followed a homoscedastic normal distribution (which is true in the case of process A, but not process B). For process C, the TTS methods assumed that

$$M \sim \text{Bernoulli}\left(\text{expit}(\beta_1 X + \gamma_{m1} Z + \gamma_{m2})\right)$$

It follows that for processes A and B, the models assumed by the G-methods are correctly specified when the corresponding misspecification indicator (s_x, s_m, s_y) is equal to zero. For process C, however, we see that (3.1) is satisfied only when $\beta_1 = 0$, therefore we expect to obtain valid estimation of the NIDE only when $\beta_1 = s_x = s_y = 0$. For the NDE, however, (3.3) is correct, thus we expect the G-estimation methods to obtain valid inference for the NDE when $s_y = 0$.

To investigate the bias and variance properties of both estimators, two parameter vectors were simulated, $\beta = (0,0,0)$ and $\beta = (1,1,1)$ with sample sizes $n = 100, 500, 1000$ and under various levels of misspecification. 1000 dataset replicates were generated for each simulation, and for each dataset the NIDE, NDE were estimated by both methods. The variance of the G-estimator was also estimated based on influence function theory, and also by bootstrap with 1000 resampling iterations. Monte Carlo estimates for the expectation and variance of each estimator were obtained across the 1000 dataset replicates.

Plots of the bias of each estimator can be seen for each data generating process in Figs. 3.1, 3.2, and 3.3. These figures show that when the conditions for the G-estimator are satisfied, the bias remains close to zero, even in small samples. For all data generating process, the standard error in the G-estimator is smaller than that of the TTS methods. This is due to the fact that the G-estimation methods exploit the assumed partial linearity to gain efficiency, whereas the TTS methods do not.

Interestingly, for process B, the TTS methods perform poorly when the mediator density is misspecified, whereas the G-estimation methods, which do not assume knowledge of the mediator density, perform similarly to data generating process A. This is likely due to large erroneous inverse density weights in the TTS methods. For data generating process C, the G-estimator performs well for the NDE as expected, however, NIDE estimation is biased when $\beta_1 \neq 0$.

In terms of variance estimation, theoretical results and bootstrap estimation performed similarly with both approximating well the empirical variance of the G-estimator. Full data tables for these simulations can be found in the online supplement to the original paper (Hines et al., 2021b).

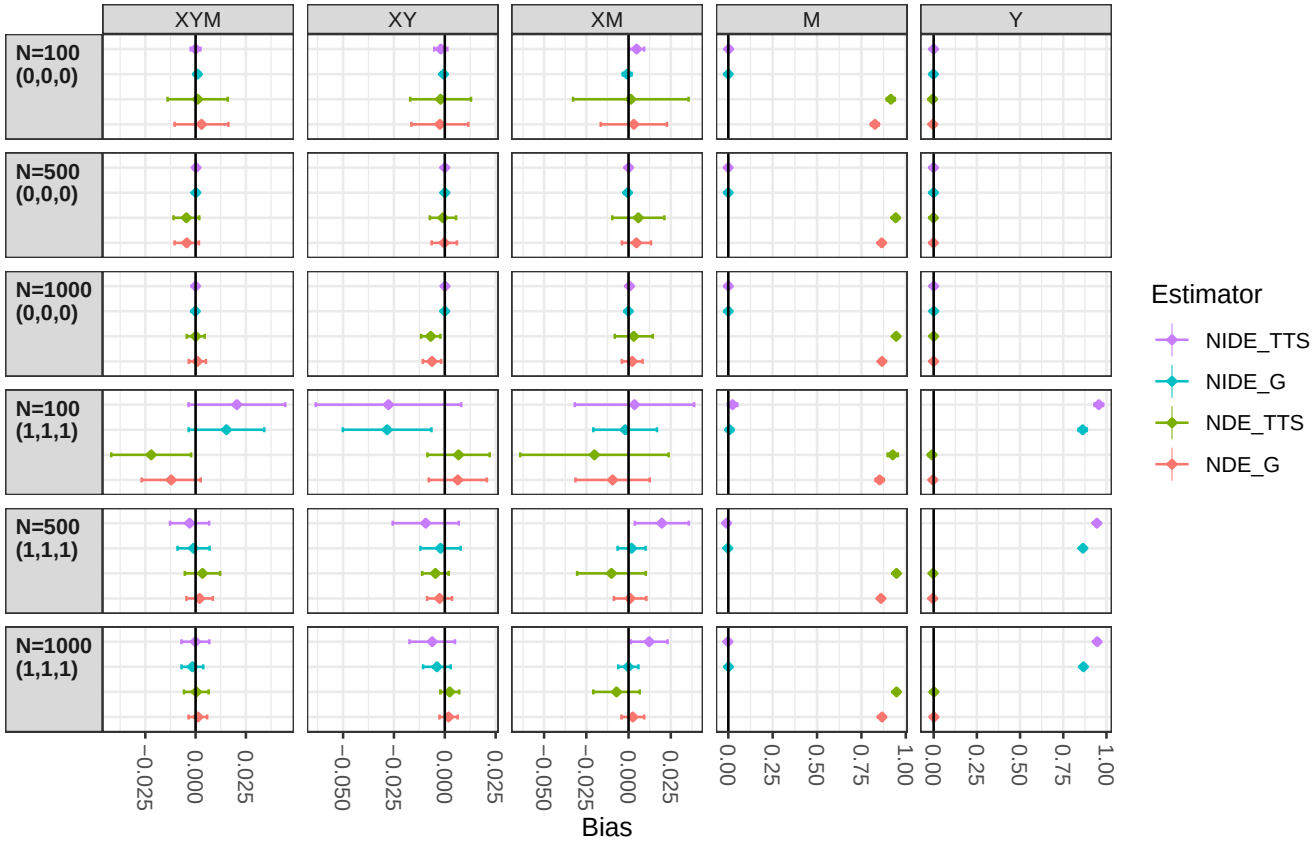


Figure 3.1: Simulated results to investigate the biases of the NDE and NIDE under data generating process A using G-estimation and TTS methods. The estimated bias from 10^3 dataset replicates is plotted on the x-axis with error bars giving a 95% confidence interval of the Monte Carlo estimate. Plots are arranged in a grid where each row represents a different sample size and true target parameter value, and the header of each row lists the correctly specified models (those for which the misspecification indicator is equal to zero). We draw the reader's attention to the different scales on the x-axis of these plots

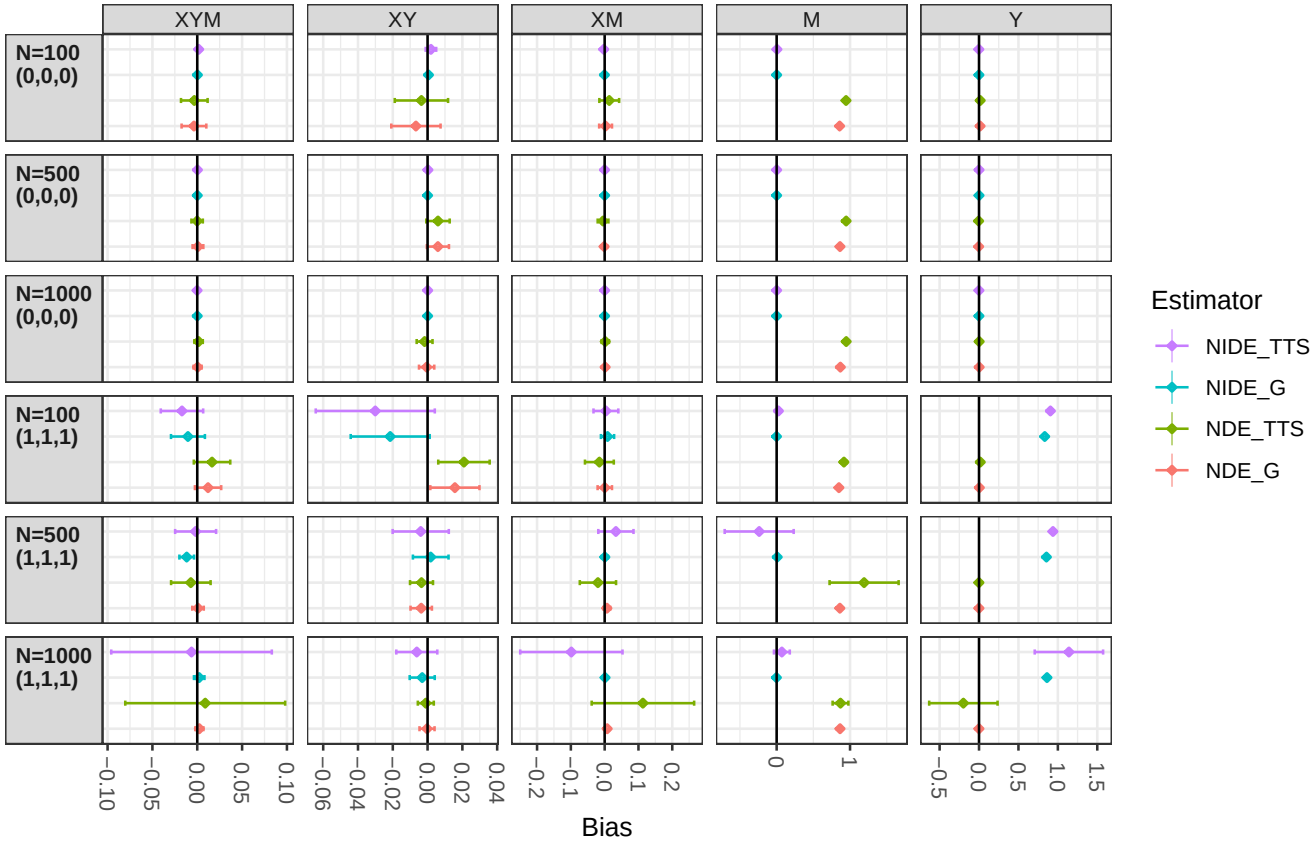


Figure 3.2: Simulated results to investigate the biases of the NDE and NIDE under data generating process B using G-estimation and TTS methods. The estimated bias from 10^3 dataset replicates is plotted on the x-axis with error bars giving a 95% confidence interval of the Monte Carlo estimate. Plots are arranged in a grid where each row represents a different sample size and true target parameter value, and the header of each row lists the correctly specified models (those for which the misspecification indicator is equal to zero). We draw the reader's attention to the different scales on the x-axis of these plots

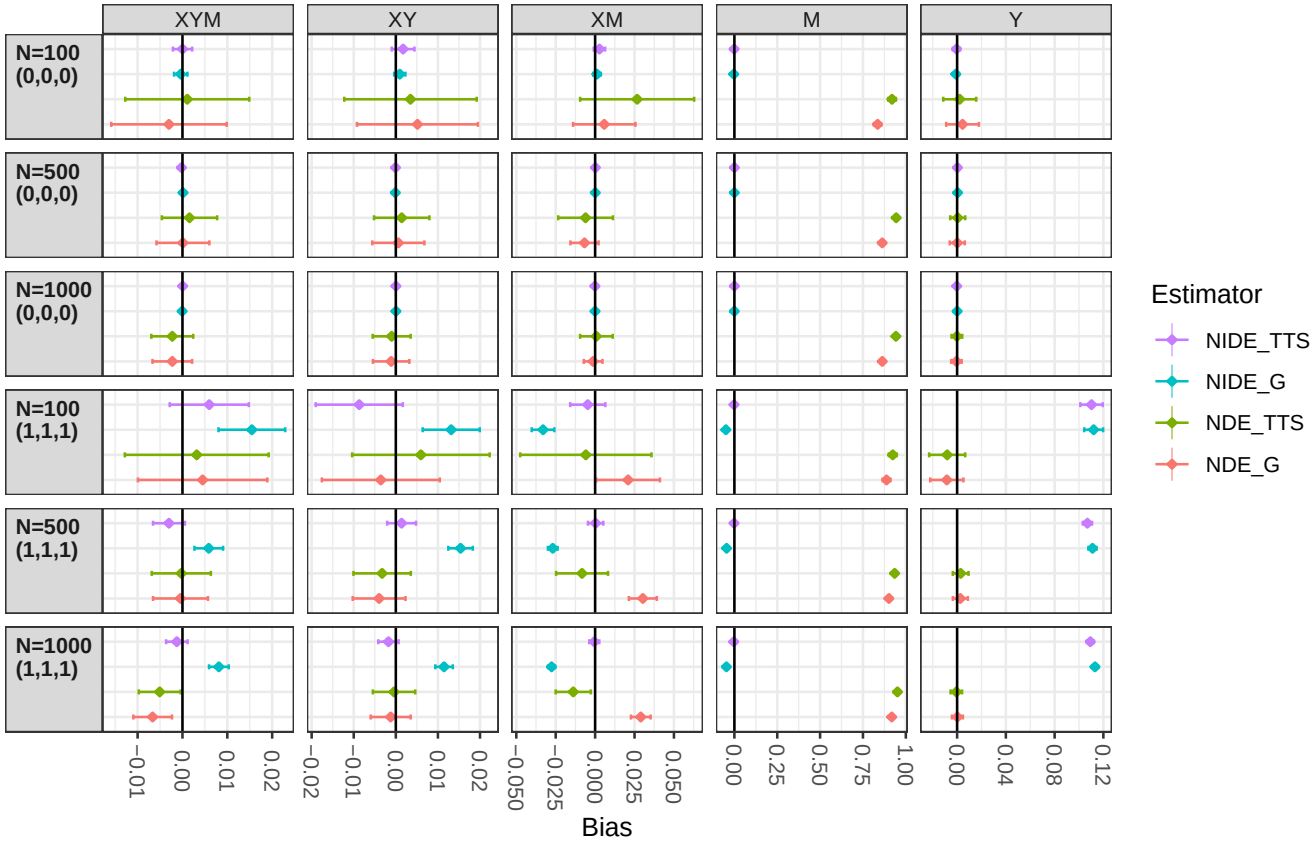


Figure 3.3: Simulated results to investigate the biases of the NDE and NIDE under data generating process C using G-estimation and TTS methods. The estimated bias from 10^3 dataset replicates is plotted on the x-axis with error bars giving a 95% confidence interval of the Monte Carlo estimate. Plots are arranged in a grid where each row represents a different sample size and true target parameter value, and the header of each row lists the correctly specified models (those for which the misspecification indicator is equal to zero). We draw the reader's attention to the different scales on the x-axis of these plots

3.6.2 Simulation study for hypothesis testing

To investigate hypothesis testing methods a greater number of resampled datasets (10^4) was used, since the computationally intensive bootstrap variance estimation procedure did not need to be carried out. The proposed tests based on G-estimation were compared with classical (non-robust) methods and Wald tests based on TTS methods. Replicate datasets were generated for n in the range 50 to 500 with various true values of β under various levels of misspecification. Figures 3.4 and 3.5 respectively show the proportion of datasets for which the tests of H_0 and H_1 were rejected at the 5% level (indicated by a grey line) for data generating process A. When the null is satisfied this rejection proportion corresponds to the Type I error rate, and otherwise corresponds to the statistical power (as in the right most column of both figures). Additional plots for data generating processes B and C can be found in the online supplement to the original paper (Hines et al., 2021b). Here we consider only data generating process A, as it is representative of all three processes.

In Fig.3.4 all testing methods fail to achieve the nominal size when the true parameter takes the value $(\beta_1, \beta_2) = (0, 0)$, as expected by theory. All tests, however, do achieve nominal size when one of β_1 or β_2 differs from zero, given the requisite misspecification conditions. Under correct specification of all working models, the G-estimation score tests display similar power to the classical LR test, dominating both the robust and classical Sobel tests, which also have similar power to each other. This supports the heuristic argument in Section 3.5.3. G-estimation based methods also perform well against those of TTS which, in many cases, seem to converge slowly to the nominal level.

Although these results suggest that the Two-step procedure has greater power over the CUE score test, the Two-step method appears to have an inflated Type I error rate in small samples, which converges more slowly to the nominal level. This may explain the power discrepancy. This behaviour is reflected also in Fig.3.5, where the Robust Wald test suffers from a slightly inflated Type I error rate in small samples. In Fig.3.5 the robust tests perform better than classical test when $E(Y|M, X, Z)$ is misspecified, as in the central row.

3.7 Illustrative example: the COPERS trial

We now illustrate our G-estimation procedure and hypothesis testing methods, by analysing data from the COPERS (COping with persistent Pain, Effectiveness Research in Self-management) trial (Taylor et al., 2016). COPERS was a multi-centre, pragmatic, randomized controlled trial examining the effectiveness of a novel non-pharmacological intervention on the management of chronic musculoskeletal pain. Participants in the intervention arm ($n = 384$) were offered to participate in group therapy sessions, while those in the control arm ($n = 300$) received usual care. The group therapy introduced cognitive behavioural approaches to promote self-efficacy in managing chronic pain. The sessions were delivered over three days within the first week with a follow-up session two weeks later. The control arm participants had no access to the active intervention sessions. Participants and group facilitators were not masked to the study arm they belonged to. The primary outcome, Y, was pain-related disability at 12 months, measured on the Chronic Pain Grade (CPG) disability sub-scale. This is a continuous measure on a scale from 0 to 100, with higher scores indicating worse pain-related disability. The original analysis found no evidence that the COPERS intervention had an effect on improving pain-related disability at 12 months (the average treatment effect on the CPG scale was -1.0, with a 95% CI of -4.8 to 2.7).

The COPERS researchers were interested in investigating whether those in the intervention arm that had attended the majority of sessions benefited more from treatment and whether the effect of therapy was mediated by feelings of self-coping with pain. To this effect, trial participants were also asked to fill out the Pain Self-Efficacy Questionnaire (PSEQ) at 12 weeks (shortly after receiving the intervention), which is intended to measure the participant's confidence to live a normal life despite chronic pain. We will use the score from this questionnaire as a continuous mediator of interest (M).

Attendance at the 24 group therapy sessions was observed to vary between participants in the intervention arm, with the original investigators considering those who had attended at least 12 sessions as receiving treatment ($A = 1, n = 260$), with the remaining patients considered as non-treated ($A = 0, n = 53$). Though planned, a mediation analysis was not performed in the primairy publication (Taylor

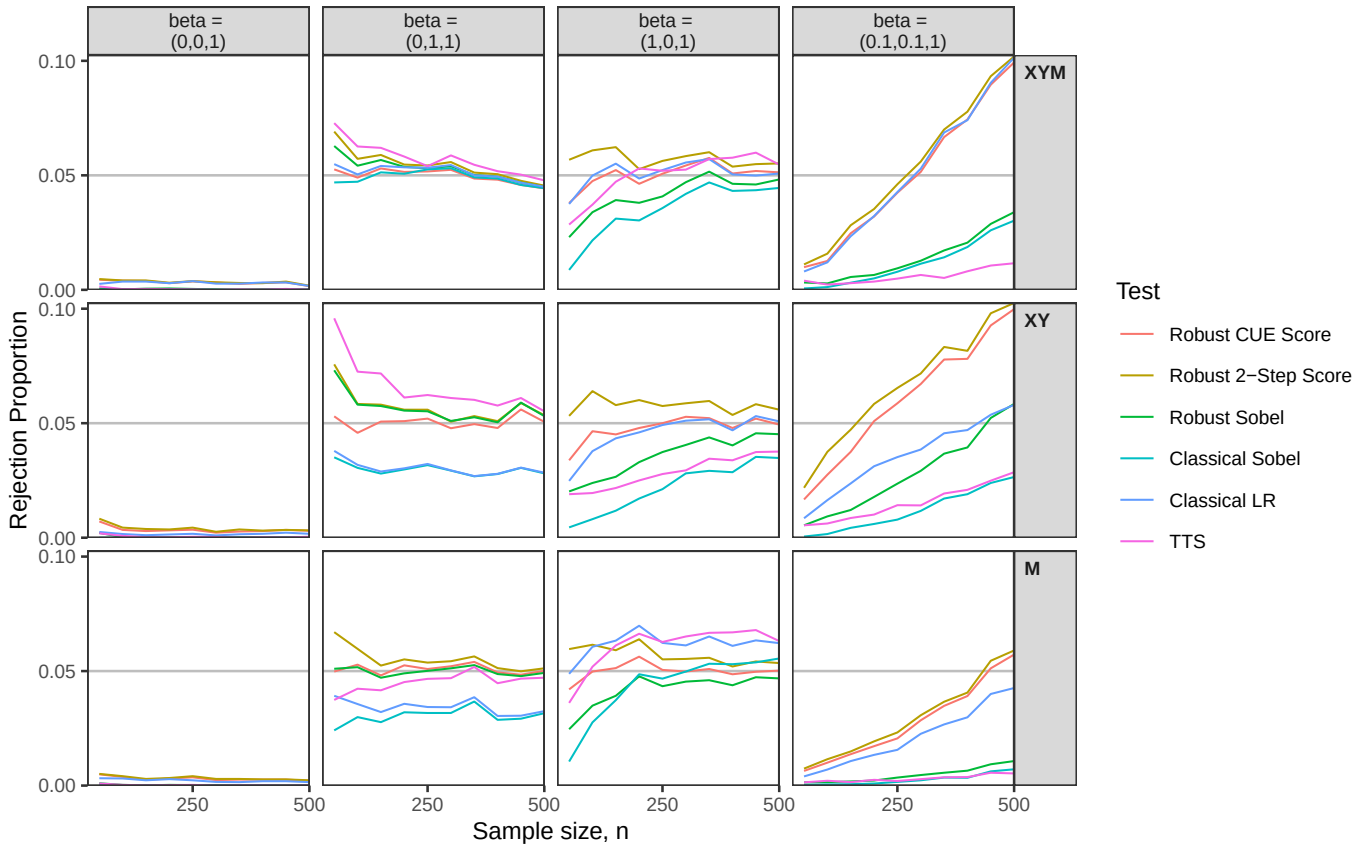


Figure 3.4: Simulated results from data generating process A of the proportion of the 10^4 datasets for which the no-mediation hypothesis (H_0) is rejected at the 5% level testing using the CUE score, Two-step score, Robust Sobel, Classical Sobel, Classical LR, and TTS methods. Each column represents a different true β parameter, whilst each row gives the models which are correctly specified (those for which the misspecification indicator is equal to zero)

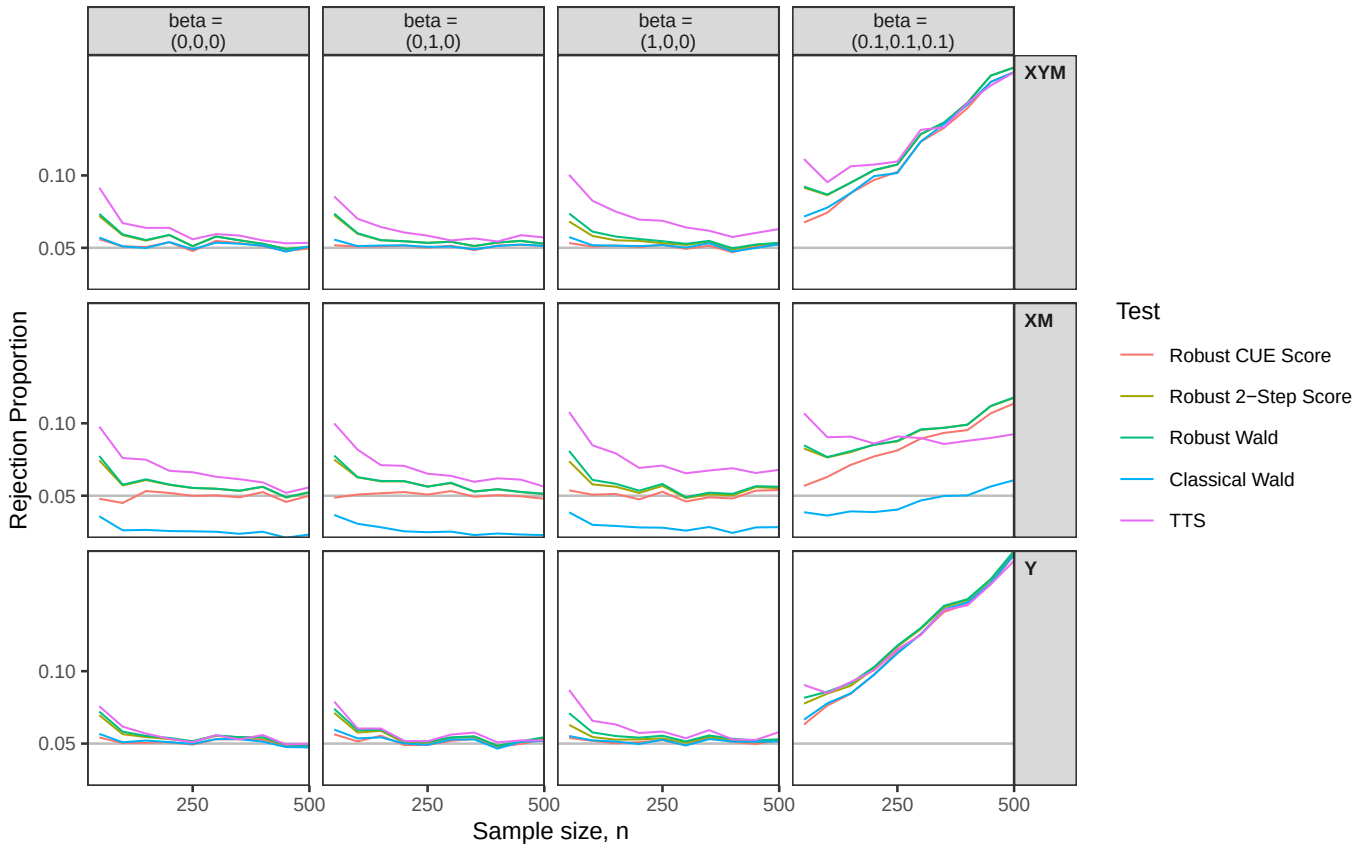


Figure 3.5: Simulated results from data generating process A of the proportion of the 10^4 datasets for which the no-direct effect hypothesis (H_1) is rejected at the 5% level testing using the CUE score, Two-step score, Robust Wald, Classical Wald, and TTS methods. Each column represents a different true β parameter, whilst each row gives the models which are correctly specified (those for which the misspecification indicator is equal to zero)

et al., 2016) due to the lack of overall treatment effect. We will examine only the patients randomized to the treatment arm, and conduct two analyses, one in which the exposure of interest, A is binary (with treatment received if attended at least half the sessions), and another where the number of sessions attended defines the continuous exposure (X).

The baseline covariates included in the original primary analysis are treated, in our analysis, as potential confounders of the three relationships of interest (treatment–mediator, treatment–outcome and mediator–outcome) and thus make up the confounder vector, Z, (which also contains an intercept term). These variables are: site of recruitment, employment status, age, gender, Hospital Anxiety and Depression Scale [HADS], Health Education Impact Questionnaire for social integration subscale, and pain-related disability at baseline. We note that these variables may be insufficient to completely adjust for confounding and it is possible that residual unobserved confounding remains. This is an important caveat for the causal interpretation of the mediated effects, however, we proceed under the assumption that residual confounding is negligible.

Several patients ($n = 10$) were excluded from our analysis as they were missing data on several baseline covariates. Of the remaining patients, some ($n = 51$) were missing data on the mediator or outcome variables. It was therefore decided to analyse complete cases ($n = 323$), weighting each observation by inverse probability weights derived from a logistic regression model for the missingness probability given Z. This method is valid assuming missing-at-random given Z. Reported standard errors do not account for the uncertainty in estimating weights, rendering them conservative (Rotnitzky et al., 2010).

For the binary exposure analysis, a logistic model was assumed for A given Z, whilst linear models for the mediator and outcome were assumed (given (A, Z) and (A, M, Z) respectively). i.e.

$$\begin{aligned} E(A|Z) &= \text{expit}(\gamma_x^{\mathrm{T}} Z) \\ E(M|A, Z) &= \beta_1 A + \gamma_m^{\mathrm{T}} Z \\ E(Y|M, A, Z) &= \beta_2 M + \beta_3 A + \gamma_y^{\mathrm{T}} Z \end{aligned}$$

The continuous exposure was analysed in a similar fashion, however, using a linear model for X given Z, that is

$$\begin{aligned} E(X|Z) &= \gamma_x^{\mathrm{T}} Z \\ E(M|X, Z) &= \beta_1 X + \gamma_m^{\mathrm{T}} Z \\ E(Y|M, X, Z) &= \beta_2 M + \beta_3 X + \gamma_y^{\mathrm{T}} Z \end{aligned}$$

Table 3.1 gives mediated effect estimates from the dichotomized exposure analysis by G-estimation, TTS methods (assuming a normally distributed mediator) and OLS. Table 3.2 gives mediated effect estimates from the continuous exposure analysis by G-estimation and OLS. Table 3.3 shows p-values for the no-mediation and no-indirect effect hypotheses from both analyses, obtained using our Robust Sobel and score tests, along with the classical Sobel and LR methods and (for the dichotomized exposure) the TTS methods.

Table 3.1: Estimated mediation effects for the COPERS trial, treating the exposure as binary and using G-estimation, TTS methods and Ordinary Least Squares

Parameter	G-estimate(95% CI)	TTS(95% CI)	OLS(95% CI)
NDE	5.31(-4.00,14.6)	-2.96(-15.9,9.97)	4.83(-2.98,12.6)
NIDE	-4.60(-7.35,-1.85)	-3.52(-7.47,0.43)	-4.32(-7.05,-1.60)
β_1	6.06(3.42,8.70)	-	5.70(2.93,8.47)
β_2	-0.76(-1.07,-0.45)	-	-0.76(-1.06,-0.45)

This mediation analysis sheds some light on the null treatment effect, with significant evidence of an indirect effect. This evidence suggests that session attendance is associated with an increased perception to cope with disability, which in turn, is associated decreased pain-related disability. Interpreting these results causally should be done with caution, due to the possibility of unobserved confounding. Nevertheless,

Table 3.2: Estimated mediation effects for the COPERS trial, treating the exposure as continuous and using G-estimation and Ordinary Least Squares

Parameter	G-estimate(95% CI)	OLS(95% CI)
NDE	0.17(-0.23,0.57)	0.17(-0.18,0.53)
NIDE	-0.22(-0.35,-0.09)	-0.22(-0.35,-0.09)
β_1	0.29(0.17,0.41)	0.29(0.17,0.42)
β_2	-0.75(-1.06,-0.44)	-0.75(-1.06,-0.45)

Table 3.3: Hypothesis testing results on the COPERS dataset for the null hypotheses H_α (H_0 is the no-mediation hypothesis and H_1 is the no-direct-effect hypothesis) for the analyses where the exposure is treated as binary and continuous

Null α value	Test	P-Value (binary)	P-Value (continuous)
0	Robust Sobel	1.18×10^{-3}	9.23×10^{-4}
0	Robust score (CUE)	1.26×10^{-5}	1.28×10^{-5}
0	Sobel	1.89×10^{-3}	8.50×10^{-4}
0	LR	5.40×10^{-5}	3.43×10^{-6}
0	TTS	8.07×10^{-2}	-
1	Robust Wald	0.264	0.396
1	Robust score (CUE)	0.256	0.397
1	Classical Wald	0.225	0.341
1	TTS	0.654	-

given the possibility of strong mediated effects, researchers interested in cognitive behavioural therapy for chronic pain may want to design add-on interventions that also change self-coping perceptions.

3.8 Extensions

Suppose that an investigator is not confident of the partially linear model in (3.2), but instead would like to conduct analysis under the semi-parametric model linear model with exposure-mediator interaction,

$$E(Y|M, X, Z) = \beta_2 M + \theta XM + g(X, Z) \tag{3.24}$$

where θ is a model parameter, such that when $\theta = 0$ there is no exposure-mediator interaction and the model in (3.2) is recovered. Under this model, and assuming consistency and sequential ignorability, the potential outcome mean $\eta(x, x^*, z)$ may be written as,

$$\eta(x, x^*, z) = (\beta_2 + \theta x)f(x^*, z) + g(x, z)$$

As in Section 3.2, one obtains the following two expressions for the conditional NIDE and NDE when (3.1) holds and when $g(x, z) = \beta_3 x + g(z)$ respectively,

$$\eta(x_0, x_1, z) - \eta(x_0, x_0, z) = \beta_1(\beta_2 + \theta x_0)(x_1 - x_0)$$
$$\eta(x_1, x_1, z) - \eta(x_0, x_1, z) = (\beta_3 + \theta f(x_1, z))(x_1 - x_0)$$

The fact that $f(x, z)$ appears in the expression for the NDE gives some indication as to why robust estimation of mediated effects is generally difficult. The solution proposed by Tchetgen Tchetgen and Shpitser (2014) in this setting would be to correctly specify a model for the NDE, $\eta(1, 1, z) - \eta(0, 1, z)$, implicitly suggesting a correct working model for the conditional expectation of the mediator. This assumption gives the impression of allowing for consistent estimation of the conditional NDE when only the outcome model is correct. The partially linear proposal in the current work is, instead, agnostic to the mediator model, but assumes that $\theta = 0$, obtaining valid estimation when the conditional expectation of the outcome is correctly specified (and partially linear in the sense of (3.3)).

For the NIDE, an estimate may be obtained by estimating $(\beta_1, \beta_2, \theta)$, since (x_1, x_0) are known. One might use G-estimation methods to estimate these three parameters and hence the NIDE itself. This might be achieved by estimation of $(\beta_1, \beta_2, \beta_3, \theta)$ in the intersection model using the set of G-estimation moment conditions

$$\begin{aligned}
U_1(\beta, \gamma) &= \{X - h(Z; \gamma_x)\}\{M - \beta_1 X - f(Z; \gamma_m)\} \\
U_2(\beta, \gamma) &= \{M - \beta_1 X - f(Z; \gamma_m)\}\{Y - \beta_2 M - \theta XM - \beta_3 X - g(Z; \gamma_y)\} \\
U_3(\beta, \gamma) &= \{X - h(Z; \gamma_x)\}\{Y - \beta_2 M - \theta XM - \beta_3 X - g(Z; \gamma_y)\} \\
U_4(\beta, \gamma) &= X\{M - \beta_1 X - f(Z; \gamma_m)\}\{Y - \beta_2 M - \theta XM - \beta_3 X - g(Z; \gamma_y)\}
\end{aligned}$$

Hence, no additional working models are required. Using methods similar to those used to show Lemma 1, one can show that these moment conditions have zero expectation (for some β_3) when (3.1) and (3.24) hold and either $f(z)$ is correctly specified, or $g(x, z) = \beta_3 x + g(z)$ and both $g(z)$ and $h(z)$ are correctly specified. Results concerning estimation and testing could also be extended to account for the fourth moment conditions.

In a similar way, additional estimating equations could also be included to estimate parameters associated with counfounder interactions, such as interactions of the form $Z_j X$ in the mediator or outcome model, or of the form $Z_j M$ in the outcome model, where j indexes the set of counfounders, Z. Alternatively, when the confounder variable, Z_j, is categorical then the partial linearity assumptions, (3.1), (3.2) and (3.3), may be satisfied within certain population subgroups, i.e. the target parameters $(\beta_1, \beta_2, \beta_3)$ differ between subgroups. In this setting, one simple strategy is to estimate mediation effects for each subgroup and take a weighted average of these effects. In practice this could be achieved by passing indicator weights to the `plmed` fitting functions.

Finally, we consider how the proposed G-estimation NIDE estimator (i.e. using moment conditions moment (3.6)–(3.8)) performs when the true data generating distribution follows (3.1) and (3.24). Lemma 1 considers the special case where $\theta = 0$. In general, however, provided $f(z)$ is correctly specified then,

$$\begin{aligned}
\beta_1^* \beta_2^* &= \beta_1(\beta_2 + \theta \bar{x}) \\
\bar{x} &= \frac{E[X \text{var}(M|X, Z)]}{E[\text{var}(M|X, Z)]}
\end{aligned}$$

See Appendix B for details. For continuous exposures $\beta_1^* \beta_2^*$ may thus be interpreted as the NIDE per unit change in X at $x_0 = \bar{x}$. For binary exposures, however, the potential outcome when $X = \bar{x}$ is not well defined. For an analogous interpretation, one might consider a conditional indirect effect defined by

$$\begin{aligned}
\Psi(x) &= \eta(x, 1, z) - \eta(x, 0, z) \\
&= \beta_1(\beta_2 + \theta x)
\end{aligned}$$

where x is some level of the exposure. For binary exposures, the G-estimator returns a weighted average of $\Psi(x)$, which retains the interpretation of an indirect effect

$$\beta_1^* \beta_2^* = \frac{E[\Psi(X) \text{var}(M|X, Z)]}{E[\text{var}(M|X, Z)]}$$

By comparison, in this setting where the outcome model is misspecified, the TTS methods return $\Psi(1)$, provided that the exposure is binary and that $h(z)$ and the conditional density of M given X and Z are both correctly specified.

3.9 Discussion

The main contribution of the current paper is a practical and robust method for carrying out inference of mediated effects in settings where partial linearity of mediator and outcome conditional expectations can be assumed and the vector of variables needed to control for confounding is low dimensional. We

recommend estimation of the NIDE and NDE by G-estimation, in settings where partial linearity may be assumed, but a mediator density function may be difficult to estimate, as required by the methods of Tchetgen Tchetgen and Shpitser (2012). This is for instance the case when analysing continuous mediators, as often encountered in applications in psychology. Compared with OLS, the G-estimators are consistent for the NIDE and NDE, under misspecification of mediator and outcome models and outcome models respectively. The variance of these estimators may be estimated by bootstrap or using asymptotic results with both giving similar results. We also make available the methods in the R package `plmed`, which calculates the NDE and NIDE by G-estimation with variances estimated using asymptotic results.

In terms of hypothesis testing we recommend the robust CUE score test over Two-step robust score methods, due to its faster convergence of the Type I error rate to the nominal size in small samples, and improved power to detect small NIDEs over Wald based testing methods, as demonstrated in simulation studies.

In future work one can hope to extend the proposed G-estimation results to the high-dimentional setting, where the number of parameters indexing nuisance models does not need to be small (compared with the number of observations). In particular, by exploiting the orthogonality of the G-estimator when the exposure and outcome models are both correct, one can show that valid inference of the average treatment effect is obtained even when cross-validated data-adaptive methods (e.g. lasso or machine learning) are used to estimate the nuisance models (such as $f(z), g(z)$, and $h(z)$ in the current paper), with the assumption that such methods will converge to the true model at a sufficiently fast rate and an appropriate sample splitting scheme is applied (Chernozhukov et al., 2017).

Other work by Dukes and Vansteelandt (2019) obtains valid inference of the average treatment effect using G-estimators when nuisance models are fitted using the bias-reduction strategy with a lasso l_1 penalty on the nuisance parameter. This work does not require sample splitting nor does it require that both the exposure and outcome models converge to the truth. Indeed their methods are valid even when the number of confounding variables is allowed to grow with sample size, provided certain sparsity assumptions on the nuisance parameter are satisfied. These methods may be applied directly to the methods in the current paper, with Theorem 2, holding even when the nuisance estimator is orthogonal and penalized (provided that it continues to be orthogonal as for instance in Dukes and Vansteelandt (2019)).

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT

T: +44 (0)20 7299 4646
F: +44 (0)20 7299 4656
www.lshtm.ac.uk

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	1804954	**Title**	Mr.
First Name(s)	Oliver		
Surname/Family Name	Hines		
Thesis Title	Assumption-Lean Inference for Causal and Statistical Questions in the Era of Machine Learning		
Primary Supervisor	Karla Diaz-Ordaz		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?	The American Statistician		
When was the work published?	2022		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	**Yes**	Was the work subject to academic peer review?	**Yes**

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	
Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	This study was conceived by all authors. I carried out the research into existing literature and writing of the manuscript under the supervision of the other authors

SECTION E

Student Signature	Ohines
Date	14 December 2022

Supervisor Signature	Karla Diaz-Ordaz
Date	14 December 2022

Chapter 4

Influence curve based inference

4.1 Introduction

The standard statistical approach of building a model, extracting one or more coefficients and reporting their estimates and associated measures of uncertainty (e.g. confidence intervals) is increasingly being criticised (see e.g. van der Laan (2015)). This standard practice encourages the use of overly simplistic, but misspecified models in order to maintain a simple interpretation of the end result (Breiman, 2001b). It moreover makes the meaning and definition of the reported coefficients dependent upon the selected models. Inference for such 'data-dependent' parameters is not straightforward; ignoring their data-dependent nature, as is commonly done, induces bias, excess variability that is not acknowledged by default standard error estimators and, as a result, overly simplistic inferences.

Building on important results on nonparametric estimation of statistical functionals (Pfanzagl and Wefelmeyer, 1985; Pfanzagl, 1990; Bickel et al., 1993), van der Laan and Rubin (2006), van der Laan and Rose (2011), Robins et al. (2008) and more recently Chernozhukov et al. (2018), showed how the aforementioned concerns can be accommodated by centering a statistical analysis around a predefined nonparametric estimand. This is a model-free functional of the observed data distribution which characterises the quantity one wishes to infer from data (Berk et al., 2021). It follows from the existing literature that root-n estimators with well understood asymptotic behaviour can often be derived (under feasible conditions) by making use of the estimand's so-called efficient influence function or canonical gradient under the nonparametric model. The resulting strategies are known as 'targeted learning' or 'debiased' machine learning, because they effectively enable the use of data-adaptive estimation strategies to model the data-generating distribution, such as variable selection procedures and machine learning algorithms, whilst permitting valid inference of the estimand of interest.

These developments are quite revolutionary in that they are changing the way in which - we believe - data will be analysed in the future. In particular, they shift the focus from model building and validation to choosing estimands that are well connected to scientific questions of interest (Petersen and van der Laan, 2014). This shift enables the analysis to be specified before data is obtained, rather than deciding which statistical quantities to report once a model has been validated, as is usually the case e.g. following model/variable selection. Furthermore, model based analyses usually assume the final model was known a priori, whereas estimand inference based on efficient influence functions tend to be 'honest' in the sense of expressing also the uncertainty around selecting the data-generating model, see e.g. Robins and van der Vaart (2006) for a precise definition of confidence set 'honesty'.

The derivation of the efficient influence function is often regarded as somewhat of a 'dark art'. One reason is that it is not given much attention in textbooks on the topic and neither is it given much focus in statistics education. Textbooks that refer to such derivations often rely on a fluency in concepts from functional analysis (e.g. Hilbert Spaces). A further reason is that the majority of research articles that derive the efficient influence function of a statistical estimand, rely on manipulating a derivative expression into a canonical form, as the integral of a product of an efficient influence function and a score function. These derivations are often complicated, with some steps appearing as if from nowhere to achieve the

desired form.

In this tutorial paper, we instead advocate an equivalent approach based on Gateaux derivatives, formalised by Ichimura and Newey (2022), which is much simpler in our opinion. We will explain this approach and show how to make use of it, while also providing intuitive insight into what an efficient influence function is. We will moreover explain how root-n converging statistical/machine-learning-based estimators can be constructed, using the efficient influence function, and what conditions are needed for these to work well. This tutorial obeys the principles of van der Laan's 'roadmap' (van der Laan and Rose, 2011). It is aimed to be broadly accessible to students and researchers who would like to derive efficient influence functions for all sorts of nonparametric estimands, using simple differentiation methods, such as the chain rule. We use diverse examples first to show the steps in calculating the efficient influence function (Section 4.3.3), and also to convey the very broad applicability of the theory (Section 4.5).

4.2 Step 1: Defining the estimand of interest

The starting point of most statistical analyses is a (semi)parametric model, which is then often interpreted as representing how nature has generated the data. For certain applications, such as in the physical sciences, this model can be the result of a deep theoretical understanding of the data-generating mechanism. However, oftentimes, especially in the spheres of medicine, psychology and economics, the model is chosen for its simplicity and convenience. Many ubiquitous models, such as the generalized linear and Cox proportional hazards models, are commonly used without reference to a mechanistic understanding, rather because the parameters indexing those models provide useful summaries of associations that are of interest to the analysis. This is problematic for various reasons. First, nature is rarely as simple as we would like it to be. This leaves many data analysts torn between reporting a simple model, which is likely misspecified, versus reporting a complex model, which is difficult to interpret (Breiman, 2001b). It demands choosing between an analysis result that is likely biased (as a result of model misspecification) versus one that is likely useless (in view of its complexity). Second, standard statistical theory for (semi)parametric models was developed for settings where the model is a priori justified by some biological, economic, ... theory (so that one can assume it to be correct) and where moreover the data analyst commits to using that model. The truth is that a given model is rarely known to be correct, and that data analysts therefore do not commit to a single model, by adopting model selection strategies. This invalidates standard statistical theory. Third, even the common attempt to infer the model from data (for instance, by relying on variable selection strategies) is overly ambitious as many competing models often fit the data nearly equally well (Breiman, 2001b). While this is generally well realised, it is also then systematically 'forgotten' in how we report and interpret statistical analysis results.

To accommodate these concerns, we will instead aim to infer so-called nonparametric estimands. These are functionals of the true observed data distribution P, which are well defined without reference to a (semi)parametric model, and target the scientific question of interest. With interest in the mean outcome Y, such estimand is unambiguously defined as

$$\Psi_1(P) = E_P(Y),$$

where the subscript P explicates that the expectation E_P is calculated w.r.t. the true distribution P of Y. We will equivalently write this as

$$\Psi_1(P) = P(Y) = \int y dP(y).$$

where $dP(y)$ denotes integration w.r.t. to the probability measure P for the random variable Y. When Y is continuous, $dP(y)$ in this expression can be replaced with $f(y)dy$ to recover the Riemann integral over the probability density function of Y. For many of the examples in this paper we work with Riemann integrals, which are likely to be familiar to most readers.

As a second example, suppose we are interested in the effect of a dichotomous exposure X (coded 0 or 1) on an outcome Y in the presence of data on a possibly high-dimensional vector of covariates Z that is sufficient to adjust for confounding. Then a relevant (statistical) estimand could be defined as

$$\Psi_2(P) = E_P\left\{E_P(Y|X=1,Z) - E_P(Y|X=0,Z)\right\},$$

where, with a slight abuse of notation, the subscript P now explicates that the expectation E_P is calculated w.r.t. the true distribution P of (Z, X, Y). This is known in the causal inference literature as the *average causal effect* or *average treatment effect*.

Alternatively, regardless of whether the exposure is dichotomous or not, its effect on Y can also be expressed using the estimand

$$\Psi_3(P) = \frac{E_P\left[\{X - E_P(X|Z)\}Y\right]}{E_P\left[\{X - E_P(X|Z)\}^2\right]},$$

which equals the expected conditional covariance between X and Y, given Z, divided by the expected conditional variance of X, given Z. Where we are happy to assume a partially linear model for $E_P(Y|X, Z)$, such as

$$E_P(Y|X, Z) = \beta X + \omega(Z),$$

for some function $\omega(.)$, $\Psi_3(P)$ reduces to β, but remains well defined outside this model (Robins et al., 2008; Vansteelandt and Dukes, 2022).

4.3 Step 2: Calculate the estimand's efficient influence function

4.3.1 Preliminaries

Throughout, we will assume that we have access to i.i.d. observed data $O_i \equiv (Z_i, X_i, Y_i)$ for subjects $i = 1, ..., n$. An estimator of the above estimands is then readily obtained by substituting P by an estimator $\hat{P}_n$, where the sub-index n denotes the sample size. For instance, choosing $\hat{P}_n$ to equal the empirical distribution, P_n, of the observations $Y_1, ..., Y_n$ gives rise to the empirical 'plug-in' estimator

$$\Psi_1(\hat{P}_n) = \hat{P}_n(Y) = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

For $\Psi_2(P)$, let $\hat{P}_n$ be any distribution of (Z, X, Y) such that the marginal distribution of Z is given by its empirical distribution, and that the conditional distribution of Y given $X = x$ for $x = 0, 1$ and $Z = Z_i$ for $i = 1, ..., n$ has conditional mean equal to a given estimator $\hat{E}(Y|X = x, Z = Z_i)$, such as the prediction from some machine learning algorithm. Then

$$\Psi_2(\hat{P}_n) = \frac{1}{n}\sum_{i=1}^{n} \hat{E}(Y|X = 1, Z = Z_i) - \hat{E}(Y|X = 0, Z = Z_i).$$

Finally, for $\Psi_3(P)$, let $\hat{P}_n$ be any distribution of (Z, X, Y) such that the conditional distribution of X given $Z = Z_i$ for $i = 1, ..., n$ has conditional mean equal to a given estimator $\hat{E}(X|Z = Z_i)$, and that the marginal distribution of $X - \hat{E}(X|Z)$ and $\left\{X - \hat{E}(X|Z)\right\}Y$ is given by its empirical distribution. Then

$$\Psi_3(\hat{P}_n) = \frac{\sum_{i=1}^{n}\left\{X_i - \hat{E}(X|Z = Z_i)\right\}Y_i}{\sum_{i=1}^{n}\left\{X_i - \hat{E}(X|Z = Z_i)\right\}^2}.$$

The key question now is whether $\Psi(\hat{P}_n)$ is a good proxy for $\Psi(P)$. To understand this, we will scale their difference by $\sqrt{n}$. When this scaled difference converges in distribution (to a non-degenerate law), then we can roughly say that $\Psi(\hat{P}_n)$ differs from $\Psi(P)$ up to a term of the order 1 over root-n. We then say that $\Psi(\hat{P}_n)$ converges to $\Psi(P)$ at parametric rate, or root-n rate, which is usually the best that we can hope to achieve.

For the sample mean Ψ_1, we have that $\hat{P}_n = P_n$ so that this scaled difference equals

$$\begin{aligned}\sqrt{n}\left\{\Psi_1(\hat{P}_n) - \Psi_1(P)\right\} &= \sqrt{n}(\hat{P}_n - P)Y \\ &= \sqrt{n}\hat{P}_n(Y - \Psi_1) = \sqrt{n}P_n(Y - \Psi_1) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Y_i - \mu) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right),\end{aligned}$$

by the classical central limit theorem, where μ and σ^2 are the mean and variance of Y, respectively. We are lucky here that the difference $\sqrt{n}\left\{\Psi_1(\hat{P}_n) - \Psi_1(P)\right\}$ can be written in terms of the operator $\sqrt{n}(P_n - P)$ applied to some Y, but this is not generally the case, for the following reasons. First, the difference $\sqrt{n}\left\{\Psi(\hat{P}_n) - \Psi(P)\right\}$ will in general depend on how much $\hat{P}_n$ differs from P. This was easy in the above example, where $\hat{P}_n$ refers to the empirical distribution P_n of Y, whose behaviour is easy to understand. It is much harder in more general cases where $\hat{P}_n$ may involve data-adaptive estimators, such as predictions $\hat{E}(Y|X = x, Z = Z_i)$ or $\hat{E}(X|Z = Z_i)$ obtained via machine learning or via parametric model building procedures. For such predictions, we may at best have access to some overall, marginal measure of prediction error, but will often have a poor understanding of the bias and imprecision in these predictions at specific covariate levels Z_i. Second, the difference $\sqrt{n}\left\{\Psi(\hat{P}_n) - \Psi(P)\right\}$ will in general also depend on how sensitive the estimand $\Psi(.)$ is to changes in the data-generating distribution. This is also generally poorly understood given that P indexing $\Psi(P)$ is an infinite-dimensional parameter (apart from exceptional cases where the observed data is discrete).

The situation thus looks a bit hopeless at this stage, and indeed, we will not succeed to understand the difference $\sqrt{n}\left\{\Psi(\hat{P}_n) - \Psi(P)\right\}$ for arbitrary estimators $\hat{P}_n$ and arbitrary estimands $\Psi(P)$. However, we will see that progress can be made for specific estimators $\hat{P}_n$, and for estimands $\Psi(P)$ that are sufficiently smooth in the data-generating law P. Before proceeding, we will first formalise the right level of smoothness that is needed.

4.3.2 Parametric submodels

To understand how sensitive $\Psi(.)$ is to changes in the data-generating distribution, we will first take a step back. Rather than examining how $\Psi(.)$ changes as we slightly perturb P towards $\hat{P}_n$, we will study the effect of such perturbation in the direction of a fixed, deterministic distribution, $\widetilde{P}$, which, for the purpose of this discussion, we shall assume is absolutely continuous with respect to P (i.e. the support of $\widetilde{P}$ is contained in the support of P). There are many ways in which we may change $\Psi(.)$ to $\widetilde{P}$. Here, we will focus on perturbations in the direction parameterised via the one-dimensional mixture model

$$P_t = t\widetilde{P} + (1 - t)P, \tag{4.1}$$

indexed by $t \in [0, 1]$, which is called a *parametric submodel*. This is not a parametric model in the usual sense (given that the true data-generating law P is unknown), but is used here as a convenient tool to formalise small perturbations away from P in the direction of $\widetilde{P}$. In particular, note that $P_0 = P$ and $P_1 = \widetilde{P}$.

The sensitivity of $\Psi(.)$ to changes in the data-generating distribution in the direction of $\widetilde{P}$ can now be formalised in terms of the pathwise or directional derivative,

$$\lim_{t \downarrow 0}\left(\frac{\Psi\left(P_t\right) - \Psi(P)}{t}\right) = \frac{d\Psi(P_t)}{dt}\Big|_{t=0},$$

evaluated at $t = 0$, and in the direction of $\widetilde{P}$. When this limit exists (i.e., is finite), it is called a Gâteaux derivative. This generalises the concept of a directional derivative to functional analysis, describing how to take the derivative of a function with respect to a function. Informally, when this derivative 'exists' for

all regular parametric submodels, then we will say that the estimand is pathwise differentiable. Here, a 'regular' parametric submodel is such that its score $\widetilde{P}(O)/P(O) - 1$ has finite variance, a mild restriction that will be needed to ensure that the derivative $dP_t/dt|_{t=0}$ is (or more precisely, inner products with this score are) well-defined. In the next paragraph, we will formalise this definition of pathwise differentiability. This formalisation will be practically useful, as it will provide insight what the so-called efficient influence function (also referred to as *canonical gradient*, or *influence curve*) is, how it can be calculated, and why it is useful.

As in Fisher and Kennedy (2020), we develop some intuition by first considering the special case of discrete data O with support $\{o_1, ..., o_k\}$. Then

$$\frac{d\Psi(P_t)}{dt}\Big|_{t=0} = \sum_{j=1}^{k} \frac{d\Psi(P_t)}{dP_t(o_j)}\Big|_{t=0} \frac{dP_t(o_j)}{dt}\Big|_{t=0}.$$

Here, $d\Psi(P_t)/dP_t(o_j)$ expresses the estimand's sensitivity to small changes in the observed data law. The second component expresses how the observed data law changes along the considered path. It is easily verified to equal

$$\frac{dP_t(o_j)}{dt}\Big|_{t=0} = \widetilde{P}(o_j) - P(o_j).$$

The resulting identity

$$\frac{d\Psi(P_t)}{dt}\Big|_{t=0} = \sum_{j=1}^{k} \frac{d\Psi(P_t)}{dP_t(o_j)}\Big|_{t=0} \left\{\widetilde{P}(o_j) - P(o_j)\right\}, \tag{4.2}$$

is limiting (by being focussed on discrete data) and ignores that the probabilities $P_t(o_1), ..., P_t(o_k)$ are not variation-independent (i.e., they sum to 1 and thus cannot be changed in arbitrary ways) (Fisher and Kennedy, 2020). We therefore appeal to Riesz's representation theorem, according to which this derivative, when it exists, can be obtained via integration of a unique 'representer' $\phi(O, P)$ with finite variance under P, w.r.t. some measure:

$$\frac{d\Psi(P_t)}{dt}\Big|_{t=0} \quad = \quad \int \phi(o, P) \left\{d\widetilde{P}(o) - dP(o)\right\} = (\widetilde{P} - P)\{\phi(O, P)\}. \tag{4.3}$$

Contrasting identity (4.3) with (4.2), we learn that the representer $\phi(O, P)$ is a functional derivative which characterises how sensitive the estimand $\Psi(P)$ is to changes in the data-generating distribution P. It is referred to as the estimand's canonical gradient, efficient influence curve or efficient influence function (under the nonparametric model). The existence of a representer with finite variance such that (4.3) holds, essentially expresses that the estimand is sufficiently smooth as a functional of the data-generating law (so that the notion of a 'derivative' is well-defined); here, the finite-variance condition expresses that the 'derivative' of the estimand w.r.t. the data-generating distribution is finite. Since identity (4.3) is insensitive to constant, additive shifts in $\phi(O, P)$, we will henceforth limit ourselves to mean zero functions (under P) without loss of generality.

We can now more formally define the estimand to be pathwise differentiable when there exists a mean-zero, finite-variance function $\phi(O, P)$ which satisfies (4.3) for all (regular) parametric submodels. Since the efficient influence function has mean zero, $P\{\phi(O, P)\} = 0$, the derivative of $\Psi(P_t)$ w.r.t. t can equivalently be represented as the average of the efficient influence function over the distribution $\widetilde{P}$:

$$\frac{d\Psi(P_t)}{dt}\Big|_{t=0} \quad = \quad \widetilde{P}\left\{\phi(O, P)\right\} = E_{\widetilde{P}}\left\{\phi(O, P)\right\}. \tag{4.4}$$

This result forms the basis of how we will calculate the efficient influence function of an estimand.

4.3.3 How to calculate the efficient influence function of an estimand

There are several ways to derive efficient influence functions. We here advocate the "point mass contamination" strategy that we find simplest. In particular, we will perturb the estimand it in the direction

$\widetilde{P}$ of a point mass at single observation $\tilde{o}$. Identity (4.4) then gives the efficient influence function at observation o directly as

$$\phi(o, P) = \frac{d\Psi(P_t)}{dt}\Big|_{t=0},$$

where the right-hand side is a so-called Gâteaux derivative. This has the same properties as ordinary derivatives, familiar from calculus, such as the chain rule. This will facilitate calculations. The following examples illustrate this.

Throughout, for convenience, we will implicitly assume that we work with continuous variables, but the results continue to hold for discrete variables, or a mix of discrete and continuous variables, upon swapping sums with integrals, indicators with Dirac delta functions, and probability mass functions with probability density functions, where needed. For our purposes, $\mathbb{1}_{\tilde{o}}(o)$ denotes the Dirac delta function w.r.t. $\tilde{o}$; i.e., the density of an idealized point mass at $\tilde{o}$, which equals zero everywhere except at $\tilde{o}$ and which integrates to 1.

Example 1 (population mean). As a first, simple example, consider the mean of Y:

$$\Psi(P) = P(Y) = E_P(Y) = \int y f(y) dy,$$

where $f(y)$ denotes the density function of Y under P (which we assume to be absolutely continuous w.r.t. the Lebesgue measure, though results hold more generally). Perturbing in the direction of a single observation $\tilde{y}$,

$$f_t(y) = t\mathbb{1}_{\tilde{y}}(y) + (1-t)f(y),$$

one obtains,

$$\Psi(P_t) = t\int y\mathbb{1}_{\tilde{y}}(y)dy + (1-t)E_P(Y) = t\tilde{y} + (1-t)\Psi(P).$$

By the chain rule, taking a derivative with respect to t at $t = 0$, gives

$$\frac{d\Psi(P_t)}{dt}\Big|_{t=0} = \tilde{y} - \Psi(P).$$

Because this has finite variance, we conclude that $E(Y)$ is pathwise differentiable with efficient influence function $Y - \Psi(P)$. □

Example 2 (density at a point y). Consider next the density at a given value y, $\Psi(P) = f(y)$. Under the parametric submodel of Example 1, we readily find that

$$\Psi(P_t) = t\mathbb{1}_{\tilde{y}}(y) + (1-t)f(y).$$

By the chain rule, taking a derivative with respect to t at $t = 0$, gives

$$\frac{d\Psi(P_t)}{dt}\Big|_{t=0} = \mathbb{1}_{\tilde{y}}(y) - \Psi(P).$$

Because the Dirac delta function is unbounded when Y is absolutely continuous w.r.t. Lebesgue measure, and therefore has infinite variance, we conclude that $f(y)$ is not pathwise differentiable. This lack of smoothness is the result of insufficient information in the data on the density $f(y)$ at the single point y. It generally implies that no root-n converging estimators can be constructed. □

Example 3 (average density). Consider next the average density of Y:

$$\Psi(P) = P\{f(Y)\} = E_P\{f(Y)\} = \int f^2(y)dy.$$

Under the parametric submodel of Example 1,

$$\Psi(P_t) = \int f_t^2(y)dy.$$

By the chain rule, taking a derivative with respect to t at $t = 0$, gives

$$\begin{aligned}\frac{d\Psi(P_t)}{dt}\Big|_{t=0} &= \int 2f(y)\frac{d}{dt}f_t(y)\Big|_{t=0}dy \\ &= 2\int f(y)\{\mathbb{1}_{\tilde{y}}(y) - f(y)\}dy \\ &= 2\{f(\tilde{y}) - \Psi(P)\}.\end{aligned}$$

Here, we use that the Dirac delta function $\mathbb{1}_y(Y)$ has average

$$P\mathbb{1}_y(Y) = \int \mathbb{1}_y(\tilde{y})f(\tilde{y})d\tilde{y} = f(y),$$

equal to the density at y under the law P. Since $2\{f(Y) - \Psi(P)\}$ has finite variance, we conclude that $E_P\{f(Y)\}$ is pathwise differentiable with efficient influence function $2\{f(Y) - \Psi(P)\}$. □

When perturbing the density $f(o)$ of a vector of observations O in the direction of a point mass at $\tilde{o}$, we have the identity

$$\frac{df_t(o)}{dt}\Big|_{t=0} = \mathbb{1}_{\tilde{o}}(o) - f(o),$$

which was also used in Example 2. This implies a simple formula for the efficient influence function at $\tilde{o}$ of the estimand $E_P\{g(O,P)\}$ for some function $g(O,P)$ of O and the true distribution:

$$\begin{aligned}\frac{d}{dt}E_{P_t}\{g(O,P_t)\}\Big|_{t=0} &= \frac{d}{dt}\left\{\int g(o,P_t)f_t(o)do\right\}\Big|_{t=0} \\ &= \left\{\int \frac{d}{dt}g(o,P_t)f_t(o)do + \int g(o,P_t)\frac{d}{dt}f_t(o)do\right\}\Big|_{t=0} \\ &= E_P\left\{\frac{d}{dt}g(o,P_t)\right\}\Big|_{t=0} + g(\tilde{o},P) - E_P\{g(O,P)\}. \qquad (4.5)\end{aligned}$$

We apply this general identity in the following example.

Example 4 (covariance). The covariance

$$\Psi(P) = E_P[\{Y - E_P(Y)\}\{X - E_P(X)\}],$$

can be written as $E_P\{g(O,P)\}$ for $O \equiv (X,Y)$ and $g(o,P) = \{y - E_P(Y)\}\{x - E_P(X)\}$. Using (4.5), we thus find that

$$\frac{d\Psi(P_t)}{dt}\Big|_{t=0} = E_P\left[\frac{d}{dt}\{Y - E_{P_t}(Y)\}\{X - E_{P_t}(X)\}\Big|_{t=0}\right] + \{\tilde{y} - E_P(Y)\}\{\tilde{x} - E_P(X)\} - \Psi(P).$$

By the chain rule, we further have that

$$\frac{d}{dt}\{Y - E_{P_t}(Y)\}\{X - E_{P_t}(X)\}\Big|_{t=0} = -\frac{d}{dt}E_{P_t}(Y)\Big|_{t=0}\{X - E_P(X)\} - \frac{d}{dt}E_{P_t}(X)\Big|_{t=0}\{Y - E_P(Y)\}.$$

Further applying (4.5) to $E_{P_t}(Y)$ and $E_{P_t}(X)$, we find that

$$\frac{d}{dt}\left\{Y - E_{P_t}(Y)\right\}\left\{X - E_{P_t}(X)\right\}\Big|_{t=0} = -\left\{\tilde{y} - E_P(Y)\right\}\left\{X - E_P(X)\right\} - \left\{\tilde{x} - E_P(X)\right\}\left\{Y - E_P(Y)\right\},$$

which has mean zero. Since $\{Y - E_P(Y)\}\{X - E_P(X)\}$ has finite variance, we conclude that the covariance $\Psi(P)$ is pathwise differentiable with efficient influence function

$$\{Y - E_P(Y)\}\{X - E_P(X)\} - \Psi(P).$$

□

Example 5 (potential outcome mean). Let Y^x denote the potential outcome under exposure level x, which expresses what value the outcome of a given individual would have taken had his/her exposure been set to x by some intervention. Under the usual identifying assumptions, (positivity, consistency, non interference and conditional exchangeability given Z) (Hernán and Robins, 2006),

$$\Psi(P) = E_P\left\{E_P(Y|X=1,Z)\right\}$$

is a statistical estimand of the population mean of Y^1.

Perturbing P in the direction of a point mass at $(\tilde{z},\tilde{x},\tilde{y})$, we find that

$$\begin{aligned}\Psi(P_t) &= \int y f_t(y|1,z) f_t(z) dy dz \\ &= \int y \frac{f_t(y,1,z) f_t(z)}{f_t(1,z)} dy dz,\end{aligned}$$

where $f_t(y|x,z)$ is the conditional density function of Y, given $X=x, Z=z$, under the parametric submodel, and $f_t(y,x,z), f_t(x,z)$, and $f_t(z)$ are the joint density functions of $(Y,X,Z),(X,Z)$, and Z, respectively under the parametric submodel. By the chain rule, we thus have that

$$\begin{aligned}\frac{d\Psi(P_t)}{dt}\Big|_{t=0} &= \int y\left\{\frac{f(z)}{f(1,z)}\frac{d}{dt}f_t(y,1,z)\Big|_{t=0} - \frac{f(y,1,z)f(z)}{f(1,z)^2}\frac{d}{dt}f_t(1,z)\Big|_{t=0} + \frac{f(y,1,z)}{f(1,z)}\frac{d}{dt}f_t(z)\Big|_{t=0}\right\} dy dz \\ &= \int y \frac{f(y,1,z)f(z)}{f(1,z)}\left(\frac{\mathbb{1}_{\tilde{y},\tilde{x},\tilde{z}}(y,1,z)}{f(y,1,z)} - \frac{\mathbb{1}_{\tilde{x},\tilde{z}}(1,z)}{f(1,z)} + \frac{\mathbb{1}_{\tilde{z}}(z)}{f(z)} - 1\right) dy dz.\end{aligned}$$

Evaluating the integral gives the canonical gradient of $\Psi(P)$ at $(\tilde{z},\tilde{x},\tilde{y})$:

$$\frac{d\Psi(P_t)}{dt}\Big|_{t=0} = \frac{\mathbb{1}_{\tilde{x}}(1)}{\pi(\tilde{z},P)}\left\{\tilde{y} - m_1(\tilde{z},P)\right\} + m_1(\tilde{z},P) - \Psi(P),$$

where $m_1(z,P) = E_P(Y|X=1,Z=z)$ and $\pi(z,P) = f(1|z) = E_P(X|Z=z)$ is the propensity score. We conclude that $\Psi(P)$ is pathwise differentiable with the above efficient influence function.

From this, it readily follows that $\Psi_2(P)$ (the average treatment effect) is pathwise differentiable with the efficient influence function given by

$$\varphi_1(O,P) - \varphi_0(O,P) - \Psi_2(P)$$

where $\varphi_x(O,P)$ is the 'uncentered' efficient influence curve

$$\varphi_x(O,P) = \frac{\mathbb{1}_X(x)}{f(x|Z)}\left\{Y - m(x,Z)\right\} + m(x,Z). \tag{4.6}$$

□

Example 6 (conditional outcome mean). Before moving on to more elaborate examples, we finally consider

$$\Psi(P) = E_P(Y|X = x),$$

for a given value x, where X may be (absolutely) continuous (w.r.t. Lebesgue measure). Perturbing P in the direction of a point mass at $(\tilde{x}, \tilde{y})$, we find that

$$\Psi(P_t) = \int y \frac{f_t(y,x)}{f_t(x)} dy,$$

where $f_t(y,x)$ and $f_t(x)$ are the joint density functions of (Y,X) and X, respectively under the parametric submodel. By the chain rule, we thus have that the canonical gradient is

$$\begin{aligned}
\phi(\tilde{o}, P) = \frac{d\Psi(P_t)}{dt}\Big|_{t=0} &= \int y \left\{ \frac{1}{f(x)} \frac{d}{dt} f_t(y,x)\Big|_{t=0} - \frac{f(y,x)}{f(x)^2} \frac{d}{dt} f_t(x)\Big|_{t=0} \right\} dy \\
&= \int \left[\frac{y}{f(x)} \left\{ \mathbb{1}_{\tilde{y},\tilde{x}}(y,x) - f(y,x) \right\} - \frac{y f(y,x)}{f(x)^2} \left\{ \mathbb{1}_{\tilde{x}}(x) - f(x) \right\} \right] dy \\
&= \frac{\mathbb{1}_{\tilde{x}}(x)}{f(x)} \left\{ \tilde{y} - E_P(Y|X = x) \right\}. \qquad (4.7)
\end{aligned}$$

An issue, however, emerges when one considers the variance of the influence function.

$$\begin{aligned}
\text{var}\left\{\phi(O,P)\right\} &= \int \left(\frac{\mathbb{1}_{\tilde{x}}(x)}{f(x)} \right)^2 \left\{ \tilde{y} - E_P(Y|X = x) \right\}^2 f(\tilde{y}|\tilde{x}) f(\tilde{x}) d\tilde{y} d\tilde{x} \\
&= \frac{\mathbb{1}_x(x)}{f(x)} \int \left\{ \tilde{y} - E_P(Y|X = x) \right\}^2 f(\tilde{y}|x) d\tilde{y} \\
&= \frac{\mathbb{1}_x(x)}{f(x)} \text{var}(Y|X = x)
\end{aligned}$$

Since the Dirac delta function $\mathbb{1}_x(x)$ takes an infinitely large value when X is continuous (i.e. when its probability distribution is absolutely continuous w.r.t. Lebesgue measure), we conclude that the conditional mean is not pathwise differentiable in that case.

When X is discrete (as in Example 5), however, then we have that the indicator function $\mathbb{1}_x(x) = 1$, so that the variance of the efficient influence function is finite (so long as $\text{var}(Y|X = x) < \infty$ and $f(x) > 0$). □

The approach that we have adopted in the above examples follows the calculation in Hampel (1974). A second, perhaps more common approach, instead uses the following, canonical form

$$\begin{aligned}
\frac{d\Psi(P_t)}{dt}\Big|_{t=0} &= \int \phi(o,P) \left\{ d\widetilde{P}(o) - dP(o) \right\} \qquad (4.8) \\
&= \int \phi(o,P) \left(\frac{d\widetilde{P}(o)}{dP(o)} - 1 \right) dP(o) \\
&= \int \phi(o,P) S(o) dP(o) \\
&= E_P\left\{ \phi(O,P) S(O) \right\} = P\left\{ \phi(O,P) S(O) \right\}. \qquad (4.9)
\end{aligned}$$

The efficient influence function is then calculated as the unique mean zero function $\phi(O,P)$ whose inner product (i.e., covariance) with the score $S(O)$ under a parametric submodel P_t equals the pathwise derivative $d\Psi(P_t)/dt|_{t=0}$, for all parametric submodels; see Levy (2019) for a tutorial. This can be quite laborious, however, since one must manipulate score functions and integral expressions, and moreover solve a functional equation like (4.9) (Ichimura and Newey, 2022).

The latter approach nonetheless appears more commonly used because it lends itself easier to semiparametric modelling, where the scores $S(O)$ can now be confined to the scores of those parametric submodels that obey the semiparametric model restrictions. A further reason for the greater popularity of this approach may be the apparent limitation of the approach advocated in this tutorial, that certain estimands (e.g., example 3) cannot be evaluated at P_t because of the use of Dirac delta functions. Ichimura and Newey (2022) note that this does not invalidate the approach, as it can be resolved by substituting the Dirac delta function in P_t by a probability measure, indexed by a bandwidth h, that approaches a point mass when the bandwidth converges to 0. This modification justifies the approach that we adopt, but for simplicity it will be left implicit in the remainder of the work.

4.4 Step 3: Construct an estimator based on the estimand's efficient influence function

4.4.1 Plug-in bias and how to remove it

The previous results help us to develop insight into the scaled difference

$$\sqrt{n}\left\{\Psi(\widetilde{P}) - \Psi(P)\right\}. \tag{4.10}$$

In particular, the canonical gradient gave us a way to express the notion of a functional derivative of the estimand w.r.t. directional changes in the data-generating law. This in turn forms the basis of a functional analog to the Taylor expansion, the so-called von Mises expansion, which is essentially derived from the Taylor series expansion of $\Psi(P_t)$ about the point $t = 1$ in the one-dimensional parametric submodel.

$$\Psi(P) \quad = \quad \Psi(\widetilde{P}) + \frac{d\Psi(P_t)}{dt}\Big|_{t=1}(0-1) + R(P,\widetilde{P}),$$

where $R(P,\widetilde{P})$ is a remainder term of the expansion. This expansion contains the pathwise derivative evaluated at $t = 1$, which may be evaluated using an anologue of the Riesz-representation theorem result in (4.4),

$$\frac{d\Psi(P_t)}{dt}\Big|_{t=1} \quad = \quad -P\left\{\phi(O,\widetilde{P})\right\} = -E_P\left\{\phi(O,\widetilde{P})\right\}, \tag{4.11}$$

details of which are given in Appendix C. It follows that the scaled difference of interest, equation (4.10), can be written as

$$\sqrt{n}\left\{\Psi(\widetilde{P}) - \Psi(P)\right\} \quad = \quad -\sqrt{n}P\left\{\phi(O,\widetilde{P})\right\} - \sqrt{n}R(P,\widetilde{P}) \tag{4.12}$$

where we note that this identity is guaranteed to hold, so long as we impose no restrictions on the remainder term, which we will consider later. Now letting $\widetilde{P}$ equal $\hat{P}_n$, we thus see that

$$\begin{aligned}\sqrt{n}\left\{\Psi(\hat{P}_n) - \Psi(P)\right\} &= -\sqrt{n}P\left\{\phi(O,\hat{P}_n)\right\} - \sqrt{n}R(P,\hat{P}_n)\\ &\approx -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi(O_i,\hat{P}_n) - \sqrt{n}R(P,\hat{P}_n),\end{aligned}$$

Here, the term

$$-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi(O_i,\hat{P}_n) \tag{4.13}$$

does not converge to zero (indeed, it would not even converge to zero if P were used in lieu of $\hat{P}_n$) and may sometimes even diverge. This tends not to cause asymptotic bias in $\Psi(\hat{P}_n)$ (because the calculation of bias requires further scaling by $1/\sqrt{n}$ and, moreover, $\phi(O,P)$ has mean zero and P_n is assumed to converge to

P). However, it biases the scaled difference $\sqrt{n}\left\{\Psi(\hat{P}_n) - \Psi(P)\right\}$, thereby invalidating naïve confidence intervals and tests. Understanding the behaviour of (4.13) is difficult as a result of the non-standard behaviour of statistical/machine learning-based estimators affecting the large sample behaviour of $\hat{P}_n$, which in turn propagates into the behaviour of $\Psi(\hat{P}_n)$. Let us reconsider Example 5 (the potential outcome mean), for instance, where

$$\Psi(P) = E_P\left\{E_P(Y|X=1,Z)\right\}.$$

A plug-in estimator is readily obtained as

$$\Psi(\hat{P}_n) = \frac{1}{n}\sum_{i=1}^{n} m_1(Z_i, \hat{P}_n).$$

Here, $m_1(z, \hat{P}_n)$ denotes a data-adaptive estimator of $m_1(z,P) = E_P(Y|X=1,Z=z)$, e.g. obtained using parametric regression models with variable selection, or via machine learning algorithms. The plug-in bias term (4.13) then equals[1]

$$-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\mathbb{1}_1(X_i)}{\pi(Z_i,P)}\left\{Y_i - m_1(Z_i,\hat{P}_n)\right\} + m_1(Z_i,\hat{P}_n) - \Psi(\hat{P}_n) = -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\mathbb{1}_1(X_i)}{\pi(Z_i,P)}\left\{Y_i - m_1(Z_i,\hat{P}_n)\right\}$$
$$= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\mathbb{1}_1(X_i)}{\pi(Z_i,P)}\left\{Y_i - m_1(Z_i,P)\right\} + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\mathbb{1}_1(X_i)}{\pi(Z_i,P)}\left\{m_1(Z_i,P) - m_1(Z_i,\hat{P}_n)\right\},$$

where the second term will often follow a non-standard distribution. For instance, when $m_1(z,\hat{P}_n)$ is obtained using parametric regression models with variable selection, it will often follow a mixture distribution for each z as a result of variation in the selected model across repeated samples.

The extent to which the plug-in bias term (4.13) causes bias is thus generally poorly understood as it inherits the behaviour of $\hat{P}_n$, which is complex when data-adaptive methods are used. Rather than attempting to understand its asymptotic behaviour, a much simpler remedy is therefore to adjust the plug-in estimator in such a way that the this bias is zero. One easy way to do this is by defining a new estimator

$$\Psi(\hat{P}_n) + \frac{1}{n}\sum_{i=1}^{n}\phi(O_i,\hat{P}_n)$$

obtained by subtracting an estimate of the plug-in bias from the plug-in estimator. Then, the scaled difference between this so-called one-step estimator and $\Psi(P)$ is governed by $-\sqrt{n}R(P,\hat{P}_n)$, which will generally be much smaller.

We will see later that there are other ways of modifying the plug-in estimator so that the resulting estimator has zero plug-in bias.

4.4.2 The von Mises expansion

In the previous section, we have built some intuition into plug-in bias and how it can be removed. In order to understand the behaviour of the scaled difference between the one-step estimator and $\Psi(P)$, a more careful derivation is needed. In particular, because P is unknown, we substituted it by the empirical distribution function P_n of the observed data, but did not express the error this is adding to the results. Let us therefore take a step back to identity (4.12). By adding and subtracting $\sqrt{n}(P_n - P)\left\{\phi(O,P)\right\}$ and $\sqrt{n}P_n\left\{\phi(O,\widetilde{P})\right\}$ to the righthand side, we obtain

$$\begin{aligned}\sqrt{n}\left\{\Psi(\widetilde{P}) - \Psi(P)\right\} &= \sqrt{n}(P_n - P)\left\{\phi(O,P)\right\} - \sqrt{n}P_n\left\{\phi(O,\widetilde{P})\right\} \\ &\quad + \sqrt{n}(P_n - P)\left\{\phi(O,\widetilde{P}) - \phi(O,P)\right\} - \sqrt{n}R(P,\widetilde{P}).\end{aligned}$$

[1]Note that we have evaluated the plug-in bias term at the true propensity score because the considered plug-in estimator does not rely on an estimated propensity score. One may alternatively evaluate the plug-in bias term at an estimated propensity score, which will then only affect the remainder term.

Setting $\widetilde{P}$ to $\hat{P}_n$ one can rewrite the plug-in bias in form of the so-called von Mises expansion:

$$\begin{aligned}\sqrt{n}\left\{\Psi(\hat{P}_n)-\Psi(P)\right\} &= -\sqrt{n}P\left\{\phi(O,\hat{P}_n)\right\}-\sqrt{n}R(P,\hat{P}_n)\\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi(O_i,P)-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi(O_i,\hat{P}_n)\\ &\quad +\sqrt{n}(P_n-P)\left\{\phi(O,\hat{P}_n)-\phi(O,P)\right\}-\sqrt{n}R(P,\hat{P}_n). \qquad (4.14)\end{aligned}$$

Here, the first term converges to a normal, mean zero variate by the central limit theorem and the unbiasedness of the canonical gradient. The empirical process term (i.e., the third term in (4.14)) and the remainder term $\sqrt{n}R(P,\hat{P}_n)$ can often be shown to converge to zero under conditions that we will come back to.

Since the asymptotic behaviour of $\hat{P}_n$, and therefore also of the second term, is often poorly understood, popular approaches are designed to remove this drift term from the expansion. This can be done in multiple possible ways.

One-step estimator. The first is to rewrite the above expansion as

$$\begin{aligned}\sqrt{n}\left\{\Psi(\hat{P}_n)+\frac{1}{n}\sum_{i=1}^{n}\phi(O_i,\hat{P}_n)-\Psi(P)\right\} &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi(O_i,P)\\ &\quad +\sqrt{n}(P_n-P)\left\{\phi(O,\hat{P}_n)-\phi(O,P)\right\}-\sqrt{n}R(P,\hat{P}_n),\end{aligned}$$

and thus to calculate the estimator of $\Psi(P)$ as the *one-step estimator*

$$\Psi(\hat{P}_n)+\frac{1}{n}\sum_{i=1}^{n}\phi(O_i,\hat{P}_n).$$

In Example 3, this delivers

$$\int f^2(y,\hat{P}_n)dy+\frac{2}{n}\sum_{i=1}^{n}\left\{f(y_i,\hat{P}_n)-\int f^2(y,\hat{P}_n)dy\right\}=\left\{\frac{2}{n}\sum_{i=1}^{n}f(y_i,\hat{P}_n)\right\}-\int f^2(y,\hat{P}_n)dy.$$

where $f(y,\hat{P}_n)$ is a density estimator. For Example 5 we consider two different cases. When the propensity score $\pi(Z_i,P)$ is known, for instance in randomized experiments, then one obtains the estimator,

$$\Psi(\hat{P}_n)+\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{1}_1(X_i)}{\pi(Z_i,P)}\left\{Y_i-m_1(Z_i,\hat{P}_n)\right\}+m_1(Z_i,\hat{P}_n)-\Psi(\hat{P}_n),$$

When the propensity score is unknown, as is the case for observational data, it must also be estimated (e.g. using a data-adaptive estimator $\pi(Z_i,\hat{P}_n)$) and the one-step estimator recovers the augmented IPW estimator,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{1}_1(X_i)}{\pi(Z_i,\hat{P}_n)}\left\{Y_i-m_1(Z_i,\hat{P}_n)\right\}+m_1(Z_i,\hat{P}_n),$$

This propensity score estimation has consequences for the remainder term; see below.

Estimating equation estimators. The second is to force the drift term to be zero by using it as an estimating equation; that is, to calculate an estimator for $\Psi(P)$ as the solution to an estimating equation given by this drift term:

$$0=\frac{1}{n}\sum_{i=1}^{n}\phi(O_i,\hat{P}_n). \qquad (4.15)$$

This is easy in the above examples, where the efficient influence function is linear in $\Psi(P)$. In Example 3, solving the identity

$$0 = \frac{2}{n}\sum_{i=1}^{n}\left\{f(y_i, \hat{P}_n) - \Psi(\hat{P}_n)\right\}$$

delivers a different estimator than the one-step estimator, namely

$$\Psi(\hat{P}_n) = \frac{1}{n}\sum_{i=1}^{n} f(y_i, \hat{P}_n),$$

with the advantage that it is guaranteed to be non-negative. In Example 5, solving the identity

$$0 = \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{1}_1(X_i)}{\pi(Z_i, \hat{P}_n)}\left\{Y_i - m_1(Z_i, \hat{P}_n)\right\} + m_1(Z_i, \hat{P}_n) - \Psi(\hat{P}_n)$$

for $\Psi(\hat{P}_n)$ delivers the same estimator as the one-step estimator.

Targeted learning. The third works instead by tuning the initial estimator $\hat{P}_n$ such that it forces (4.15) to hold, which is the focus of targeted learning approaches (van der Laan and Rubin, 2006; van der Laan and Rose, 2011). For instance, tuning the estimator $\hat{P}_n$ in Example 5 to a retargeted estimator $\hat{P}_n^*$ that satisfies

$$0 = \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{1}_1(X_i)}{\pi(Z_i, \hat{P}_n^*)}\left\{Y_i - m_1(Z_i, \hat{P}_n^*)\right\},$$

ensures that the one-step estimator reduces to the simple plug-in estimator

$$\frac{1}{n}\sum_{i=1}^{n} m_1(Z_i, \hat{P}_n^*),$$

which then has standard asymptotic behaviour. This tuning can be achieved in many ways; for instance, one may leave the propensity score model unchanged by defining $\pi(Z_i, \hat{P}_n^*) = \pi(Z_i, \hat{P}_n)$ and tune the outcome model by defining,

$$m_1(Z_i, \hat{P}_n^*) = m_1(Z_i, \hat{P}_n) + \hat{\epsilon}\frac{1}{\pi(Z_i, \hat{P}_n)},$$

where $\hat{\epsilon}$ is chosen to set the plug-in bias to zero, i.e., it is the solution to

$$0 = \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{1}_1(X_i)}{\pi(Z_i, \hat{P}_n^*)}\left\{Y_i - m_1(Z_i, \hat{P}_n) - \hat{\epsilon}\frac{1}{\pi(Z_i, \hat{P}_n)}\right\}.$$

Retargeting an initial density estimator in Example 3 is less straightforward because of the difficulty of ensuring that the retargeted density continues to be a proper density.

Under sufficient conditions that ensure the empirical process and remainder terms to converge to zero, it follows from the above expansion that all 3 above approaches deliver an estimator $\Psi(\hat{P}_n^*)$ whose asymptotic distribution obeys

$$\sqrt{n}\left\{\Psi(\hat{P}_n^*) - \Psi(P)\right\} \xrightarrow{d} \mathcal{N}\left(0, P\left\{\phi(Y, P)^2\right\}\right). \tag{4.16}$$

This is a powerful result, which means that the asymptotic efficiency bound for a nonparametric estimand can be derived as the expected square of the efficient influence function. Heuristically, this bound is a nonparametric analogue of the Cramer-Rao lower bound, and estimators of the type in (4.16) are said to be asymptotically efficient, in the sense that they are asymptotically equivalent to the estimator obtained by solving an estimating equation with known rather than estimated influence function:

$$0 = \frac{1}{n}\sum_{i=1}^{n}\phi(O_i, P).$$

It thus tells us that the influence function behaves like the score function in parametric estimation (Wasserman, 2006). It motivates why the variance of $\Psi(\hat{P}_n)$ can be estimated as 1 over n times the sample variance of the efficient influence function (evaluated at $\hat{P}_n$), without needing to account for the uncertainty in $\hat{P}_n$. Identity (4.16) also motivates why the definition of pathwise differentiability includes the requirement of an efficient influence function with finite variance. Pathwise differentiability of an estimand is therefore tantamount to the existence of (regular) root-n consistent estimators of that estimand.

4.4.3 Controlling the empirical process term

The asymptotic behaviour of the empirical process term

$$\sqrt{n}(P_n - P)\left\{\phi(O, \hat{P}_n) - \phi(O, P)\right\}$$

is generally difficult to understand when data-adaptive statistical methods are used. However, it becomes much simpler to understand when the estimator $\hat{P}_n$ is trained on an independent dataset, as one can then reason conditional on that estimator. Reasoning as such, a direct application of Chebyshev's inequality shows that the empirical process term converges to zero in probability when the conditional variance of $\phi(O, \hat{P}_n) - \phi(O, P)$, i.e.,

$$P\left[\left\{\phi(O, \hat{P}_n) - \phi(O, P)\right\}^2\right]$$

given $\hat{P}_n$, converges to zero in probability. The latter can often be shown to hold when the estimator $\hat{P}_n$ converges to P in probability (or even weaker conditions that certain functionals of $\hat{P}_n$ converge to the corresponding functionals of P in probability) and certain positivity conditions hold (see for instance Chernozhukov et al. (2017); Vansteelandt and Dukes (2022) for detailed examples). The use of an independent sample in this way is important for shrinking the empirical process term, but contrary to what popular wisdom sometimes seems to suggest, does not eliminate the leading plug-in bias terms on which we have focused.

Because one rarely has independent data available to train $\hat{P}_n$, Zheng and van der Laan (2011) and Chernozhukov et al. (2018) recommend a cross-fitting procedure, whereby the data is split into K folds. For each individual i from fold $k = 1, ..., K$, the efficient influence function for that individual is then evaluated in an estimator $\hat{P}_n$ trained on the data for all individuals, except those in the k-th fold. This usually results in a better asymptotic approximation, as reflected by more accurate standard error estimators obtained as 1 over root-n times the sample standard deviation of those influence functions. However, it may induce some finite-sample bias in the estimator as a result of the data-adaptive estimator $\hat{P}_n$ being trained on a smaller sample of data.

4.4.4 Controlling the remainder term

To understand the remainder term $\sqrt{n}R(P, \hat{P}_n)$, we return to the von Mises expansion (4.14), from which it is seen to equal

$$\sqrt{n}R(P, \hat{P}_n) \quad = \quad -\sqrt{n}P\left\{\phi(O, \hat{P}_n)\right\} - \sqrt{n}\left\{\Psi(\hat{P}_n) - \Psi(P)\right\}.$$

In Example 5, this is

$$\begin{aligned}
\sqrt{n}R(P, \hat{P}_n) &= -\sqrt{n}E_P\left[\frac{\mathbb{1}_1(X)}{\pi(Z, P)}\left\{Y - m_1(Z, \hat{P}_n)\right\} + m_1(Z, \hat{P}_n) - \Psi(\hat{P}_n)\right] - \sqrt{n}\left\{\Psi(\hat{P}_n) - \Psi(P)\right\} \\
&= -\sqrt{n}E_P\left[\frac{\mathbb{1}_1(X)}{\pi(Z, P)}\left\{Y - m_1(Z, \hat{P}_n)\right\} + m_1(Z, \hat{P}_n) - \Psi(P)\right] \\
&= -\sqrt{n}E_P\left[\left\{\frac{\pi(Z, P)}{\pi(Z, P)} - 1\right\}\left\{m_1(Z, P) - m_1(Z, \hat{P}_n)\right\}\right] = 0.
\end{aligned}$$

Hence when the propensity score $\pi(Z, P)$ is known, the remainder term is zero. Other estimands are also known to have a zero remainder, such as the average density (see Example 3).

When substituting $\pi(Z, \hat{P}_n)$ for $\pi(Z, P)$, by the Cauchy-Schwarz inequality, the remainder can be upper bounded by

$$\sqrt{n}E_P\left[\left\{\frac{\pi(Z,P)}{\pi(Z,\hat{P}_n)}-1\right\}^2\right]^{1/2} E_P\left[\left\{m_1(Z,P)-m_1(Z,\hat{P}_n)\right\}^2\right]^{1/2}.$$

This converges to zero in probability when $\pi(Z, \hat{P}_n)$ and $m_1(Z, \hat{P}_n)$ converge to $\pi(Z, P)$ and $m_1(Z, P)$, respectively, at faster than n to the quarter rate (and $\pi(Z, \hat{P}_n)$ is bounded away from zero), which is a typical requirement in the non/semiparametric literature. In this specific example, the remainder also shrinks to zero under more general conditions; $m_1(Z, \hat{P}_n)$ can be allowed to converge at a slow rate, so long as $\pi(Z, \hat{P}_n)$ is fast converging. This additional flexibility is sometimes known as 'rate double-robustness', and does not apply to remainder terms in general, although it does apply for many common estimands in causal inference/missing data problems (Rotnitzky et al., 2021). To obtain fast rates of convergence with flexible methods, we typically rely on strong smoothness/sparsity assumptions (e.g. when Z is high dimensional, $\pi(Z, P)$ and/or $m_1(Z, P)$ should depend on a small number of the covariates), in addition to well-chosen tuning parameters for the learners.

We refer the reader to Fisher and Kennedy (2020) for a rigorous treatment of the remainder terms of the von Mises expansion, which are usually analysed on a case-by-case basis (see for instance Chernozhukov et al. (2017); Vansteelandt and Dukes (2022) for detailed examples).

4.5 Examples

In this section, we illustrate the calculation of the canonical gradient for the expected conditional covariance and the average derivative effect, deriving the one-step estimators in both cases. Further examples are provided in Appendix C, which also contains results that readers may find helpful for reference.

4.5.1 General results

For notational convenience we define an operator, ∂_t, applied to an arbitrary function, $g(t)$, as

$$\partial_t g(t) = \frac{dg(t)}{dt}\Big|_{t=0}.$$

For instance, let $f_t(y, x)$ denote a parametric submodel which disturbs the density $f(y, x)$ of (Y, X) at (y, x) in the direction of a point mass at $(\tilde{y}, \tilde{x})$. Then from

$$f_t(y|x) = \frac{f_t(y,x)}{f_t(x)}$$

and using the chain rule and the quotient rule for derivatives, we obtain

$$\begin{aligned}
\partial_t f_t(y|x) &= \partial_t\left\{\frac{f_t(y,x)}{f_t(x)}\right\} \\
&= \frac{\partial_t f_t(y,x) f(x) - f(y,x)\partial_t f_t(x)}{f^2(x)} \\
&= \frac{1}{f(x)}\left[\mathbb{1}_{\tilde{y},\tilde{x}}(y,x) - f(y,x) - \frac{f(y,x)}{f(x)}\left\{\mathbb{1}_{\tilde{x}}(x) - f(x)\right\}\right] \\
&= \frac{\mathbb{1}_{\tilde{x}}(x)}{f(x)}\left\{\mathbb{1}_{\tilde{y}}(y) - f(y|x)\right\}.
\end{aligned}$$

Similarly to (4.5), this expression may be used to derive the following identity for the conditional expectation of an arbitrary function $g(o, P)$, where $o = (y, x')'$:

$$\begin{aligned}\partial_t E_{P_t}\left\{g(O,P_t)|X=x\right\} &= \partial_t \int g(o,P_t) f_t(y|x) dy \\ &= \frac{\mathbb{1}_{\tilde{x}}(x)}{f(x)}\left[g(\tilde{o},P) - E_P\left\{g(O,P)|X=x\right\}\right] + E_P\left\{\partial_t g(O,P_t)|X=x\right\}. \qquad (4.17)\end{aligned}$$

Such generic expressions are helpful to relate to, as they can be used to speed up derivations. For instance, for the potential outcome mean, defining $m_1(Z, P) = E_P(Y|X = 1, Z)$, it readily follows from (4.5) that

$$\partial_t E_{P_t}\left\{m_1(Z,P_t)\right\} = m_1(\tilde{z},P) - E_P\left\{m_1(Z,P)\right\} + E_P\left\{\partial_t m_1(Z,P_t)\right\},$$

and by (4.17), that

$$\partial_t m_1(Z,P_t) \;= \frac{\mathbb{1}_{\tilde{x},\tilde{z}}(1,z)}{f(1,z)}\left\{\tilde{y} - m_1(z,P)\right\} + 0.$$

Averaging over the distribution of Z then delivers

$$E_P\left\{\partial_t m_1(Z,P_t)\right\} = \frac{\mathbb{1}_{\tilde{x}}(1)}{f(1|z)}\left\{\tilde{y} - m_1(z,P)\right\}.$$

Hence, we recover the same result as before.

Example 7 (expected conditional covariance). Consider the expected conditional covariance,

$$\Psi(P) = E_P\left[\left\{Y - E_P(Y|Z)\right\}\left\{X - E_P(X|Z)\right\}\right]$$

which appears in hypothesis testing (Shah and Peters, 2018) and in parameter estimation in generalized linear models (Vansteelandt and Dukes, 2022). Define

$$\mathrm{cov}_t(Y,X|Z) = E_{P_t}\left[\left\{Y - E_{P_t}(Y|Z)\right\}\left\{X - E_{P_t}(X|Z)\right\}|Z\right].$$

Upon noting that $\Psi(P) = E_P\left\{\mathrm{cov}(Y, X|Z)\right\}$ is of the form in (4.5), we find that

$$\partial_t \Psi(P_t) = \mathrm{cov}(Y,X|\tilde{z}) - \Psi(P) + E_P\left\{\partial_t \mathrm{cov}_t(Y,X|Z)\right\}.$$

The complication is clearly in the final term, which is of the form in (4.17), hence,

$$\begin{aligned}\partial_t \mathrm{cov}_t(Y,X|z) = \frac{\mathbb{1}_{\tilde{z}}(z)}{f(z)}&\left[\left\{\tilde{y} - E_P(Y|\tilde{z})\right\}\left\{\tilde{x} - E_P(X|\tilde{z})\right\} - \mathrm{cov}(Y,X|z)\right] \\ &+E[\partial_t\left\{Y - E_{P_t}(Y|Z)\right\}\left\{X - E_{P_t}(X|Z)\right\}|Z=z].\end{aligned}$$

Similarly to the covariance example previously, the final term above turns out to be zero. It follows, therefore, that the canonical gradient is

$$\phi(O,P) = \partial_t \Psi(P_t) = \left\{Y - E_P(Y|Z)\right\}\left\{X - E_P(X|Z)\right\} - \Psi(P),$$

and since this has finite variance, the expected conditional covariance is pathwise differentiable.

Constructing a one-step estimator or estimating equations estimator based on the canonical gradient of the expected conditional covariance is relatively straightforward and in fact both methods will provide the same result in this example. The one-step estimator takes an original plug-in estimator $\Psi(\hat{P}_n)$ and adds a correction term

$$\Psi(\hat{P}_n) + \frac{1}{n}\sum_{i=1}^{n}\phi(O_i,\hat{P}_n) = \frac{1}{n}\sum_{i=1}^{n}\left\{Y_i - \hat{m}(Z_i)\right\}\left\{X_i - \hat{\pi}(Z_i)\right\}.$$

where $\hat{m}(z) = E_{\hat{P}_n}(Y|z)$ and $\hat{\pi}(z) = E_{\hat{P}_n}(X|z)$. Estimation by this strategy therefore requires additional modelling to obtain the functions $\hat{m}(z)$ and $\hat{\pi}(z)$. □

Example 8 (average derivative effect). This example concerns the average derivative effect estimand of Härdle and Stoker (1989) with canonical gradient given by Newey and Stoker (1993). We let $m(x, z, P) = E_P(Y|X = x, Z = z)$ be a conditional response surface which is assumed to be differentiable w.r.t. x, with derivative $m'(x, z, P)$, and we also introduce a known weight function, $w(x, z)$. The average derivative effect estimand is written

$$\Psi(P) = E_P \{w(X, Z)m'(X, Z, P)\}.$$

Powell et al. (1989) showed that, for a differentiable function $g(x, z)$, with derivative w.r.t. x, $g'(x, z)$,

$$E_P \{w(X, Z)g'(X, Z)\} = E_P \{l(X, Z, P)g(X, Z)\}$$

under regularity conditions, which require that X is a continuous random variable and that $w(x, z)f(x, z)$ is differentiable w.r.t. x and is zero on the boundary of the support of X, where $f(x, z)$ used to denote the joint distribution of (X, Z) under P. In the above expression,

$$l(x, z, P) \equiv -w'(x, z) - w(x, z)f'(x, z)/f(x, z),$$

and, as before, superscript prime denotes the derivative with respect to x. Using (4.5),

$$\partial_t \Psi(P_t) = w(\tilde{x}, \tilde{z})m'(\tilde{x}, \tilde{z}) - \Psi(P) + E_P \{w(X, Z)\partial_t m'(X, Z, P_t)\}.$$

For the final term, we rely on Powell's identity:

$$\begin{aligned} E_P \{w(X, Z)\partial_t m'(X, Z, P_t)\} &= \partial_t E_P \{w(X, Z)m'(X, Z, P_t)\} \\ &= \partial_t E_P \{l(X, Z, P)m(X, Z, P_t)\} \\ &= E_P \left[l(X, Z, P) \frac{\mathbb{1}_{\tilde{x},\tilde{z}}(X, Z)}{f(X, Z)} \{\tilde{y} - m(X, Z, P)\} \right] \\ &= l(\tilde{x}, \tilde{z}, P) \{\tilde{y} - m(\tilde{x}, \tilde{z}, P)\}. \end{aligned}$$

Since this has finite variance, the average derivative effect is pathwise differentiable with canonical gradient

$$\phi(O, P) = \partial_t \Psi(P_t) = l(X, Z, P) \{Y - m(X, Z, P)\} + w(X, Z)m'(X, Z, P) - \Psi(P).$$

Using this efficient influence function, an efficient estimator may be easily derived following the one-step or estimating equation strategy. In this case both will result in the same estimator. Setting the sample average of $\phi(O_i, \hat{P}_n)$ to zero results in the estimator

$$\Psi(\hat{P}_n) = \frac{1}{n} \sum_{i=1}^{n} l(X_i, Z_i, \hat{P}_n) \left\{ Y_i - m(X_i, Z_i, \hat{P}_n) \right\} + w(X_i, Z_i)m'(X_i, Z_i, \hat{P}_n)$$

This estimator therefore requires modelling the functions $m(x, z, P), m'(x, z, P)$ and $l(x, z, P)$.

We include some extra examples in Appendix C.

4.6 Implementation

We begin by summarising the steps that need to be followed to go from scientific question to (data-adaptive) estimation described in the previous sections.

Step 1: Defining the estimand of interest.
The estimand $\Psi(P)$ is a nonparametrically defined statistical functional which is chosen with reference to the scientific question of interest. The estimand might be motivated for a variety of reasons, such as with reference to causal inference (e.g. example 5), independence testing (e.g. example 7), variable importance (e.g., Williamson et al. (2021a)), etc.

Step 2: Calculating its efficient influence function (under the nonparametric model). There are several ways to do this.

1. *Point-mass contamination.* We compute the Gâteaux/ pathwise derivative of $\Psi(P)$ at P in direction of a probability point mass $\widetilde{P}$. We consider a parametric submodel $P_t = (1-t)P + t\widetilde{P}$ for $t \in [0,1]$, which we use to evaluate the efficient influence function,

$$\phi(o, P) = \frac{d\Psi\left(P_t\right)}{dt}\Big|_{t=0}.$$

2. The most general method is to work from the definition of pathwise differentiability. Define a rich class of submodels P_t for $t \in (-\epsilon, \epsilon)$ such that $P_0 = P$ and $S(O) = \frac{d}{dt}\log f_t(o)|_{t=0}$ is the score function of t, where $f_t(o)$ denotes the density/probability point mass of O under P_t. Next one writes the derivative of $\Psi(P_t)$, as the integral,

$$\frac{d\Psi\left(P_t\right)}{dt}\Big|_{t=0} = \int \phi(o, P) S(o) dP(o).$$

   By the Riesz representation theorem, $\phi(o, P)$ is the efficient influence function.

3. For many new estimands, we can manipulate expressions so they can be expressed as simpler components (by applying the chain and product rules) and use known influence functions as building blocks.

Step 3: Obtaining an estimator based on the efficient influence function (such as one-step or TMLE) which admits a first-order representation, provided we use sample splitting or cross-fitting.

We recommend cross-fitting in the following way: First the data is partitioned into K *folds*, i.e. K smaller data sets of (roughly) equal size. Next, for each fold k, estimate the nuisance functionals (e.g. via data-adaptive methods) using the rest of all the data, excluding that in fold k. Use these nuisance functionals to evaluate the efficient influence function for each observations in fold k. After repeating for each fold, one is left with an estimate of the efficient influence function for all observations in the dataset. Finally, use these efficient influence function estimates to evaluate the estimator, and the error in the estimator. See Zheng and van der Laan (2011) and Chernozhukov et al. (2018) for more on cross-fitting.

4.7 Discussion

Statistical education still focuses primarily on parametric statistical models, which are assumed to reflect how the data is generated. The inferential theory that is taught does not reflect how data is usually analysed, where models are chosen data-adaptively and different models may fit equally well, especially nowadays, given the increased popularity of machine learning. We therefore believe that many courses would be better focused on translating a scientific question into an nonparametric estimand, and basing inference on its efficient influence function under the nonparametric model.

Courses and textbook treatments on the calculus of influence functions often focus on (semi)parametric models (Tsiatis, 2006). The resulting derivations can be challenging, as they require one to respect the restrictions that the model imposes on the observed data distribution. Moreover, they show how one can use these restrictions in order to make efficiency gains. Extracting information from modelling assumptions nevertheless comes at the risk of invalid inference when assumptions are violated. By contrast, our focus is on inference under a nonparametric model. This not only makes the resulting inferences more honest, but can dramatically simplify calculations. Additional efficiency gains are then reserved for special cases when restrictions are known to hold by the study design (Zhang et al., 2008), or reflect strong pre-existing scientific knowledge (Liu et al., 2021).

It is difficult, however, to proceed entirely nonparametrically and avoid regularity conditions all together. Indeed, without assumptions on distribution tails, inference of the mean, Example 1 in the current paper, is impossible (Bahadur and Savage, 1956; Bickel and Lehmann, 1975). Likewise, many of

our examples rely on working models for statistical functionals, necessitating certain regularity (Robins and Ritov, 1997). For instance, the one-step estimator for Example 5 requires estimating $m_1(Z_i, \hat{P}_n)$. Whilst flexible data-adaptive/ machine learning estimators can be used, these are better thought of as very highly parametric rather than nonparametric, and make assumptions on the true functional $m_1(z, P)$, e.g. that it is smooth in z. The crucial difference, however, is that compared with the parametric modelling approach, estimators based on the nonparametric model do not 'extract efficiency' from highly parametric modelling assumptions.

Because of the crucial role that efficient influence functions play, we focused on their derivation. Whilst the formal justification of the von Mises expansion relies on concepts from advanced mathematics, calculating the efficient influence function can often be done using techniques covered in a basic calculus course. We have illustrated this for several causal and non-causal statistical functionals (estimands); the method of derivation described can lead to simpler proofs than those in the original research papers.

Influence functions have applications beyond using them to define estimators with zero plug-in bias. Influence functions capture the stability of estimators to outliers (in fact this is one of their original purposes), which makes them additionally useful to diagnose outliers (as measurements with large influence function values). Recently, influence functions have started to be used in the machine learning literature too. For example, Koh and Liang (2017) used influence functions for interpretability of black-box models, by characterising the impact a data point has on the black-box's predictions. Curth et al. (2020) and Kennedy (2020) use influence functions as the outcome in machine learning procedures of conditional (e.g. subgroup-specific) estimands.

We hope that our contribution helps demystify the calculation of influence functions and thus encourages their wider adoption.

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT

T: +44 (0)20 7299 4646
F: +44 (0)20 7299 4656
www.lshtm.ac.uk

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	1804954	**Title**	Mr.
First Name(s)	Oliver		
Surname/Family Name	Hines		
Thesis Title	Assumption-Lean Inference for Causal and Statistical Questions in the Era of Machine Learning		
Primary Supervisor	Karla Diaz-Ordaz		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?			
When was the work published?			
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Choose an item.	Was the work subject to academic peer review?	Choose an item.

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	Journal of the American Statistical Association
Please list the paper's authors in the intended authorship order:	Oliver Hines, Karla Diaz-Ordaz, Stijn Vansteelandt
Stage of publication	**Undergoing revision**

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	This study was conceived by all authors. I carried out the mathematical research, computational simulations and writing of the manuscript under the supervision of the other authors.

SECTION E

Student Signature	Ohines
Date	14 December 2022

Supervisor Signature	Karla Diaz-Ordaz
Date	14 December 2022

Chapter 5

Variable importance estimands

5.1 Introduction

In the medical and social sciences there has been a longstanding interest in quantifying heterogeneity in the effects of treatments or interventions between groups of individuals. Understanding such heterogeneity is essential, for instance, in informing scientific research and optimizing treatment decisions. Suggestions to exploit this for personalised medicine have been expressed as early as the 1970's (if not before) (Senn, 2001; Rosenkranz, 2020), leading up to the domain of pharmacogenetics. Attention focused initially on subgroup analyses, which identify population subgroups (defined in terms of pre-treatment covariates) which benefit most/least from treatment, to be evaluated further in potential future studies; see e.g. Rothwell (2005) for a review, and Slamon et al. (2001) for the first clinical trial in oncology that was restricted to targeted trial populations. Typical challenges of subgroup analyses are how to select stratification variables in a systematic way, and how to handle the resulting multiplicity problem. Endeavours to address these were soon followed by methodological developments on personalised medicine in the causal inference literature, pioneered by Murphy (2003).

For many years, the primary focus was on policy learning; that is, determining the optimal treatment policy (which assigns the same treatment to individuals with the same measured covariate values) with the aim to minimise (some measure of) the population risk that would be seen if that policy were applied, uniformly in the population (van der Laan and Luedtke, 2014; Kallus, 2020; Athey and Wager, 2021). More recently, attention has shifted towards (machine-learning based) estimation of conditional average treatment effects (CATEs) (Abrevaya et al., 2015; Athey and Imbens, 2016; Nie and Wager, 2021; Kallus et al., 2018; Wager and Athey, 2018; Künzel et al., 2019; Kennedy, 2020; Knaus et al., 2021). Letting Y^a denoting the outcome that would be observed if treatment A were set to the value $a \in \{0, 1\}$, and X a vector of pre-treatment covariates, the CATE can be defined as $\tau(x) \equiv E(Y^1 - Y^0 | X = x)$ (Rubin, 1974). Estimates of the CATE can also been used for policy learning, for instance, the so-called optimal dynamic treatment rule (OTR) for an individual with covariate value x, assigns treatment based on the sign of $\tau(x)$. The CATE, however, additionally provides insight into the magnitude of the treatment effect for these individuals.

These foregoing developments are extremely useful and important, but they leave unanswered a key question that researchers commonly have when being presented an estimated CATE or OTR: namely, what are the key drivers of treatment effect heterogeneity? In the context of policy learning, attempts to find the optimal policy within a restricted class of 'simple' policies go some way towards answering this question, albeit at the risk of targeting a suboptimal policy (Zhang et al., 2015). Indeed, by comparing the mean outcome under the optimal policy, with the mean outcome under a suboptimal policy, which depends only on certain covariates, one may quantify the importance of the excluded covariates in determining the optimal policy (Williamson et al., 2021b).

In this paper, we will address this question by instead quantifying the importance of variable subsets in determining the CATE. By shifting focus from the OTR to the CATE, we argue that the resulting variable importance measures are easier to infer as well as more interesting from a scientific perspective,

since they provide greater insight into effect modifiers. For instance, it may be the case that a particular treatment is uniformly beneficial, in which case the optimal policy is to always treat, despite extremely large treatment effect heterogeneity. An understanding of such heterogeneity may, e.g. provide information about treatment mechanism, suggest future therapies, be used to compare clinical trial populations, and be used to quantify systematic treatment biases (such as on the basis of race or socio-economic status).

One existing proposal for CATE variable attribution is based on the 'causal forest' CATE estimator, which extends the widely used random forest algorithm to CATE estimation (Athey et al., 2019; Wager and Athey, 2018; Athey and Imbens, 2016; Breiman, 2001a). The resulting variable importance measures (VIMs) rely on the 'tree architecture' of causal forest models, thus are inherently tied to the CATE estimation strategy, and have also been criticised as they tend to assign greater importance to continuous variables, or categorical variables with many categories (Grömping, 2009; Strobl et al., 2007). Generally, VIMs which depend on a particular modelling strategy (e.g. random forests/ linear regression) are referred to as 'algorithmic' (Williamson and Feng, 2020), with the disadvantage that algorithmic VIMs are well defined and interpretable only within their particular modelling strategy.

The need to define generic nonparametric VIMs is closely related to the need for methods which explain the output of 'black box' machine learning prediction algorithms, itself an active area of research in the computer science literature. One popular approach is based on so-called Shapley additive explanation (SHAP) values (Lundberg and Lee, 2017), which quantify the direction and magnitude of each covariate in obtaining model predictions. Applications of SHAP to CATE estimation are rare, with Syrgkanis et al. (2019) being a notable recent example. Debate remains, however, over exactly how SHAP values should be defined and interpreted causally (Janzing et al., 2020; Chen et al., 2020).

In view of these issues, we propose treatment effect variable importance measures (TE-VIMs), which are model-free scalar summary statistics intended to measure the importance of subsets of covariates in predicting the CATE. In particular, our proposed estimands quantify the contribution of covariate subsets towards the variance $\text{var}\{\tau(X)\}$ of the treatment effect (VTE) (Levy et al., 2021); the latter quantifies treatment effect heterogeneity by capturing the extent to which varying effects of treatment can be explained by observed covariates. More precisely, we will consider the estimand

$$\Theta_s \equiv E[\text{var}\{\tau(X)|X_{-s}\}] = \text{var}\{\tau(X)\} - \text{var}\{\tau_s(X)\}, \tag{5.1}$$

where $X \in \mathbb{R}^p$ represents a p-dimensional covariate vector, the symbol u_{-s} denotes the vector of all the components of u with index not in $s \subseteq \{1, ..., p\}$, and $\tau_s(x) \equiv E\{\tau(X)|X_{-s} = x_{-s}\}$ denotes the CATE conditional on X_{-s}; note that $\tau_s(x)$ only depends on x_{-s}, but we write it as a function of x for simplicity of notation. We interpret $\Theta_s \geq 0$ as a difference in VTEs, quantifying the amount by which the VTE changes when variables in the set s are excluded from the model. More formally, it expresses the additional treatment effect heterogeneity explained by X_s, over and above that already explained by X_{-s}, where for a vector u, we let u_s denote the vector of all components of u with index in s.

The proposed TE-VIMs rescale Θ_s by the VTE to express this difference as a proportion,

$$\Psi_s \equiv \frac{\Theta_s}{\text{var}\{\tau(X)\}} = 1 - \frac{\text{var}\{\tau_s(X)\}}{\text{var}\{\tau(X)\}}. \tag{5.2}$$

Assuming that the VTE is non-zero, we interpret $\Psi_s \in [0, 1]$ as the proportion of treatment effect heterogeneity explained by X_{-s} compared with X. This interpretation is analogous to the familiar coefficient of determination (R^2 statistic). The proposed TE-VIMs connect VTE estimands (Levy et al., 2021) to recently proposed regression-VIMs (Williamson et al., 2021a; Zhang and Janson, 2020), also referred to as 'leave out covariates' (Verdinelli and Wasserman, 2021; Lei et al., 2018), and the more general VIM framework by Williamson et al. (2021b). In this way, our work represents a step towards extending VIMs to the analysis of more general statistical functionals, as discussed in Section 5.5.

In Section 5.2 we motivate TE-VIMs, and provide estimators which are efficient under the nonparametric model. These rely on estimating working models, relating our proposal to the DR-learner of the CATE through an interpretation based on so-called pseudo-outcomes (Kennedy, 2020; Luedtke and van der Laan, 2016; van der Laan, 2013). Experimental results on simulated data are provided in Section 5.3 and Section 5.4 demonstrates an application to the AIDS Clinical Trials Group Protocol 175 (Hammer et al., 1996).

5.2 Methodology

5.2.1 Motivating the estimand

Suppose we have n i.i.d. observations $(z_1, ..., z_n)$ of a random variable Z distributed according to an unknown distribution P, such that Z consists of (Y, A, X), where $Y \in \mathbb{R}$ is an 'outcome', $A \in \{0, 1\}$ is an 'exposure' or 'treatment' and $X \in \mathbb{R}^p$ is a p-dimensional vector of covariates. Under standard identification assumptions of consistency ($A = a \implies Y = Y^a$), conditional exchangeability ($Y^a \perp\!\!\!\perp A|X$ for $a = 0, 1$), and positivity ($0 < \pi(X) < 1$ w.p.1), the CATE is identified by $\tau(x) \equiv E(Y^1 - Y^0|X = x) = \mu(1, x) - \mu(0, x)$, where $\mu(a, x) \equiv E(Y|A = a, X)$ and $\pi(x) \equiv E(A|X = x)$ denotes the 'propensity score'.

Assume that $||\tau|| < \infty$, where $||f|| \equiv E\{f(X)^2\}^{1/2}$ is the $L_2(P)$ norm. With this choice our estimand is finite and well defined, since

$$\Theta_s = E[\{\tau(X) - \tau_s(X)\}^2] = ||\tau - \tau_s||^2 < \infty,$$

Notice that the VTE is Θ_p, where, with a slight abuse of notation, p denotes the index set $\{1, ..., p\}$ and τ_p is the ATE. We further assume that the VTE is non-zero, i.e. $\Theta_p > 0$ and since $\Theta_p \geq \Theta_s$, it follows that $\Psi_s = \Theta_s/\Theta_p \in [0, 1]$.

The regression-VIM in Williamson et al. (2021a) is analogous to our proposal, in the sense that the former replaces $\tau(x)$ with $\mu(x)$ and $\tau_s(x)$ with $\mu_s(x) \equiv E(Y|X_{-s} = x_{-s})$. Specifically, they consider,

$$\theta_s \equiv E[\text{var}\{\mu(X)|X_{-s}\}] = E[\{\mu(X) - \mu_s(X)\}^2]$$

which is analogous to Θ_s. The two proposals, however, differ in how this mean conditional variance is scaled. Williamson et al. (2021a) consider scaling by the outcome variance, i.e. by defining $\psi_s \equiv \theta_s/\text{var}(Y)$, whereas we scale by the VTE, i.e. $\Psi_s = \Theta_s/\Theta_p$. We scale by the VTE because the treatment effect variance, $\text{var}(Y^1 - Y^0)$, is generally not identifiable without strong assumptions (Levy et al., 2021; Ding et al., 2016; Heckman et al., 1997). The VTE, however, is a convenient scaling parameter which bounds $\Psi_s \in [0, 1]$, aiding interpretability since, when the VTE is non-zero, Ψ_s behaves like a coefficient of determination (R^2).

In practice, the scaling factor makes little difference to the interpretation of our estimands, since investigators are likely to compare the relative importance of covariate sets s and s' by comparing the magnitudes of Ψ_s and $\Psi_{s'}$. This approach is demonstrated in Section 5.4, where the importance of each covariate is ranked individually using Ψ_s where s is a set containing a single covariate of interest. Quantifying variable importance in this way, however, may be problematic when covariates are themselves highly correlated. An alternative could be to define importance with reference to Ψ_s where s is the set of all covariates except the covariate of interest, or else to define variable importance with reference to so-called Shapley population VIMs, which consider all possible covariate permutations that do not include the variable of interest (Owen and Prieur, 2017; Williamson and Feng, 2020). To rank all covariates, the latter are highly computationally intensive, requiring CATE estimates for each of the 2^p possible covariate subsets. Instead, Williamson et al. (2021b) recommend using domain specific knowledge to group covariates, e.g. one might compare the relative importance of biological factors vs. non-biological factors in determining the CATE. We remark that decreasing the index set can never increase the TE-VIM, in the sense that $s' \subseteq s$ implies that $\Psi_{s'} \leq \Psi_s$, i.e. the covariate set s' cannot be more important than s.

5.2.2 CATE estimation

Estimation of the proposed TE-VIM will rely on initial CATE estimates, obtained via flexible machine learning based methods (Knaus et al., 2021), which we review first. CATE estimation is challenging since common machine learning algorithms (random forests, neural networks, boosting etc.) are instead designed for mean outcome regression, such as by minimising the mean squared error loss. CATE estimation strategies therefore either modify existing machine learning methods to target CATEs, e.g. Athey et al. (2019); Wager and Athey (2018) and Athey and Imbens (2016) modify the random forest algorithm for CATE estimation. Alternatively, 'metalearning' strategies decompose CATE estimation into a sequence

of sub-regression problems, which can be solved using off-the-shelf machine learning algorithms, see e.g. Künzel et al. (2019); Nie and Wager (2021); Kennedy (2020).

In the current work we focus on two metalearning algorithms which, following the naming convention of Künzel et al. (2019), we refer to as the T-learner and the DR-learner. The T-learner is based on the decomposition $\tau(x) = \mu(1,x) - \mu(0,x)$, and the T-learner estimate of the CATE is $\hat{\tau}^{(T)}(x) \equiv \hat{\mu}(1,x) - \hat{\mu}(0,x)$, where $\hat{\mu}(a,x)$ represents an estimate of $\mu(a,x)$ obtained by a regression of Y on (A,X). The T-learner, however is problematic for two main reasons. Firstly, whilst regularisation methods can be used to control the smoothness of $\hat{\mu}(a,x)$, the same is not true of $\hat{\tau}^{(T)}(x)$ which may be highly erratic. Slow convergence rates affecting $\hat{\mu}(a,x)$ may therefore propagate into the CATE estimator $\hat{\tau}^{(T)}(x)$. Secondly, $\hat{\mu}(1,x)$ is chosen to make an optimal bias-variance trade-off over the covariate distribution of the treated population. Likewise, $\hat{\mu}(0,x)$ is chosen to make an optimal bias-variance trade-off over the covariate distribution of the untreated population. When there is poor overlap between the treated and untreated subgroups, then the difference $\hat{\mu}(1,x) - \hat{\mu}(0,x)$ may fail to deliver an optimal bias-variance trade-off over the population covariate distribution, making the T-learner potentially poorly targeted towards CATE estimation.

The DR-learner (Kennedy, 2020; Luedtke and van der Laan, 2016; van der Laan, 2013) is an alternative metalearning algorithm based on the decomposition $\tau(x) = E\{\varphi(Z)|X=x\}$ where

$$\varphi(z) \equiv \{y - \mu(a,x)\}\frac{a - \pi(x)}{\pi(x)\{1-\pi(x)\}} + \mu(1,x) - \mu(0,x).$$

is called the 'pseudo outcome', or the augmented inverse propensity weighted score (Robins, 1994), which acts like the causal contrast, $Y^1 - Y^0$, in expectation. The DR-learning procedure first estimates $\mu(a,x)$ and $\pi(x)$ to obtain the pseudo-outcome plug-in estimator

$$\hat{\varphi}(z) \equiv \{y - \hat{\mu}(a,x)\}\frac{a - \hat{\pi}(x)}{\hat{\pi}(x)\{1-\hat{\pi}(x)\}} + \hat{\mu}(1,x) - \hat{\mu}(0,x),$$

In a second step, the estimated pseudo-outcome, $\hat{\varphi}(Z)$ is regressed on covariates X to obtain $\hat{\tau}^{(DR)}(x)$. A sample splitting scheme is also recommended, whereby the regression steps to obtain $\hat{\mu}(a,x)$, $\hat{\pi}(x)$, and $\hat{\tau}^{(DR)}(x)$ are performed on three independent samples, see Kennedy (2020) for details.

The DR-learner alleviates the issues related to the T-learner since the complexity of $\hat{\tau}^{(DR)}(x)$ can be controlled by regularising the regression in the final stage of the procedure, mitigating concerns regarding the smoothness of the T-learner. With regard to consistency, the square of $E\{\hat{\varphi}(Z)|X=x\} - \tau(x)$ is bounded by at most the product of the squared errors of the propensity score and regression estimators (up to constant scaling). In practice, this means that the final regression step, where $\hat{\varphi}(Z)$ is regressed on X, mimics the oracle regression of $\varphi(Z)$ on X provided that

(A1) The propensity score and outcome estimators are 'rate double robust' in the sense that $\{\pi(x) - \hat{\pi}(x)\}\{\mu(a,x) - \hat{\mu}(a,x)\}$ is $o_P(n^{-1/2})$ in $L_2(P)$ norm for $a = 0, 1$.

This requirement implies that one can trade-off accuracy in the outcome and propensity score estimators, a property which is known as rate double robustness, hence the name 'DR-learner'.

Estimation of the CATE $\tau_s(x)$ is complicated by the fact that one cannot assume that $Y \perp\!\!\!\perp A|X_{-s}$ for an arbitrary subset of covariates s, a problem that is sometimes referred to as 'runtime confounding' (Coston et al., 2020). The DR-learner readily extends to the setting of runtime confounding through the decomposition $\tau_s(x) = E\{\varphi(Z)|X_{-s} = x_{-s}\}$. This expression implies that one may estimate $\tau_s(x)$ by regressing $\hat{\varphi}(Z)$ on X_{-s}, i.e. by modifying the final regression step of the DR-learner.

In this paper we propose a metalearner for $\tau_s(x)$ based on the decomposition $\tau_s(x) = E\{\tau(X)|X_{-s} = x_{-s}\}$. Specifically we propose regressing an initial estimate of the CATE, $\hat{\tau}(X)$ on X_{-s} to obtain an estimate of $\tau_s(x)$. This approach is agnostic the initial CATE estimator used and, like the DR-learner, one can control the complexity of the resulting CATE estimator $\hat{\tau}_s(x)$ by regularisation. We advocate this approach since it generally results in estimates of $\tau_s(x)$, which are compatible with those of $\tau(x)$, similar to Williamson et al. (2021a), who recommend estimating $\mu_s(x)$ by regressing an estimate $\hat{\mu}(x)$ of $\mu(x)$ on X_{-s}.

5.2.3 TE-VIM estimation

Efficient estimators of Θ_s

Next we consider estimators based on the efficient influence curve (IC) of Θ_s under the nonparametric model. Briefly, ICs are model-free, mean zero, functionals that characterise the sensitivity of an estimand to small changes in the data generating law. As such, ICs are useful for constructing efficient estimators and determining their asymptotic distribution, see e.g. Hines et al. (2022); Fisher and Kennedy (2020) for an introduction to these methods. In Appendix D we derive the IC of Θ_s at a single observation $z = (y, a, x)$ of Z to be

$$\phi_s(z) = \{\varphi(z) - \tau_s(x)\}^2 - \{\varphi(z) - \tau(x)\}^2 - \Theta_s. \tag{5.3}$$

The interpretation of $\varphi(Z)$ as a pseudo-outcome which plays the role of the unobserved causal contrast $Y^1 - Y^0$ holds in the present context. To see why, we compare the IC in (5.3) with that of θ_s given by Williamson et al. (2021a),

$$\{y - \mu_s(x)\}^2 - \{y - \mu(x)\}^2 - \theta_s.$$

The IC of θ_s has the same form as (5.3), since the latter is recovered by replacing the outcome y with the pseudo-outcome, $\varphi(z)$, replacing the conditional mean outcomes with CATEs, i.e. conditional means of $\varphi(Z)$, and replacing θ_s with Θ_s.

Efficient estimating equations estimators can be derived from ICs by setting the sample mean of (an estimate of) the IC to zero. In the current setting, this strategy is equivalent to the so-called one-step correction which we outline in Appendix D. For Θ_s and θ_s, we thus obtain the estimators

$$\begin{aligned}\hat{\Theta}_s &\equiv n^{-1}\sum_{i=1}^{n}\{\hat{\varphi}(z_i) - \hat{\tau}_s(x_i)\}^2 - \{\hat{\varphi}(z_i) - \hat{\tau}(x_i)\}^2 \\ \hat{\theta}_s &\equiv n^{-1}\sum_{i=1}^{n}\{y_i - \hat{\mu}_s(x_i)\}^2 - \{y_i - \hat{\mu}(x_i)\}^2,\end{aligned} \tag{5.4}$$

where $\hat{\tau}(x)$ and $\hat{\tau}_s(x)$ and $\hat{\varphi}(z)$ are consistent estimators fitted on an independent sample. In practice, a cross-fitting approach may be used to obtain the fitted models and evaluate $\hat{\Theta}_s$ from a single sample (Zheng and van der Laan, 2011; Chernozhukov et al., 2018). Theorem 1 below states the asymptotic distribution of $\hat{\Theta}_s$ under (A1) and when,

(A2) The differences $\tau(x) - \hat{\tau}(x)$ and $\tau_s(x) - \hat{\tau}_s(x)$ are both $o_P(n^{-1/4})$ in $L_2(P)$ norm.

The requirement for $n^{1/4}$ rate convergence in (A2) is standard in the recent VIM framework of Williamson et al. (2021b), for instance, an assumption similar to (A2) appears in the regression-VIM setting for the outcome models $\hat{\mu}(x)$ and $\hat{\mu}_s(x)$. In the TE-VIM setting, however, one is also required to estimate the pseudo-outcomes, with (A1) required to control the error which arises from doing so.

Assumptions (A1) and (A2) suggest that the DR-learner of the CATE may be preferable to the T-learner due to its robustness properties. In particular, consider that the T-learner of the CATE $\hat{\tau}^{(T)}(x)$ satisfies the first condition of (A2), provided that $\mu(a, x) - \hat{\mu}(a, x)$ is $o_P(n^{-\delta})$ in $L_2(P)$ norm, with $\delta \geq 1/4$. (A1) then implies that $\pi(x) - \hat{\pi}(x)$ must be at least $o_P(n^{-1/2+\delta})$ in $L_2(P)$ norm. In other words, the propensity score estimator is allowed to converge at a slower rate, provided the outcome estimator converges at a faster rate, but the converse is not true. This is unsatisfying for example in clinical trial settings, where the exposure is randomised and the propensity score model is known, since the T-learner of the CATE would still require $n^{1/4}$ rate convergence of the outcome model.

Conversely, the DR-learner $\hat{\tau}^{(DR)}(x)$ satisfies the first condition of (A2), provided that (A1) holds and $E\{\hat{\varphi}(Z)|X = x\} - \hat{\tau}^{(DR)}(x)$ is $o_P(n^{-1/4})$ in $L_2(P)$ norm, i.e. provided the final DR-learning regression estimator is consistent at $n^{-1/4}$ rate. Applying the same reasoning as before, we see that (A1) implies that $\mu(a, x) - \hat{\mu}(a, x)$ can be $o_P(n^{-\delta})$ in $L_2(P)$ norm, provided that $\pi(x) - \hat{\pi}(x)$ is $o_P(n^{-1/2+\delta})$ in $L_2(P)$ norm, for any $\delta \geq 0$. In other words, the outcome estimator is allowed to converge at a slower rate, provided the propensity score estimator converges at a faster rate and vice-versa, which marks an improvement over the T-learner. The asymptotic distribution of Θ_s under (A1) and (A2) is given Theorem 3 below.

Theorem 3 *Under (A1), (A2) and regularity assumptions given in Appendix D, $\hat{\Theta}_s$ is asymptotically linear with IC, $\phi_s(Z)$, and hence $\hat{\Theta}_s$ converges to Θ_s in probability, and for $\Theta_s > 0$ then $n^{1/2}(\hat{\Theta}_s - \Theta_s)$ converges in distribution to a mean-zero normal random variable with variance $E\{\phi_s^2(Z)\}$.*

Importance testing

As well as being used to derive efficient estimators, the IC characterises the asymptotic distribution of such estimators, as in Theorem 3. Since $\hat{\Theta}_s$ and $\hat{\theta}_s$ are analogous, we expect them to share similar properties. One such property concerns the behaviour under the zero-importance null hypothesis, $H_0 : \Theta_s = 0$, i.e. when $\tau(x) = \tau_s(x)$. In practice, this hypothesis corresponds to treatment effect homogeneity over X_s given X_{-s}, thus the IC in (5.3) is exactly zero. The fact that the IC degenerates in this way makes H_0 difficult to test, since Wald-type tests, i.e. those based on the asymptotic distribution of $\hat{\Theta}_s$, tend to overestimate the variance of $\hat{\Theta}_s$ and are therefore conservative. For this reason, Theorem 3 gives the behaviour of the estimator only when $\Theta_s > 0$.

One solution to the IC degeneracy problem under H_0, proposed by Williamson et al. (2021b), is to estimate $\text{var}\{\tau(X)\}$ and $\text{var}\{\tau_s(X)\}$ using efficient estimators in separate samples, with each estimand having a non-zero IC provided that $\text{var}\{\tau_s(X)\} > 0$, despite the two ICs being identical under H_0. With this condition, both estimators would be independent and asymptotically normal, hence their difference, which represents an estimator of Θ_s, would also be asymptotically normal even when $\Theta_s = 0$. One therefore obtains a valid Wald-type test for H_0, at the expense of using an estimator for Θ_s, which is inefficient due to the requirement for sample splitting. Similarly, one could test the zero-VTE null hypothesis ($\text{var}\{\tau(X)\} = 0$) by estimating $E\{\tau^2(X)\}$ and $E\{\tau(X)\}^2$ using efficient estimators in separate samples and taking their difference. Fundamentally, however, the distribution of $\hat{\Theta}_s$ under H_0 depends on higher-order pathwise derivatives of the estimand, see e.g. Carone et al. (2018), and remains generally an open problem.

Targeted maximum likelihood estimation

For the index set $s = p$, $\hat{\Theta}_p$ is an estimator of the VTE, which is distinct from the targeted maximum likelihood estimation (TMLE) VTE estimator proposed by Levy et al. (2021). Both are based on initial estimates of $\mu(a, x)$ and $\pi(x)$ and are regular asymptotically linear in the sense of ensuring that the sample mean of the estimated IC is negligible. The TMLE estimator achieves this by replacing the initial estimates $\hat{\mu}(a, x)$, with 'retargeted' estimates $\hat{\mu}^*(a, x)$. These retargeted estimates are used to estimate the CATE as $\hat{\tau}^*(x) \equiv \hat{\mu}^*(1, x) - \hat{\mu}^*(0, x)$, with the ATE and VTE obtained respectively as the sample mean and sample variance of $\hat{\tau}^*(X)$. Unlike $\hat{\Theta}_p$, the TMLE VTE estimator is a 'plug-in' estimator, in the sense that it maps a probability distribution (implied by the empirical measure of the covariates, the propensity score estimator $\hat{\pi}(x)$ and the targeted outcome estimator $\hat{\mu}^*(a, x)$) onto an estimated value using the definition of Θ_p. See e.g. Hines et al. (2022) for an introductory comparison of TMLE estimators with one-step bias correction estimators.

Efficient estimators of Ψ_s

Our main goal is to consider efficient estimation of the rescaled TE-VIM $\Psi_s = \Theta_s/\Theta_p$, which has IC,

$$\Phi_s(z) = \{\phi_s(z) - \Psi_s \phi_p(z)\}/\Theta_p, \tag{5.5}$$

where $\phi_p(.)$ denotes (5.3) for the index set $s = p$. This IC implies an estimating equations estimator, $\hat{\Psi}_s = \hat{\Theta}_s/\hat{\Theta}_p$. Like $\phi_s(z)$, the IC $\Phi_s(z)$ is also degenerate, with $\Phi_s(z) = 0$ when $\Psi_s = 0$ and when $\Psi_s = 1$, corresponding to $\Theta_s = 0$ and $\Theta_s = \Theta_p$ respectively. For this reason, the asymptotic result in Theorem 4 below, holds only when $\Psi_s \in (0, 1)$, i.e. when the covariate set s accounts for some, but not all, CATE variability. We make the additional assumption (B1) that the difference $\tau_p - \hat{\tau}_p$ is $o_P(n^{-1/4})$.

Theorem 4 *Assume that $\Theta_p \neq 0$. Under (A1), (A2), (B1) and regularity assumptions given in Appendix D, $\hat{\Psi}_s$ is asymptotically linear with IC, $\Phi_s(Z)$, and hence $\hat{\Psi}_s$ converges to Ψ_s in probability, and for*

$\Psi_s \in (0,1)$ *then* $n^{1/2}(\hat{\Psi}_s - \Psi_s)$ *converges in distribution to a mean-zero normal random variable with variance* $E\{\Phi_s(Z)^2\}$.

Proposed algorithms for estimating Ψ_s

The estimator $\hat{\Theta}_s$ is indexed by the choice of pseudo-outcome and CATE estimators. Generally, we are not constrained to any particular learning method. Throughout, we consider a plug-in estimator for $\varphi(.)$, where models for $\mu(a,x)$ and $\pi(x)$ are obtained from the same sample, and these are used to construct $\hat{\varphi}(z)$. We consider two approaches to learning $\tau(x)$ based on the T- and DR-learning strategies, though for computational simplicity, our 'DR-learner' does not use the sample splitting scheme proposed by Kennedy (2020).

We expect the estimator based on the DR-learning approach to be less biased in finite samples than that based on the T-learning approach, due to the robustness properties of the DR-learner and since it explicitly seeks to minimise the term, $n^{-1}\sum_{i=1}^{n}\{\hat{\varphi}(z_i) - \hat{\tau}(x_i)\}^2$, which appears in (5.4). This term can be problematic in practice, since it may give negative $\hat{\Theta}_s$ estimates, when $\hat{\tau}(.)$ converges slowly to $\tau(.)$.

In the regression-VIM work of Williamson et al. (2021a), the authors found that regressing the observed outcome on X and X_{-s} returns conditional mean models $\hat{\mu}(x)$ and $\hat{\mu}_s(x)$ which do not take into account that the two conditional means are related, generally resulting in incompatible estimates. Their solution was to first regress the outcome on X then regress predictions from the resulting conditional mean model on X_{-s}. We propose a similar approach to learning $\hat{\tau}_s(x)$, effectively using the 'runtime confounding' CATE metalearning strategy described in Section 5.2.2.

The proposed working function estimators are implemented in Algorithms 1 and 2 below. These algorithms return pseudo-outcome and CATE estimates, $\{\hat{\varphi}_i\}_{i=1}^n, \{\hat{\tau}_i\}_{i=1}^n,$, $\{\hat{\tau}_{s,i}\}_{i=1}^n$, and $\{\hat{\tau}_{p,i}\}_{i=1}^n$, which can be used to obtain $\hat{\Psi}_s = \hat{\Theta}_s/\hat{\Theta}_p$, with variance estimated by $n^{-2}\sum_{i=1}^n \hat{\phi}_i^2$, where,

$$\hat{\Theta}_s = n^{-1}\sum_{i=1}^{n}\{\hat{\varphi}_i - \hat{\tau}_{s,i}\}^2 - \{\hat{\varphi}_i - \hat{\tau}_i\}^2$$

$$\hat{\phi}_i = \frac{1}{\hat{\Theta}_p}\left[\{\hat{\varphi}_i - \hat{\tau}_{s,i}\}^2 - \hat{\Psi}_s\{\hat{\varphi}_i - \hat{\tau}_{p,i}\}^2 + (\hat{\Psi}_s - 1)\{\hat{\varphi}_i - \hat{\tau}_i\}^2\right]$$

and $\hat{\Theta}_p$ is defined in a similar manner. Algorithm 2 uses a cross-fitting regime to ensure that $\hat{\varphi}_i, \hat{\tau}_i, \hat{\tau}_{s,i}$, and $\hat{\tau}_{p,i}$ are constructed from estimators which are not fitted using the ith observation. This is useful in controlling the so-called empirical process term, see e.g. Newey and Robins (2018); Hines et al. (2022). Both algorithms are also indexed by the choice of CATE learner in steps 2 and 3 of each algorithm respectively, with the substeps marked (A) and (B) referring to the T- and DR-learning strategies. We note that where the algorithms require models to be 'fitted', any suitable regression method can be used.

Algorithm 1 - Without sample splitting

(1) Fit $\hat{\mu}(.,.)$ and $\hat{\pi}(.)$. Use these fitted models to obtain $\hat{\varphi}_i \equiv \hat{\varphi}(z_i)$.

(2) (A) Use the model for $\hat{\mu}(.,.)$ from Step 1, to obtain $\hat{\tau}(x) \equiv \hat{\mu}(1,x) - \hat{\mu}(0,x)$. Or (B) Fit $\hat{\tau}(.)$ by regressing $\hat{\varphi}(Z)$ on X. After doing (A) or (B), use the fitted models to obtain $\hat{\tau}_i \equiv \hat{\tau}(x_i)$.

(3) Fit $\hat{\tau}_s(.)$ by regressing $\hat{\tau}(X)$ on X_{-s}. Use the fitted model to obtain $\hat{\tau}_{s,i} \equiv \hat{\tau}_s(x_i)$.

(4) Repeat Step 3 for the covariate set p and (optionally) any other covariate sets of interest.

Algorithm 2 - With sample splitting

(1) Split the data into K folds.

(2) **For** each fold k: Fit $\hat{\mu}(.,.)$ and $\hat{\pi}(.)$ using the data set excluding fold k. Use these fitted models to obtain $\hat{\varphi}_i \equiv \hat{\varphi}(z_i)$ for i in fold k.

(3) (A) Use the model for $\hat{\mu}(.,.)$ from Step 2, to obtain $\hat{\tau}(x) \equiv \hat{\mu}(1,x) - \hat{\mu}(0,x)$. Or (B) Fit $\hat{\tau}(.)$ by regressing $\hat{\varphi}(Z)$ on X using the data excluding fold k. After doing (A) or (B), use the fitted models to obtain $\hat{\tau}_i \equiv \hat{\tau}(x_i)$ for i in fold k.

(4) Fit $\hat{\tau}_s(.)$ by regressing $\hat{\tau}(X)$ on X_{-s} using the data excluding fold k. Use the fitted model to obtain $\hat{\tau}_{s,i} \equiv \hat{\tau}_s(x_i)$ for i in fold k.

(5) Repeat Step 4 for the covariate set p and (optionally) any other covariate sets of interest. **End for.**

5.3 Simulation study

In our simulation study we compared Algorithms 1A, 1B, 2A and 2B on generated data in finite samples, using $K = 5$ fold sample splitting. We generated 1000 datasets of size $n \in \{500, 1000, 2000, 3000, 4000\}$ from the following structural equation model

$$\begin{aligned} X_1, X_2 &\sim \text{Uniform}(-1, 1) \\ A &\sim \text{Bernoulli}\{\text{expit}(-0.4X_1 + 0.1X_1X_2)\} \\ Y &\sim \mathcal{N}\left(\{X_1X_2 + 2X_2^2 - X_1\} + A\tau(X), 1\right) \end{aligned}$$

where the CATE is given by $\tau(X) = X_1^2(X_1 + 7/5) + 25X_2^2/9$. Analytically computing the true estimand values gives that the TE-VIMs are $\Psi_1 = 0.32$ and $\Psi_2 = 0.68$, with the ATE and VTE taking the values $\tau_p = 1.39$ and $\Theta_p = 1.00$ respectively.

For each dataset, $\hat{\Psi}_s$ was estimated along with its variance (using the variance estimators above), standard error (as the square root of the variance), and associated Wald based (95%) confidence intervals for the index sets, $s = \{1\}, \{2\}$. Two regression algorithms were considered for estimation of $\mu(a,x)$, $\pi(x)$, $\tau_s(x)$, and in the case of the DR-learner, $\tau(x)$. The first regression algorithm used generalised additive models, as implemented through the `mgcv` package in R (Wood et al., 2016). These are flexible spline smoothing models with interaction terms and for the propensity score model a logistic link function was used. The second regression algorithm used random forest learners available through the `ranger` package in R (Wright and Ziegler, 2017).

Figure 5.1 shows empirical estimates of the bias and variance of $\hat{\Psi}_1$ scaled by $n^{1/2}$ and n respectively, as well as 95% Wald based confidence-interval coverage probabilities. Similar plots for $\hat{\Psi}_2$ are in Appendix D. Comparing Algorithms 1 and 2 (i.e. no sample splitting vs sample splitting), we notice a greater difference in the results when random forest learning is used, with sample splitting generally reducing bias, increasing variance and improving confidence interval coverage.

Additionally, the DR-learning approach (Algorithm B) outperforms the T-learning approach (Algorithm A) in the sense of reducing the bias and achieving confidence interval coverage closer to the 95% level across all sample sizes. This improvement is due to the DR-learner making better use of the propensity score model to improve estimation of the CATE. On the basis of these results, we recommend Algorithm 2B for TE-VIM learning.

5.4 Applied example: variable importance of treatment effect heterogeneity in HIV

We demonstrate our estimators on data from the AIDS Clinical Trials Group Protocol 175 (ACTG175) (Hammer et al., 1996), which evaluated 2139 patients infected with HIV whose CD4 T-cell count was in the range 200 to 500 mm^{-3}. Patients were randomised to 4 treatment groups: (i) zidovudine (ZDV) monotherapy, (ii) ZDV+didanosine (ddI), (iii) ZDV+zalcitabine, and (iv) ddI monotherapy. We compare treatment groups (iv) and (ii) as in Lu et al. (2013); Cui et al. (2020). These two groups are represented with the binary indicator, $A = 0, 1$, with $n = 561$ and $n = 522$ patients in each group respectively.

Previous studies have used ACTG175 data to analyse the causal effect of A on a survival time endpoint, and the data is available through the `speff2trial` package in R. We consider CD4 count at 20±5 weeks

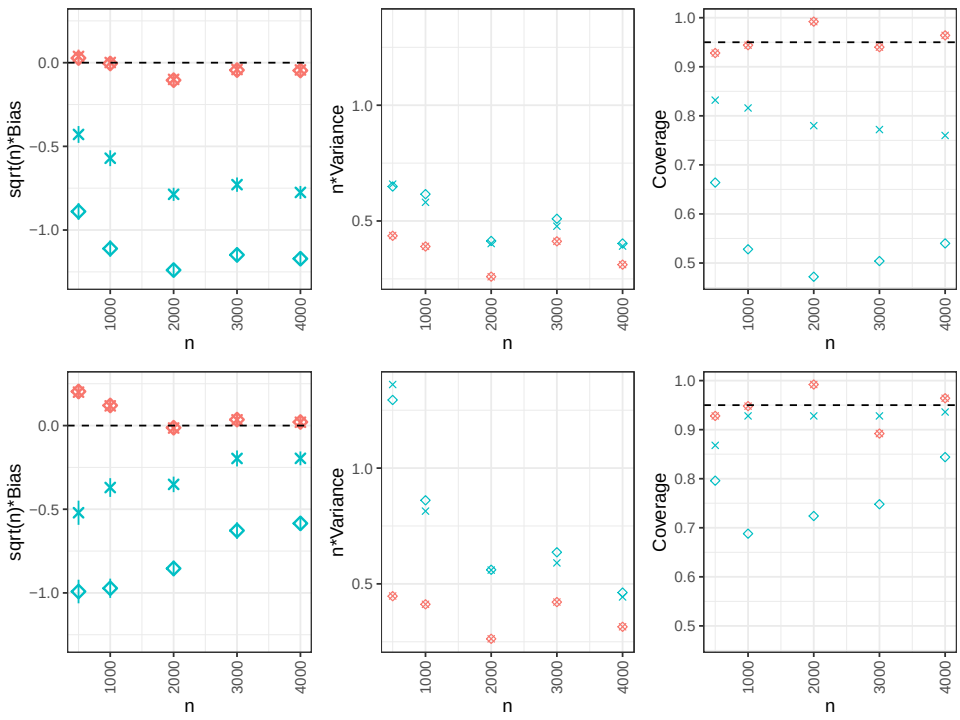


Figure 5.1: Bias, variance and coverage for $\hat{\Psi}_1$ using 1000 sampled datasets. Red and blue points indicate that working models are fitted using generalised additive modelling and random forests respectively. Top row of plots corresponds to Algorithm 1 (no sample splitting) and the bottom row corresponds to Algorithm 2 (sample splitting). Square and crossed points indicate that the algorithm used the T-learner and DR-learner respectively for CATE estimation. Similar plots for $\hat{\Psi}_2$ are given in Appendix D.

as a continuous outcome, Y, and consider 12 baseline covariates, 5 continuous: age, weight, Karnofsky score, CD4 count, CD8 count; and 7 binary: sex, homosexual activity (y/n), race (white/non-white), symptomatic status (symptomatic/asymptomatic), history of intravenous drug use (y/n), hemophilia (y/n), and antiretroviral history (experienced/naive).

TE-VIMs for each covariates were estimated using all algorithms with $K = 20$ folds (between 10 to 20 folds is typical for cross-fitting procedures). Propensity score estimates were obtained as the mean of the treatment indicator in the training set. This model is correctly specified since treatment is randomised. Other fitted models (i.e. those for the outcome and CATEs) were obtained using the Super Learner (van der Laan et al., 2007), an ensemble learning method, implemented in the `SuperLearner` package in R. This used 20 cross validation folds, and a 'learner library' containing various routines (`glm`, `glmnet`, `gam`, `xgboost`, `ranger`). Additional results which use the 'discrete' Super Learner for model fitting are presented in Appendix D. The discrete Super Learner selects the regression algorithm in the learner library which minimises a cross validated estimate of e.g. the mean squared error loss, whereas the Super Learner minmizes the same loss by taking a convex combination of learners.

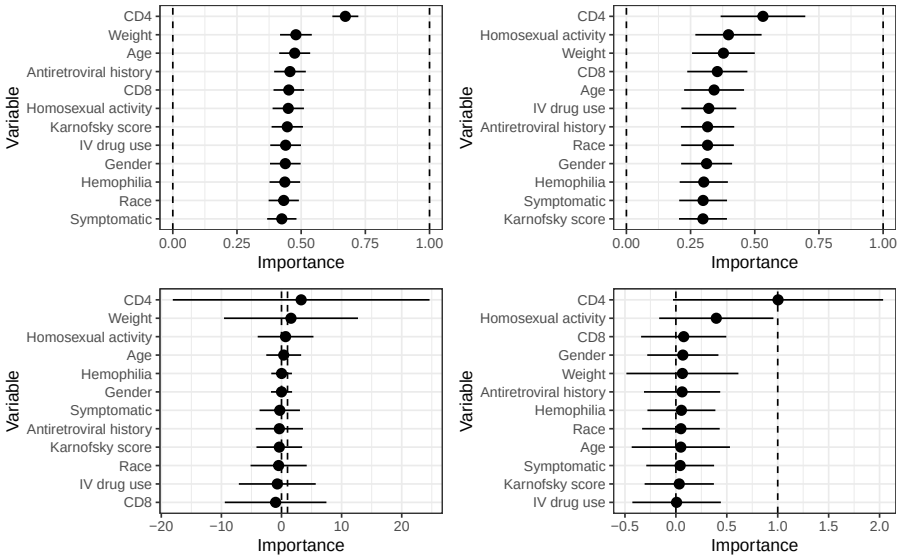


Figure 5.2: TE-VIM estimates from the ACTG175 study using the Super Learner for functional estimation. Top row: no sample splitting (Algorithm 1). Bottom row: with sample splitting (Algorithm 2). Left col: T-learner (A). Right col: DR-learner (B). Black lines indicate 95% Confidence intervals. In each plot, covariates are sorted according to their TE-VIM point estimate. Dashed lines indicate the $[0, 1]$ support of the TE-VIM.

The cross-fitted augmented inverse propensity weighted estimate of the ATE was $29.4mm^{-3}$ (CI: 14.8, 44.1; p<0.01) where all confidence-intervals (CIs) are reported at 95% significance and Wald-type p-values are reported. CIs for the VTE were constructed using both algorithms, and we report the square root of these intervals after truncating at zero. Algorithm 2A gave a root-VTE estimate of $6.44mm^{-3}$ (CI: 0,55.3; p=0.79) and Algorithm 2B gave a root-VTE estimate of $28.6mm^{-3}$ (CI: 0, 74.2; p=0.21). By comparison, the CV-TMLE estimator of Levy et al. (2021) gave an ATE estimate of $29.3mm^{-3}$ (CI: 14.7, 39.3; p<0.01) and a root-VTE estimate of $6.39mm^{-3}$ (CI: 0,39.3; p=0.70). The results from the CV-TMLE algorithm are more similar to those of Algorithm 2A than Algorithm 2B, since the CV-TMLE algorithm uses the T-learner to obtain initial CATE estimates. CV-TMLE is so-called since the initial propensity score and outcome regression estimates obtained in a 'cross validated' way.

These VTE estimates suggest that treatment effect homogeneity is a possible concern for our analysis, however, large treatment effect heterogeneity (relative to the ATE) cannot be ruled out. In practice, TE-VIMs are known to lie on the interval $[0, 1]$ therefore, we recommend truncating CIs at these values. For the purposes of comparing the algorithms, however, we have chosen not do so here. The results (Figure

5.2) suggest CD4 T-cell count at baseline is the most important variable in determining an individual's treatment effect, with weak evidence that any other single factor is a good predictor of CATE variability. Algorithms 2A and 2B respectively estimated the TE-VIM for CD4 T-cell count to be 3.26 (CI: -18.1, 24.6; p=0.76) and 1.00 (CI: -0.028, 2.03; p=0.057).

On balance, Algorithm B provides more credible TE-VIM estimates in this applied example than Algorithm A. This is evident in the extremely wide confidence intervals provided by Algorithm 1B in Figure 5.2. Algorithms A and B both use the same pseudo-outcome estimates, however the former makes better use of the propensity score model (i.e. the fact that this is a random trial), when estimating the CATE, whereas the latter estimates the CATE based on outcome regression alone. This distinction means that we expect the CATE estimates from Algorithm B to be robust to slow convergence of the outcome model with sample size, whereas those from Algorithm A are not. The cost of this robustness, however, is that Algorithm B relies on $n^{1/4}$ rate convergence of the pseudo-outcome regression, i.e. the final regression step of the DR-learner.

Additionally, Algorithm 2 (with sample splitting) provides more credible confidence intervals than Algorithm 1 (without sample splitting), effectively by using out-of-sample predictions to control for overfitting. This is evident from the narrow confidence intervals provided by Algorithms 1A and 1B in Figure 5.2, which suggest that all covariates account for a substantial amount of treatment effect heterogeneity. As with the simulation study, we recommend Algorithm 2B for TE-VIM inference.

5.5 Related work and extensions

Optimal treatment rule - VIMs

The proposed TE-VIMs complement VTEs and ATEs as an additional set of scalar summary estimands for the CATE. Fundamentally, these estimands summarise different aspects of the CATE function which are of scientific interest. Whilst TE-VIMs capture the importance of variable subsets in explaining the CATE, we remark that this should not be confused with the importance of those variables in contributing to optimal treatment decisions. For example, it is possible that a single covariate is important in explaining the magnitude of the CATE, but unimportant in determining the effect direction.

From the perspective of policy learning, one might instead be interested in the importance of variable subsets in explaining the optimal dynamic treatment rule (OTR), defined as $d(x) \equiv \mathbb{I}\{\tau(x) > 0\}$, where $\mathbb{I}(.)$ denotes an indicator function and we assume w.l.o.g. that a more positive outcome is preferred. The current approach might be extended by considering an OTR-VIM estimand,

$$\Gamma_s \equiv E[\text{var}\{d(X)|X_{-s}\}] = E[d_s(X)\{1 - d_s(X)\}]$$

where $d_s(x) \equiv E\{d(X)|X_{-s} = x_{-s}\} = Pr\{\tau(X) > 0|X_{-s} = x_{-s}\}$, and we note that $d(X) \in \{0, 1\}$ and $d_s(X) \in [0, 1]$. We argue that $\Gamma_s \in [0, 0.25]$ is analogous to Θ_s and θ_s, with the OTR $d(X)$ used in place of $\tau(X)$ and $\mu(X)$ respectively. Alternatively, Williamson et al. (2021b) propose an OTR-VIM based on the estimand,

$$\Gamma^*_s \equiv E\{\mu(d(X), X)\} - E\{\mu(d^*_s(X), X)\} = E\left[\tau(X)\{d(X) - d^*_s(X)\}\right]$$

where $d^*_s(x) \equiv \mathbb{I}\{\tau_s(x) > 0\}$ is the optimal treatment rule given covariates X_{-s}, and $\Gamma^*_s \geq 0$. Unlike the TE-VIM, both Γ_s and Γ^*_s are not pathwise differentiable, which complicates inference. Estimators for e.g. Γ^*_s, however, are typically based on efficient estimators derived in the setting where the OTR is a known function.

Treatment effect cumulative distribution function

Another related proposal considers the treatment effect cumulative distribution function (TE-CDF) (Levy and van der Laan, 2018), which is a curve $\beta : \mathbb{R} \mapsto [0, 1]$, with

$$\beta(t) = Pr\{\tau(X) \leq t\}.$$

Motivated by optimal treatments, the value $\beta(0)$ is of particular interest since it captures the marginal probability that an individual has a negative CATE, and therefore the proportion of the population which is not treated under the OTR. We note that $\beta(0)$ is not the same as $Pr(Y^1 - Y^0 \leq 0)$ which suffers similar identifiability issues regarding the joint distribution of (Y^1, Y^0) as the quantity $\text{var}(Y^1 - Y^0)$ mentioned previously. Like the OTR-VIMs above, the TE-CDF is generally not pathwise differentiable, hence Levy and van der Laan (2018) focus instead on a kernel smoothed analogue of $\beta(t)$. It is mentioned by Levy et al. (2021) that, provided $\tau_p > 0$, then Chebyshev's inequality implies

$$\beta(0) \leq \Lambda \equiv \frac{\Theta_p}{\tau_p^2}.$$

Thus, the VTE is also of scientific interest since it can be used to bound $\beta(0)$, informing investigators about the marginal probability of negative CATEs once a positive ATE has been established. Estimation of Λ could be carried out using estimating equations estimators, as in the current work, or targeted methods (Levy et al., 2021), using the IC for Λ,

$$\frac{\{\varphi(Z) - \tau_p\}^2 - \{\varphi(Z) - \tau(X)\}^2 - \Lambda\tau_p\{2\varphi(Z) - \tau_p\}}{\tau_p^2}.$$

Such estimators are beyond the scope of the current work, though we refer the interested reader to Appendix D for a sketch of the details for the former.

Continuous treatments

The idea of treating the CATE as a statistical functional that we would like to summarise enables similar VIM estimands to be defined in settings where one is interested in other statistical functionals. For instance, Hines et al. (2021a) propose an analogue of the CATE

$$\lambda(x) \equiv \frac{\text{cov}(A, Y|X = x)}{\text{var}(A|X = x)},$$

which is well defined even when A is a continuous exposure, and which identifies the CATE under standard causal assumptions (consistency, positivity, exchangeability) when A is binary, i.e. $\lambda(x) = \mu(1, x) - \mu(0, x)$. One might, therefore, extend the ATE, VTE, and TE-VIMs to continuous exposures by defining the estimands: $E\{\lambda(X)\}$, $\text{var}\{\lambda(X)\}$, and

$$\frac{E[\text{var}\{\lambda(X)|X_{-s}\}]}{\text{var}\{\lambda(X)\}},$$

which reduce to their CATE counterparts when A is binary. The ICs for these estimands are obtained by replacing the pseudo-outcome $\varphi(z)$ with

$$[y - \mu(x) - \lambda(x)\{a - \pi(x)\}] \frac{a - \pi(x)}{\text{var}(A|X = x)} + \lambda(x)$$

which reduces to $\varphi(z)$ when A is binary. See Appendix D for details.

5.6 Conclusion

We propose TE-VIMs, which extend the VIM framework of Williamson et al. (2021b) to include the CATE as a functional of interest. These have immediate applications to the analysis of observational and clinical trial data, and provide insight into scientific questions related to treatment effect heterogeneity. Our methods complement VTE analyses, which quantify treatment effect heterogeneity (Levy et al., 2021). We derive efficient estimating equation estimators which are amenable to data-adaptive estimation of

working models. These are broadly applicable, since they are not tied to a particular model or regression algorithm, unlike existing proposals based on causal random forests (Athey et al., 2019).

We elucidate links between our estimators and regression-VIM counterparts, by interpreting our estimators in terms of pseudo-outcomes (Kennedy, 2020). We believe pseudo-outcome based approaches might generalise to other statistical functionals, where analogous pseudo-outcomes could be derived. For instance, 'derivative effect VIMs' could be derived which are well defined for continuous exposures and which identify the proposed TE-VIMs when the exposure is binary (Hines et al., 2021a), or VIMs for policy learning could be developed using double robust scores (Athey and Wager, 2021).

We recommend that TE-VIM inference be incorporated into a more broad treatment effect analysis, where primary interest is in inferring the ATE and VTE. We believe that VTE inference should form part of a primary analysis, since the ATE and VTE may be used to bound the marginal probability of adverse CATEs, and since it is possible that the population ATE is zero, but some (or indeed all) individuals experience a large CATE. One may then infer TE-VIMs as part of a secondary analysis, when large treatment effect heterogeneity cannot be ruled out, since TE-VIM estimands are not of scientific interest when there is little variability in the CATE to account for.

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT

T: +44 (0)20 7299 4646
F: +44 (0)20 7299 4656
www.lshtm.ac.uk

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	1804954	**Title**	Mr.
First Name(s)	Oliver		
Surname/Family Name	Hines		
Thesis Title	Assumption-Lean Inference for Causal and Statistical Questions in the Era of Machine Learning		
Primary Supervisor	Karla Diaz-Ordaz		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?			
When was the work published?			
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Choose an item.	Was the work subject to academic peer review?	Choose an item.

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	Biometrika
Please list the paper's authors in the intended authorship order:	Oliver Hines, Jonathan Levy, Karla Diaz-Ordaz, Stijn Vansteelandt, Mark van der Laan
Stage of publication	**Not yet submitted**

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	This study was conceived by all authors. I carried out the mathematical research, computational simulations under the supervision of Diaz-Ordaz, Vansteelandt, and van der Laan. I wrote the first draft of the manuscript, with comments and editing from the other authors.

SECTION E

Student Signature	Ohines
Date	14 December 2022

Supervisor Signature	Karla Diaz-Ordaz
Date	14 December 2022

Chapter 6

Nonparametric score testing

6.1 Introduction

Over recent decades, a theory of nonparametric inference has developed which allows regular asymptotically linear (RAL) estimators to be constructed for nonparametrically defined estimands, even when data adaptive/ machine learning methods are used to fit nuisance functionals. Estimation of the average treatment effect (ATE) of a binary treatment on outcome represents a canonical example in which data adaptive/ machine learning methods can be used to estimate e.g. the conditional mean outcome or the propensity score given covariates. Provided that the target estimand is 'pathwise differentiable', there are two prevailing strategies for constructing efficient point estimators: one-step bias correction estimators (also called 'double/ debiased machine learning' estimators) (Chernozhukov et al., 2018) and 'targeted learning' estimators (TMLE) (van der Laan and Rubin, 2006). Both strategies produce point estimators that are RAL and hence consistent and asymptotically normal with a variance characterised by the efficient influence curve (IC) of the target estimand (also called the influence function or pathwise derivative).

Asymptotic normality of the estimator usually forms the basis of subsequent inference, with Wald confidence intervals (CIs) and hypothesis tests constructed using an estimate of the variance of the point estimator. Like in the parametric setting, nonparametric Wald CIs have several limitations. Firstly, they are generally not invariant to 'differentiable reparametrisations' of the estimand. For example, consider the variance of treatment effect (VTE) estimand, which has recently been proposed as a (non-negative) global measure of treatment effect heterogeneity (Levy et al., 2021). In general, the Wald CI obtained by treating the VTE as the target estimand is different to the CI obtained by squaring the values in a Wald CI for the square root of the VTE, even though both intervals represent a CI for the VTE. This is problematic since it can lead to conclusions which depend on the scale on which the target estimand is defined.

Secondly, Wald CIs require knowledge of the IC up to constants of proportionality that are the same for all observations. The population quantile of a continuous random variable is a canonical example where this is the case. It is well documented, for example, that estimation of the variance of the population quantile estimator requires estimation of the probability density at the quantile of interest (Mosteller, 1946).

Finally, an appealing property of TMLE estimators, compared with one-step bias correction estimators, is that they deliver 'plug-in' point estimates that respect the constraints of the estimand and which are invariant to differentiable reparameterisations of the target estimand. For instance, a TMLE estimator of the VTE is guaranteed to be non-negative (Levy et al., 2021), and the square root of this VTE estimator provides a RAL estimator for the root-VTE. This is not the case for one-step bias correction estimators of the VTE (Chapter 5). Whilst TMLE point estimators necessarily represent plug-in values, the same cannot be said of the values contained in the subsequent Wald CIs, even when centred on a TMLE plug-in estimator. This means that, in finite samples, Wald CIs can include values outside the parameter space e.g. negative VTE estimates or probability estimates outside $[0, 1]$ etc..

The TMLE strategy is so-called since it treats the IC as if it were the parametric score function in a

likelihood model. TMLE estimators essentially construct and then maximise a parametric quasi-likelihood function with the IC as its score. Interpreting the IC as if it were a score function also forms the basis of our approach, where we formally extend score testing results to nonparametric estimands, and propose a CI estimation strategy which inverts the proposed nonparametric score tests.

Our approach builds on related hypothesis testing frameworks from the generalised methods of moments (GMM) literature (Dufour et al., 2017; Newey and Smith, 2004; Hansen et al., 1996; Newey and West, 1987), where a set of moment conditions is used to perform inference. In the parametric model setting, these moment conditions usually correspond to derivatives of the log-likelihood function (i.e score functions) and classical parametric Wald, score, and likelihood ratio tests are obtained. In the nonparametric setting, we propose deriving similar hypothesis tests by treating the IC itself as a moment condition, i.e. a function of the data distribution which is known to have mean zero. A brief review of the relevant GMM literature can be found in Appendix E.3.

The remainder of the paper will be structured as follows. In Section 6.2 we review nonparametric Wald confidence interval estimation. We show how one can obtain requisite point estimators and variance estimators using so-called one-step bias correction and TMLE. In Section 6.3 we propose a new score test statistic and demonstrate, through worked examples, how this alleviates concerns regarding the fact that Wald intervals are not invariant to reparameterisations of the estimand under differentiable mappings. In Section 6.4 we propose a novel interval estimation procedure, for 'complicated settings', i.e. those where the statistic depends on infinite dimensional parameters (functions) of the unknown data generating mechanism. Our proposal makes use of parametric submodel results, similar to those found in the TMLE literature. Finally in Section 6.5 a simulation study is carried out.

6.2 Preliminaries

6.2.1 Wald type confidence sets

Suppose we have n iid observations $z_1, ..., z_n$ of $Z \sim P_0$ which follows an unknown distribution $P_0 \in \mathcal{M}$, where $\mathcal{M}$ denotes the nonparametric model. We consider a d-dimensional estimand $\boldsymbol{\Psi} : \mathcal{M} \mapsto \mathbb{R}^d$, with associated IC, $\boldsymbol{\phi}(Z, P_0)$, which is a mean-zero statistical functional, i.e. $P_0\{\boldsymbol{\phi}(Z, P_0)\} = 0$, where $P_0\{.\}$ denotes expectation under P_0. An estimator $\hat{\boldsymbol{\Psi}}$ of $\boldsymbol{\Psi}(P_0)$ is said to be regular asymptotically linear (RAL) if

$$\begin{aligned} \hat{\boldsymbol{\Psi}} - \boldsymbol{\Psi}(P_0) &= \boldsymbol{U}_n(P_0) + \boldsymbol{r}_n \\ \boldsymbol{U}_n(P) &\equiv n^{-1} \sum_{i=1}^{n} \boldsymbol{\phi}(z_i, P) \end{aligned} \tag{6.1}$$

where $\boldsymbol{r}_n = o_P(n^{-1/2})$ is a remainder term and $\boldsymbol{U}_n : \mathcal{M} \mapsto \mathbb{R}^d$ is defined for any $P \in \mathcal{M}$. In Section 6.2.2 we describe how RAL estimators can be constructed from an initial distribution estimator $\hat{P}_n$ of P_0, though for our discussion of Wald type confidence sets, the exact form of $\hat{\boldsymbol{\Psi}}$ is not important. Letting $I_0 \equiv P_0\left\{\boldsymbol{\phi}(Z, P_0)\boldsymbol{\phi}^\top(Z, P_0)\right\}$, it follows from the central limit theorem that, as $n \to \infty$

$$\begin{aligned} \sqrt{n}\boldsymbol{U}_n(P_0) &\xrightarrow{d} \mathcal{N}(0, I_0) \\ \implies \sqrt{n}\{\hat{\boldsymbol{\Psi}} - \boldsymbol{\Psi}(P_0)\} &\xrightarrow{d} \mathcal{N}(0, I_0) \end{aligned}$$

where $\mathcal{N}(\mu, \Sigma)$ denotes a multivariate normal distribution with mean μ and covariance matrix Σ. When I_0 is non-singular, this result can be expressed through the quadratic form

$$n\{\hat{\boldsymbol{\Psi}} - \boldsymbol{\Psi}(P_0)\}^\top I_0^{-1}\{\hat{\boldsymbol{\Psi}} - \boldsymbol{\Psi}(P_0)\} \xrightarrow{d} \chi^2_d \tag{6.2}$$

where χ^2_d denotes a χ^2 distribution with d degrees of freedom. Supposing a consistent non-singular estimate of the variance $\hat{I}_n = I_0 + o_P(1)$ the left hand side above is estimated by the 'Wald statistic'

$$n\hat{W}_n(\boldsymbol{\Psi}(P_0)) \equiv n\{\hat{\boldsymbol{\Psi}} - \boldsymbol{\Psi}(P_0)\}^\top \hat{I}_n^{-1}\{\hat{\boldsymbol{\Psi}} - \boldsymbol{\Psi}(P_0)\} \xrightarrow{d} \chi^2_d.$$

This result forms the basis of nonparametric Wald tests for null hypotheses of the form $H_0 : \boldsymbol{\Psi}(P_0) = \boldsymbol{\psi}_0$. The Wald test for H_0 rejects at significance level α if

$$n\hat{W}_n(\boldsymbol{\psi}_0) > c_\alpha^2$$

where c_α^2 denotes the $1-\alpha$ quantile of the χ_d^2 distribution. A Wald confidence set at significance level α is defined as the set of $\boldsymbol{\psi}_0$ values which do not satisfy the inequality above, i.e. the set of estimand values which cannot be rejected at level α by the Wald test. For $d = 1$ this results in the familiar Wald CI

$$\hat{\Psi} \pm c_\alpha \sqrt{\frac{\hat{I}_n}{n}} \tag{6.3}$$

where $a \pm b$ denotes the interval $(a-b, a+b)$.

6.2.2 One-step bias correction and TMLE point estimators

A common strategy for constructing RAL estimators is to consider the so-called plug-in estimator $\boldsymbol{\Psi}(\hat{P}_n)$, where $\hat{P}_n$ represents an initial estimate of P_0 obtained from the observed data (Hines et al., 2022; Kennedy, 2022). Consider the decomposition

$$\boldsymbol{\Psi}(\hat{P}_n) - \boldsymbol{\Psi}(P_0) = \boldsymbol{U}_n(P_0) - \boldsymbol{U}_n(\hat{P}_n) + \boldsymbol{r}_n \tag{6.4}$$

which holds in general if we make no assumptions about the remainder term $\boldsymbol{r}_n$. This decomposition is the so-called Von Mises expansion of the estimand and usually it can be shown that $\boldsymbol{r}_n = o_p(n^{-1/2})$ when certain statistical functionals of $\hat{P}_n$ converge (in the large sample limit) to their P_0 counterparts at sufficiently fast rates. Additionally it is often assumed that some statistical functionals of $\hat{P}_n$ and P_0 are bounded and that $\boldsymbol{\phi}(z, \hat{P}_n)$ forms a Donsker class. Remainder terms of the type in (6.4) are usually analysed on a case-by-case basis and we will assume that requisite assumptions on $\hat{P}_n$ are satisfied such that $\boldsymbol{r}_n = o_p(n^{-1/2})$.

The expansion in (6.4) is very nearly that of the RAL estimator in (6.1), except for the term $\boldsymbol{U}_n(\hat{P}_n)$, which we call the 'plug-in bias'. The plug-in bias is generally non-zero and generally not $o_p(n^{-1/2})$. Fundamentally there are two popular strategies to account for plug-in bias and hence derive RAL estimators for $\Psi(P_0)$.

The first strategy, known as one-step bias correction or double/ debiased machine learning (Chernozhukov et al., 2018), considers debiasing the plug-in estimator in the estimand space. In particular, (6.4) implies that the estimator $\hat{\boldsymbol{\Psi}} = \boldsymbol{\Psi}(\hat{P}_n) + \boldsymbol{U}_n(\hat{P}_n)$ is RAL. Since the IC represents a pathwise derivative, the one-step estimation strategy can be thought of as performing a single Newton-Raphson update to the initial plug-in estimator. This strategy is popular due to its simplicity once the IC has been derived, and once it has been shown that $\boldsymbol{r}_n = o_p(n^{-1/2})$.

The second strategy, TMLE (van der Laan and Rubin, 2006) considers debiasing the plug-in estimator in the distribution space $\mathcal{M}$ rather than the estimand space. Specifically, the goal of TMLE is to obtain a new estimate $\hat{P}_n^*$ of P_0 which is, in some sense, 'close' to the initial estimate $\hat{P}_n$ and is chosen such that $\boldsymbol{U}_n(\hat{P}_n^*) = 0$. Some TMLE estimators actually set $\boldsymbol{U}_n(\hat{P}_n^*) = o_p(n^{-1/2})$, though we will mostly ignore this detail. In Section 6.4.1 we describe in more detail how $\hat{P}_n^*$ is obtained with reference to a parametric submodel containing $\hat{P}_n$. By (6.4), the resulting TMLE estimator $\boldsymbol{\Psi}(\hat{P}_n^*)$ represents a RAL plug-in estimator.

TMLE estimators immediately provide an estimator $I_n(\hat{P}_n^*)$ for the covariance I_0 where

$$I_n(P) \equiv n^{-1} \sum_{i=1}^{n} \boldsymbol{\phi}(z_i, P)\boldsymbol{\phi}^\top(z_i, P). \tag{6.5}$$

In contrast, the variance estimator associated with one-step estimators must also account for plug-in bias, e.g. by using the 'mean corrected' covariance estimator $\tilde{I}_n(\hat{P}_n)$ of I_0 where

$$\tilde{I}_n(P) \equiv n^{-1} \sum_{i=1}^{n} \{\boldsymbol{\phi}(z_i, P) - \boldsymbol{U}_n(P)\}\{\boldsymbol{\phi}(z_i, P) - \boldsymbol{U}_n(P)\}^\top \tag{6.6}$$

or else by using a non-plug-in estimator for $\boldsymbol{\phi}(z, P_0)$. Note that $I_n(P)$ and $\tilde{I}_n(P)$ are positive definite by construction, provided that none of the components of $\boldsymbol{\phi}(z_i, P)$ are zero (or for the mean corrected variance estimator, constant) for all i, or linearly dependant. Additionally, $\tilde{I}_n(P) = I_n(P) - \boldsymbol{U}_n(P)\boldsymbol{U}_n^\top(P)$ by definition.

6.2.3 Invariance to reparametrization of point estimates

Consider a differentiable function $h : \mathbb{R}^d \mapsto \mathbb{R}^q$, with $q \leq d$, and suppose that our goal is to find a point estimate for $h\{\boldsymbol{\Psi}(P_0)\}$. Since the IC is a pathwise derivative, it follows by the chain rule that the IC of $P \mapsto h\{\boldsymbol{\Psi}(P)\}$ at P_0 is

$$\nabla h\{\boldsymbol{\Psi}(P_0)\}\boldsymbol{\phi}(Z, P_0)$$

where $\nabla h(.)$ denotes the Jacobian of $h(.)$. Following the strategies in the previous Section, we presume an initial estimate $\hat{P}_n$ of P_0 has been obtained, and that the remainder term in (6.4) is $o_p(n^{-1/2})$. The one-step bias corrected estimate of $h\{\boldsymbol{\Psi}(P_0)\}$ is

$$h\{\boldsymbol{\Psi}(\hat{P}_n)\} + \nabla h\{\boldsymbol{\Psi}(\hat{P}_n)\}\boldsymbol{U}_n(\hat{P}_n) = h\{\boldsymbol{\Psi}(\hat{P}_n)\} + \nabla h\{\boldsymbol{\Psi}(\hat{P}_n)\}\{\hat{\boldsymbol{\Psi}} - \boldsymbol{\Psi}(\hat{P}_n)\}$$

where $\hat{\boldsymbol{\Psi}} = \boldsymbol{\Psi}(\hat{P}_n) + \boldsymbol{U}_n(\hat{P}_n)$ is the one-step bias correction estimate of $\boldsymbol{\Psi}(P_0)$. We interpret the right hand side above as a first-order Taylor approximation of $h(\hat{\boldsymbol{\Psi}})$, i.e. the estimate of $h\{\boldsymbol{\Psi}(P_0)\}$ obtained by first estimating $\boldsymbol{\Psi}(P_0)$ by $\hat{\boldsymbol{\Psi}}$ then applying $h(.)$. Unfortunately, evaluating $h(\hat{\boldsymbol{\Psi}})$ is not always feasible since there is no general guarantee that $\hat{\boldsymbol{\Psi}}$ lies in the domain of $h(.)$, for instance if $h(u) = \log(u)$ is the logarithmic function and $\hat{\boldsymbol{\Psi}}$ is a negative scalar, i.e. $\boldsymbol{\Psi}(\hat{P}_n) < -\boldsymbol{U}_n(\hat{P}_n)$. This non-invariance is problematic since it generally leads to conclusions that depend on the scale on which the estimand is defined, though this problem diminishes with sample size.

Conversely, the TMLE estimator is invariant to differentiable reparameterisations of the estimand. To see why, note that if a targeted distribution $\hat{P}_n^*$ has been obtained such that $\boldsymbol{U}_n(\hat{P}_n^*) = o_p(n^{-1/2})$, then

$$\nabla h\{\boldsymbol{\Psi}(\hat{P}_n^*)\}\boldsymbol{U}_n(\hat{P}_n^*) = o_p(n^{-1/2})$$

and $h\{\boldsymbol{\Psi}(\hat{P}_n^*)\}$ is an RAL estimator of $h\{\boldsymbol{\Psi}(P_0)\}$. Intuitively, the invariance of the TMLE estimator arises since the debiasing of the plug-in estimator $\boldsymbol{\Psi}(\hat{P}_n)$ occurs in the distribution space rather than the estimand space, and hence the targeted distribution estimator $\hat{P}_n^*$ inadvertently debiases the initial distribution estimator $\hat{P}_n$ for any estimand that has an IC $\propto \boldsymbol{\phi}(Z, P_0)$. By the linearity of the pathwise derivative, this class includes estimands of the form $h\{\boldsymbol{\Psi}(P_0)\}$.

In Appendix E.1 we show how Wald CIs are not invariant to differentiable reparameterisations of the type $h(.)$ above, even when centred on a TMLE point estimator that is invariant to such reparameterisations.

6.3 Score intervals

6.3.1 Score statistic proposal

Here we propose non-parametric score-type confidence sets, so called because they rely on test statistics similar to those used in score tests from likelihood based inference. Following the discussion on Wald intervals in Section 6.2.1, we consider a Wald statistic where the initial estimator $\hat{\boldsymbol{\Psi}}$ is replaced with the asymptotically equivalent 'estimator' $\hat{\boldsymbol{\Psi}}^* \equiv \hat{\boldsymbol{\Psi}} - \boldsymbol{r}_n$. Whilst the $\hat{\boldsymbol{\Psi}}$ estimator is RAL in the sense that the remainder term $\boldsymbol{r}_n$ vanishes asymptotically (at faster than 1 over root-n rate), $\hat{\boldsymbol{\Psi}}^*$ is exactly linear in the sense that there is no remainder term at all.

It is, however, not possible to evaluate $\hat{\boldsymbol{\Psi}}^*$ since $\boldsymbol{r}_n$, cannot be estimated without knowing the true distribution P_0. With this caveat, we consider a Wald-type statistic centred on $\hat{\boldsymbol{\Psi}}^*$

$$\{\hat{\boldsymbol{\Psi}}^* - \boldsymbol{\Psi}(P_0)\}^\top \hat{I}_n^{-1}\{\hat{\boldsymbol{\Psi}}^* - \boldsymbol{\Psi}(P_0)\} = \boldsymbol{U}_n^\top(P_0)\hat{I}_n^{-1}\boldsymbol{U}_n(P_0)$$

Next we suppose that one could estimate the covariance I_0 with the linear estimator $\hat{I}_n = I_n(P_0)$. Like $\hat{\boldsymbol{\Psi}}^*$ the estimator $I_n(P_0)$ can only be evaluated when P_0 is known. This procedure results in a the statistic $nM_n(P_0)$, where we define the 'score statistic'

$$M_n(P) \equiv \boldsymbol{U}_n^\top(P) I_n^{-1}(P) \boldsymbol{U}_n(P). \tag{6.7}$$

with $nM_n(P_0) \xrightarrow{d} \chi^2_d$. We therefore interpret $M_n(P_0)$ as a Wald statistic that uses unobtainable linear estimators for $\Psi(P_0)$ and I_0. Compared with the Wald statistic $\hat{W}_n(\boldsymbol{\Psi}(P_0))$ previously, the score statistic $M_n(P_0)$ can be function of P_0 through any arbitrary statistical functionals, not just the target estimand $\boldsymbol{\Psi}(P_0)$. This makes testing the null hypothesis $\boldsymbol{\Psi}(P_0) = \boldsymbol{\psi}_0$, and constructing the implied confidence for $\boldsymbol{\Psi}(P_0)$, generally non-trivial. Instead, we generally propose using the score statistic to construct a confidence set in distribution space $\hat{\mathcal{M}} \subseteq \mathcal{M}$, which implies a confidence set for $\boldsymbol{\Psi}(P_0)$ in the estimand space. In particular, we reject distributions for which the score statistic exceeds the threshold c^2_α/n, though in Appendix E.2 we discuss alternative, but asymptotically equivalent, score statistic thresholds. The validity of this approach is justified by the Theorem 5 below.

Theorem 5 (Confidence Set Mapping) *Let $M(P)$ be a statistic such that $Pr\{M(P_0) \leq c^2_\alpha/n\} \geq 1-\alpha$. Let $\hat{\mathcal{M}} \equiv \left\{P \in \mathcal{M} | M(P) \leq c^2_\alpha/n\right\}$ denote a confidence set over $\mathcal{M}$ then,*

$$Pr\left[\boldsymbol{\Psi}(P_0) \in \boldsymbol{\Psi}[\hat{\mathcal{M}}]\right] \geq 1-\alpha$$

where $\boldsymbol{\Psi}[.]$ denotes the image under $\boldsymbol{\Psi}(.)$.

Proof 1 *Letting $\boldsymbol{\Psi}^{-1}[A]$ denote the preimage of A under $\boldsymbol{\Psi}(.)$, i.e. the set of distributions which map to estimand values in A, then*

$$Pr\{\boldsymbol{\Psi}(P_0) \in \boldsymbol{\Psi}[\hat{\mathcal{M}}]\} = Pr\{P_0 \in \boldsymbol{\Psi}^{-1}[\boldsymbol{\Psi}[\hat{\mathcal{M}}]]\}$$

By definition, $\hat{\mathcal{M}} \subseteq \boldsymbol{\Psi}^{-1}[\boldsymbol{\Psi}[\hat{\mathcal{M}}]]$ hence,

$$Pr\{P_0 \in \boldsymbol{\Psi}^{-1}[\boldsymbol{\Psi}[\hat{\mathcal{M}}]]\} = Pr(P_0 \in \hat{\mathcal{M}}) + Pr\{P_0 \in \boldsymbol{\Psi}^{-1}[\boldsymbol{\Psi}[\hat{\mathcal{M}}]] \setminus \hat{\mathcal{M}}\} \geq 1-\alpha.$$

where we have used the fact that $Pr(P_0 \in \hat{\mathcal{M}}) \geq 1-\alpha$

Theorem 5 is significant as it justifies using the score statistic $M_n(P)$ to construct confidence sets in the space of distributions, and that these sets necessarily imply valid confidence sets in the space of estimand values.

6.3.2 Invariance to reparameterisation of the score statistic

The score statistic $M_n(P_0)$ is particularly appealing since it is invariant to smooth reparameterisations of the estimand. To see why, consider replacing $\boldsymbol{\phi}(Z,P)$ with $\tilde{\boldsymbol{\phi}}(Z,P) = J\boldsymbol{\phi}(Z,P)$, where J is a q by d matrix with $\text{Rank}(J) = q$ and $q \leq d$. Under this transformation

$$\begin{aligned}\tilde{\boldsymbol{U}}_n(P) &= J\boldsymbol{U}_n(P) \\ \tilde{I}_n(P) &= JI_n(P)J^\top\end{aligned}$$

Hence,

$$\begin{aligned}\tilde{M}_n(P) &= \tilde{\boldsymbol{U}}_n^\top(P)\tilde{I}_n^{-1}(P)\tilde{\boldsymbol{U}}_n(P) \\ &= \boldsymbol{U}_n^\top(P)J^\top(JI_n(P)J^\top)^{-1}J\boldsymbol{U}_n(P) \\ &= \boldsymbol{U}_n^\top(P)B_n(P)I_n^{-1}(P)\boldsymbol{U}_n(P)\end{aligned}$$

where $B_n(P) = J^\top (JI_n(P)J^\top)^{-1} JI_n(P)$ is a d by d projection matrix with $\text{Rank}\{B_n(P)\} = \text{Trace}\{B_n(P)\} = q$. It follows that $\tilde{M}_n(P_0) \xrightarrow{d} \chi^2_q$, and, when $q = d$ then $B_n(P)I_n^{-1}(P) = I_n^{-1}(P)$ and hence $\tilde{M}_n(P) = M_n(P)$.

This invariance is useful, since it implies that the IC $\boldsymbol{\phi}(Z, P)$ must be known only up to constants of proportionality, and hence $M_n(P)$ is invariant under differentiable reparameterisations of the type $\tilde{\boldsymbol{\Psi}}(P_0) = h\{\boldsymbol{\Psi}(P_0)\}$ where $h : \mathbb{R}^d \mapsto \mathbb{R}^d$ is a known differentiable function with non-singular Jacobian $J(P) = \nabla h\{\boldsymbol{\Psi}(P)\}$. This is the case since the IC of $\tilde{\boldsymbol{\Psi}}(P_0)$ is $J(P_0)\boldsymbol{\phi}(Z, P_0)$ and $J(P_0)$ can be treated as a constant of proportionality.

The following two examples make use of this property to derive CIs for the simple weighted mean and for the population quantile. For the latter, our proposed CIs do not involve density estimation as is usually required for estimating the standard error in the quantile.

Example 1: Weighted mean

Consider that $Z = (Y, W)$ consists of an outcome $Y \in \mathbb{R}$ and a known weight $W \in \mathbb{R}$. Define the target estimand as the weighted mean $\Psi(P_0) = P_0(WY)/P_0(W)$, where it is assumed that $P_0(W) \neq 0$. This estimand has IC

$$\phi(Z, P_0) = \frac{YW - \Psi(P_0)W}{P_0(W)} \propto YW - \Psi(P_0)W$$

In constructing the score statistic we note that $P_0(W)$ is a non-zero constant of proportionality which could be discarded, however for the purposes of illustration we include it in our analysis here. The IC above implies

$$U_n(P) = \frac{\theta_n}{P\{W\}}\{\psi_n - \Psi(P)\}$$
$$I_n(P) = \left(\frac{\theta_n}{P\{W\}}\right)^2 \left[\sigma_n^2 + \beta_n\{\psi_n - \Psi(P)\}^2 + 2\gamma_n\{\psi_n - \Psi(P)\}\right]$$

where we define the summary statistics

$$\theta_n \equiv n^{-1}\sum_{i=1}^n w_i, \qquad \psi_n \equiv \theta_n^{-1} n^{-1}\sum_{i=1}^n y_i w_i,$$
$$\sigma_n^2 \equiv \theta_n^{-2} n^{-1}\sum_{i=1}^n (y_i w_i - \psi_n w_i)^2, \qquad \beta_n \equiv \theta_n^{-2} n^{-1}\sum_{i=1}^n w_i^2,$$
$$\gamma_n \equiv \theta_n^{-2} n^{-1}\sum_{i=1}^n (y_i w_i - \psi_n w_i) w_i.$$

It follows that $M_n(P) \leq c_\alpha^2/n$ becomes

$$M_n(P) = \frac{\{\psi_n - \Psi(P)\}^2}{\sigma_n^2 + \beta_n\{\psi_n - \Psi(P)\}^2 + 2\gamma_n\{\psi_n - \Psi(P)\}} \leq \frac{c_\alpha^2}{n}$$

For this simple nonparametric estimand, the statistic $M_n(P)$ depends on P only through the target estimand $\Psi(P)$. It is therefore easy to solve the inequality above to construct a CI for $\Psi(P_0)$. Specifically, the inequality holds for all $\Psi(P)$ in the interval $\psi_n + a_n \pm b_n$ where

$$a_n = \frac{c_\alpha^2 \gamma_n}{n - c_\alpha^2 \beta_n}$$
$$b_n = \sqrt{\frac{c_\alpha^2 \sigma_n^2}{n - c_\alpha^2 \beta_n} + a_n^2}$$

We see that for $c_\alpha = 0$ we obtain the point estimate ψ_n that would be expected. The term a_n, however, acts as an $O(n^{-1})$ term that shifts the centre of the CI away from ψ_n. The term b_n is of the form

$$b_n = \sqrt{\frac{c_\alpha^2 \sigma_n^2}{n}} + o_P(n^{-1/2})$$

hence to order $o_P(n^{-1/2})$ we obtain the standard Wald interval $\psi_n \pm c_\alpha \sigma_n n^{-1/2}$.

Example 2: Population quantile

Let $Z = Y \in \mathbb{R}$ be a continuously distributed random variable. Let $\Psi(P_0)$ be the τth quantile of Y, which has IC,

$$\phi(Y, \Psi) = \frac{\tau - \{1 - \Theta(Y - \Psi(P_0))\}}{p_0(\Psi(P_0))} \propto \tau - \{1 - \Theta(Y - \Psi(P_0))\}$$

where we define the step function $\Theta(u) = 1$ for $u \geq 0$ and 0 otherwise, and $p_0(y)$ is the marginal density of Y under P_0. Constructing Wald CI for $\Psi(P_0)$ is difficult since one is required to estimate the density at the point estimate, $p_0(\hat{\Psi})$. The density, however, appears as a proportionality constant in the IC, hence can be disregarded in the score statistic. Discarding this factor gives

$$\begin{aligned} U_n(P) &= \tau - F_n(P) \\ I_n(P) &= \tau^2 + (1 - 2\tau) F_n(P) \end{aligned}$$

where $F_n(P) = n^{-1} \sum_{i=1}^n 1 - \Theta(y_i - \Psi(P))$ represents the empirical CDF at $\Psi(P)$. It follows that

$$M_n(P) = \frac{\{\tau - F_n(P)\}^2}{\tau^2 + (1 - 2\tau) F_n(P)} \leq \frac{c_\alpha^2}{n}$$

which is an inequality that depends on P only through $\Psi(P)$. This inequality is satisfied by all $\Psi(P)$ such that $F_n(P)$ is in the interval

$$F_n(P) \in \tau - \frac{c_\alpha^2}{n}\left(\tau - \frac{1}{2}\right) \pm \frac{c_\alpha}{\sqrt{n}} \sqrt{\tau(1-\tau) + \frac{c_\alpha^2}{n}\left(\tau - \frac{1}{2}\right)^2}$$

To illustrate this result we consider setting $\tau = 1/2$, in which case $\Psi(P_0)$ is the population median and the interval for $F_n(P)$ above reduces to

$$F_n(P) \in \frac{1}{2} \pm \frac{c_\alpha}{2\sqrt{n}}$$

We argue that this is asymptotically equivalent to the Wald CI for the median since, letting $Q_n(\tau)$ denote the τth empirical quantile and $\hat{p}(y)$ denote an estimate of $p_0(y)$, we use the hand-waving argument that

$$\begin{aligned} \Psi(P) &\in Q_n\left(\frac{1}{2} \pm \frac{c_\alpha}{2\sqrt{n}}\right) \\ &\approx Q_n\left(\frac{1}{2}\right) \pm \frac{c_\alpha}{2\sqrt{n}} \frac{dQ_n(\tau)}{d\tau}\Big|_{\tau=1/2} \\ &\approx \hat{\Psi} \pm \frac{c_\alpha}{\sqrt{n}} \frac{1}{2\hat{p}(\hat{\Psi})}. \end{aligned}$$

This argument is 'hand-waving' since the empirical quantile function is not differentiable. The score based CIs derived in this example are similar to existing nonparametric quantile CI estimators e.g. (Frey and Zhang, 2017; Hutson, 1999; Lanke, 1974), in the sense that both methods deliver CIs which are bounded by a pair of empirical quantiles. Existing methods, however, appeal to a theory of so-called fractional order statistics (Stigler, 1977) to derive the boundary quantiles of interest, whereas we frame the problem using score tests.

6.4 Complicated estimands

Examples 1 and 2 above can be thought of as 'simple' nonparametric estimands in the sense that the score statistics $M_n(P)$ associated with each of these estimands is a function of P only through the estimand $\mathbf{\Psi}(P)$. This is typically not the case as we illustrate in the next example, where we construct the score statistic associated with the ATE.

Example 3: Average treatment effect (ATE)

Let $Z = (Y, A, X)$ consist of an outcome $Y \in [0, 1]$, an exposure $A \in \{0, 1\}$, and a vector of covariates $X \in \mathbb{R}^p$. Generally the ATE is defined in the same way for $Y \in \mathbb{R}$, however for the latter we can scale the outcome to be in $[0, 1]$ via the transformation $Y \mapsto (Y - a)/(b - a)$ where a and b are minimum and maximum outcomes respectively, obtained from the data or known a priori. Letting Y_a denote the outcome that would have been observed if one sets $A = a$, then the ATE is defined as $E(Y_1 - Y_0)$. Under standard causal assumptions of consistency ($A = a \implies Y = Y^a$), conditional exchangeability ($Y^a \perp\!\!\!\perp A|X$ for $a = 0, 1$), and positivity ($0 < \pi(X) < 1$ w.p.1), the ATE is identified by $\Psi(P_0) \equiv P_0\{\mu_0(1, X) - \mu_0(0, X)\}$, where $\mu_0(a, x) \equiv E_{P_0}(Y|A = a, X = x)$. The ATE has IC,

$$\phi(Z, P_0) = \mu_0(1, X) - \mu_0(0, X) - \Psi(P_0) + \{Y - \mu_0(A, X)\}\beta_0(A, X)$$

where, letting $\pi_0(x) \equiv E_{P_0}(A|X = x)$, we define,

$$\beta_0(A, X) \equiv \frac{A - \pi_0(X)}{\pi_0(X)\{1 - \pi_0(X)\}}.$$

This IC is more complicated than those in Examples 2 and 3 since it depends on P_0 through $\Psi(P_0)$ and through the unknown statistical functionals $\mu_0(1, X), \mu_0(0, X)$ and $\pi_0(X)$. It follows that for some distribution P

$$\begin{aligned} U_n(P) &= \psi_n(P) - \Psi(P) + \theta_n(P) \\ I_n(P) &= \sigma_n^2(P) + \{\psi_n(P) - \Psi(P)\}^2 + 2\theta_n(P)\{\psi_n(P) - \Psi(P)\} \end{aligned}$$

where, letting $\mu(a, x)$ and $\beta(a, x)$ denote analogues of $\mu_0(a, x)$ and $\beta_0(a, x)$ under P, we define

$$\begin{aligned} \psi_n(P) &= n^{-1} \sum_{i=1}^{n} \mu(1, x_i) - \mu(0, x_i) \\ \theta_n(P) &= n^{-1} \sum_{i=1}^{n} \{y_i - \mu(a_i, x_i)\}\beta(a_i, x_i) \\ \sigma_n^2(P) &= n^{-1} \sum_{i=1}^{n} \left[\mu(1, x_i) - \mu(0, x_i) - \psi_n(P) + \{y_i - \mu(a_i, x_i)\}\beta(a_i, x_i)\right]^2 \end{aligned}$$

Hence the score statistic is written

$$M_n(P) = \frac{\{\psi_n(P) - \Psi(P) + \theta_n(P)\}^2}{\sigma_n^2(P) + \{\psi_n(P) - \Psi(P)\}^2 + 2\theta_n(P)\{\psi_n(P) - \Psi(P)\}}$$

which is a complicated function of P. We will return to this example after a discussion of inference strategies related to TMLE.

6.4.1 Parametric submodels and TMLE

TMLE point estimators, described in Section 6.2.2, require obtaining a targeted distribution $\hat{P}_n^*$ from an initial distribution $\hat{P}_n$ such that $\boldsymbol{U}_n(\hat{P}_n^*) = 0$. In the last couple of decades, targeting learning methods

have been developed which address this task (van der Laan and Gruber, 2016). Building on these methods, we consider a parametric submodel $\hat{P}_{n,\boldsymbol{\epsilon}}$, which is a set of distributions with indexing parameter $\boldsymbol{\epsilon} \in \mathbb{R}^d$, that contains the initial distribution estimate at $\boldsymbol{\epsilon} = 0$, i.e. $\hat{P}_{n,0} = \hat{P}_n$. In the original description of TMLE and one-step TMLE, $\epsilon \in \mathbb{R}$ is a univariate parameter, however, we consider a vector parameter $\boldsymbol{\epsilon}$, so that $\hat{P}_{n,\boldsymbol{\epsilon}}$ describes a richer submodel.

TMLE point estimators also require specification of a 'loss-function' $l(z, P)$, which implies a sample loss

$$L_n(P) \equiv n^{-1} \sum_{i=1}^{n} l(z_i, P).$$

When $l(z, P)$ is a negative log-likelihood, the resulting estimator is referred to as a targeted maximum likelihood estimation estimator. The key insight, which follows from similar results on so-called 'one-step TMLE', is that one can construct the submodel to satisfy the differential equation

$$\frac{\partial L_n(\hat{P}_{n,\boldsymbol{\epsilon}})}{\partial \boldsymbol{\epsilon}} = \boldsymbol{U}_n(\hat{P}_{n,\boldsymbol{\epsilon}}). \tag{6.8}$$

with the boundary condition $\hat{P}_{n,0} = \hat{P}_n$. If this equation is satisfied then for $\hat{P}^*_n \equiv \hat{P}_{n,\boldsymbol{\epsilon}^*}$, where

$$\boldsymbol{\epsilon}^* = \underset{\boldsymbol{\epsilon} \in \mathbb{R}^d}{\arg\min}\, L_n(\hat{P}_{n,\boldsymbol{\epsilon}}),$$

$\boldsymbol{U}_n(\hat{P}^*_n) = 0$ and $\boldsymbol{\Psi}(\hat{P}^*_n)$ is a RAL estimator of $\boldsymbol{\Psi}(P_0)$. This estimator is a 'minimum loss-based' in the sense that the sample loss $L_n(\hat{P}_{n,\boldsymbol{\epsilon}})$ is minimised over the submodel. Variations in this method include 'local' TMLE and 'one-step' TMLE, where the differential equation (6.8) is respectively replaced with

$$\frac{\partial L_n(\hat{P}_{n,\boldsymbol{\epsilon}})}{\partial \boldsymbol{\epsilon}}\Big|_{\boldsymbol{\epsilon}=0} = \boldsymbol{U}_n(\hat{P}_n) \tag{6.9}$$

$$\frac{\partial L_n(\hat{P}_{n,\epsilon})}{\partial \epsilon} = ||\boldsymbol{U}_n(\hat{P}_{n,\epsilon})|| \tag{6.10}$$

where $||.||$ denotes the euclidean norm, and $\epsilon \in \mathbb{R}$ is a univariate parameter. The euclidean norm appears in (6.10) so that the resulting parametric submodel is indexed by a univariate parameter ϵ, even when the estimand (and hence plug-in bias) have dimenstion $d > 1$. Parametric submodels satisfying (6.10), and the boundary condition $\hat{P}_{n,0} = \hat{P}_n$ are referred to as 'universally least favourable' submodels, whereas those satisfying (6.9), with the same boundary condition, are referred to as 'locally least favourable'. We remark that the locally least favourable submodels are not guaranteed to contain distributions for which the plug-in bias $\boldsymbol{U}_n(P)$ is small.

Example 3 continued: ATE point estimators

Point estimation of the ATE using one-step bias correction estimators and TMLE estimators is now somewhat of a canonical problem in the literature on inference for nonparametric estimands. It is helpful to revisit this problem here to inform the discussion on interval estimation in Section 6.4.2. Rather than estimating a full distribution $\hat{P}_n$ of P_0, $M_n(P_0)$ is a function of P_0 only through $\mu_0(a, x)$, $\pi_0(x)$ and $\Psi(P_0)$. It is sufficient therefore to obtain initial estimates for $\mu_0(a, x)$ and $\pi_0(x)$, which we denote $\hat{\mu}(a, x)$ and $\hat{\pi}(x)$, and to define $\hat{P}_n$ such that the marginal covariate distribution follows the empirical covariate distribution. This implies an initial plug-in estimator of the ATE $\Psi(\hat{P}_n) = \psi_n(\hat{P}_n)$, where the previous

summary statistics under $\hat{P}_n$ evaluate to

$$\psi_n(\hat{P}_n) = n^{-1}\sum_{i=1}^{n}\hat{\mu}(1,x_i) - \hat{\mu}(0,x_i)$$
$$\theta_n(\hat{P}_n) = n^{-1}\sum_{i=1}^{n}\{y_i - \hat{\mu}(a_i,x_i)\}\hat{\beta}(a_i,x_i)$$
$$\sigma_n^2(\hat{P}_n) = n^{-1}\sum_{i=1}^{n}\Big[\hat{\mu}(1,x_i) - \hat{\mu}(0,x_i) - \psi_n(\hat{P}_n) + \{y_i - \hat{\mu}(a_i,x_i)\}\hat{\beta}(a_i,x_i)\Big]^2$$

and where $\hat{\beta}(a,x)$ is obtained by replacing $\pi_0(x)$ with $\hat{\pi}(x)$ in the expression for $\beta_0(a,x)$. It follows that $U_n(\hat{P}_n) = \theta_n(\hat{P}_n)$ and $I_n(\hat{P}_n) = \sigma_n^2(\hat{P}_n)$, hence the one-step bias correction estimator of the ATE is $\hat{\Psi} = \psi_n(\hat{P}_n) + \theta_n(\hat{P}_n)$. This estimator is RAL given assumptions on the remainder term, with variance estimated by

$$n^{-1}\tilde{I}_n(\hat{P}_n) = n^{-2}\sum_{i=1}^{n}\Big[\hat{\mu}(1,x_i) - \hat{\mu}(0,x_i) - \hat{\Psi} + \{y_i - \hat{\mu}(a_i,x_i)\}\hat{\beta}(a_i,x_i)\Big]^2$$

The one-step bias correction estimator of the ATE suffers from the invariance problems described in Section 6.2.3. Instead, one may prefer a TMLE estimator which is invariant to differentiable reparameterisations of the ATE. Consider parametric submodels associated with the logistic loss function

$$l(z_i,\hat{P}_{n,\epsilon}) \equiv -y_i\log\{\hat{\mu}_\epsilon(a_i,x_i)\} - (1-y_i)\log\{1-\hat{\mu}_\epsilon(a_i,x_i)\}$$

where $\hat{\mu}_\epsilon(a_i,x_i)$ denotes the estimate of $\mu_0(a_i,x_i)$ under the parametric submodel $\hat{P}_{n,\epsilon}$, with univariate parameter $\epsilon \in \mathbb{R}$ such that $\hat{P}_{n,0} = \hat{P}_n$. In the current setting, the differential equation in (6.8) becomes

$$\frac{\partial}{\partial\epsilon}\left\{n^{-1}\sum_{i=1}^{n} l(z_i,\hat{P}_{n,\epsilon})\right\} = \theta_n(\hat{P}_{n,\epsilon})$$
$$\implies n^{-1}\sum_{i=1}^{n}\frac{\partial l(z_i,\hat{P}_{n,\epsilon})}{\partial\epsilon} - \{y_i - \hat{\mu}_\epsilon(a_i,x_i)\}\hat{\beta}_\epsilon(a_i,x_i) = 0$$

where $\hat{\beta}_\epsilon(a_i,x_i)$ denotes $\beta_0(a,x)$ evaluated under $\hat{P}_{n,\epsilon}$. This differential equation is satisfied, along with the boundary condition $\hat{P}_{n,0} = \hat{P}_n$, by setting the marginal covariate distribution under $\hat{P}_{n,\epsilon}$ equal to the empirical covariate distribution, setting $\hat{\beta}_\epsilon(a,x) = \hat{\beta}(a,x)$, and defining

$$\text{Logit}\{\hat{\mu}_\epsilon(a,x)\} = \text{Logit}\{\hat{\mu}(a,x)\} + \epsilon\hat{\beta}(a,x). \tag{6.11}$$

where Logit represents the logistic function. The targeted distribution estimator, for the logistic loss function, is then obtained as $\hat{P}_n^* = \hat{P}_{n,\epsilon^*}$ where $\epsilon^* = \arg\min_{\epsilon\in\mathbb{R}} L_n(\hat{P}_{n,\epsilon})$. This targeted distribution implies the TMLE point estimator for the ATE $\Psi(\hat{P}_n^*) = \psi_n(\hat{P}_n^*)$ with variance estimated by $n^{-1}I_n(\hat{P}_n^*) = n^{-1}\sigma_n^2(\hat{P}_n^*)$. In the next section, we outline how a score type interval can be constructed using similar parametric submodels.

6.4.2 Proposal: Detargeted interval estimation

Our proposed score testing procedure, which we call detargeted interval estimation (DIE), is outlined in Algorithm 1. This algorithm makes use of parametric submodels, of the type described in Section 6.4.1 centred on a targeted distribution estimator $\hat{P}_n^*$ rather than an initial estimator $\hat{P}_n$. DIE is so called since, we imagine moving in distribution space (but within a parametric submodel) away from the TMLE distribution point estimator $\hat{P}_n^*$, until the resulting 'detargeted' distribution estimator can be rejected on

the basis of a score test at significance level α. This procedure results in a set of non-rejected distributions (which represent a subset of the parametric submodel), which can be mapped to a set of non-rejected estimand values, i.e. a confidence set for $\boldsymbol{\Psi}(P_0)$ with significance level α.

We consider a d dimensional estimand $\boldsymbol{\Psi} : \mathcal{M} \mapsto \mathbb{R}^d$, and assume that an initial targeted distribution $\hat{P}_n^*$, such that $\boldsymbol{U}_n(\hat{P}_n^*) = 0$, has already been obtained through a previous TMLE estimation procedure. The DIE procedure is agnostic to exactly how the targeted distribution estimator is obtained, making the procedure quite generic, i.e. applicable to various initial distribution estimators and TMLE targeting strategies. Next we define a parametric submodel $\hat{P}_{n,\epsilon}^*$ indexed by $\boldsymbol{\epsilon} \in \mathbb{R}^d$ such that $\hat{P}_{n,0}^* = \hat{P}_n^*$. In the 'universal' version of our procedure (universal-DIE), we construct the parametric submodel to satisfy the differential equation

$$\frac{\partial L_n(\hat{P}_{n,\boldsymbol{\epsilon}}^*)}{\partial \boldsymbol{\epsilon}} = \boldsymbol{U}_n(\hat{P}_{n,\boldsymbol{\epsilon}}^*). \tag{6.12}$$

which is analogous to (6.8). In the 'local' version of our procedure (local-DIE), the parametric submodel is constructed such that

$$\frac{\partial L_n(\hat{P}_{n,\boldsymbol{\epsilon}}^*)}{\partial \boldsymbol{\epsilon}}\Big|_{\epsilon=0} = \boldsymbol{U}_n(\hat{P}_n^*). \tag{6.13}$$

which is analogous to the locally least-favourable model in (6.9).

Algorithm 1: Detargeted interval estimation

1. Use a TMLE algorithm to obtain a targeted distribution estimator $\hat{P}_n^*$ such that $\boldsymbol{U}_n(\hat{P}_n^*) = 0$.
2. Define a parametric submodel $\hat{P}_{n,\boldsymbol{\epsilon}}^*$, indexed by $\boldsymbol{\epsilon} \in \mathbb{R}^d$, such that $\hat{P}_{n,\boldsymbol{\epsilon}}^* = \hat{P}_n^*$ and either (6.12) or (6.13) is satisfied (this choice determines which flavour of the DIE procedure is used).
3. Use a numerical search procedure to find the values of $\boldsymbol{\epsilon}$ such that $M_n(\hat{P}_{n,\boldsymbol{\epsilon}}^*) \leq c_\alpha^2/n$. In the case of a $d = 1$ scalar estimand, it is sufficient to find the two value of ϵ such that $M_n(\hat{P}_{n,\epsilon}^*) = c_\alpha^2/n$.
4. Return the set of estimand values $\Psi(\hat{P}_{n,\epsilon}^*)$ which satisfy the inequality in step 3. In the case of a $d = 1$ scalar estimand, this will correspond to the interval bounded by the two ϵ values described in step 3.

We remark that these parametric submodels, centred on the TMLE distribution estimator $\hat{P}_n^*$, require specification of a loss-function $l(z, P)$ to obtain the sample loss $L_n(P)$. Due to the 'TMLE agnostic' nature of our proposal, the DIE loss-function does not need to be the same as the loss-function that was used to obtain $\hat{P}_n^*$ from an initial distribution estimator $\hat{P}_n$, though we will only consider examples for which this is the case. Examples of the DIE procedure applied to standard causal estimands are provided after the discussion of theoretical results below.

6.4.3 Theoretical properties

Consider a targeted distribution estimator $\hat{P}_n^* \in \mathcal{M}$ such that $\boldsymbol{U}_n(\hat{P}_n^*) = 0$. The targeted distribution estimator implies a targeted point estimator $\boldsymbol{\Psi}(\hat{P}_n^*)$ of $\boldsymbol{\Psi}(P_0)$, which is RAL when requisite assumptions hold, e.g. consistency/ boundedness assumptions on $\hat{P}_n^*$, and (e.g. Donsker) assumptions on the IC. Under such assumptions, the covariance of $\boldsymbol{\Psi}(\hat{P}_n^*)$ is estimated as $n^{-1}I_n(\hat{P}_n^*)$, implying a Wald confidence set for $\boldsymbol{\Psi}(P_0)$

$$\hat{W}_n \equiv \left\{ \boldsymbol{\psi}_0 \text{ such that } W_n^*(\boldsymbol{\psi}_0) \leq \frac{c_\alpha^2}{n} \right\} \tag{6.14}$$

where we denote the Wald statistic

$$W_n^*(\boldsymbol{\psi}_0) \equiv \{\boldsymbol{\Psi}(\hat{P}_n^*) - \boldsymbol{\psi}_0\}^\top I_n^{-1}(\hat{P}_n^*)\{\boldsymbol{\Psi}(\hat{P}_n^*) - \boldsymbol{\psi}_0\}$$

and c_α^2 represents the $1-\alpha$ quantile from the χ_d^2 distribution. Theorem 6 below describes how this Wald statistic is asymptotically related to the score statistic when evaluated at certain plug-in estimand values. In particular, the Theorem concerns distribution estimators which are in, some sense, close to the targeted distribution estimator.

Theorem 6 (Score-Wald Asymptotic Equivalence) *Let $\hat{P}_n^*$ be a targeted distribution estimator such that $\boldsymbol{U}_n(\hat{P}_n^*) = 0$, the plug-in estimator $\boldsymbol{\Psi}(\hat{P}_n^*)$ is RAL, and the covariance estimator $I_n(\hat{P}_n^*)$ is consistent, i.e. $I_n(\hat{P}_n^*) \xrightarrow{p} I_0$ as $n \to \infty$. Further let $\tilde{P}_n$ be an alternative distribution estimator such that for $\delta > 0$, $||\boldsymbol{U}_n(\tilde{P}_n)|| < \delta/\sqrt{n}$, the covariance estimator $I_n(\tilde{P}_n) \xrightarrow{p} I_0$ is consistent, and the one-step bias correction estimator $\boldsymbol{\Psi}(\tilde{P}_n) + \boldsymbol{U}_n(\tilde{P}_n)$ is RAL.*

Under these conditions, $W_n^*(\boldsymbol{\Psi}(\tilde{P}_n)) - M_n(\tilde{P}_n) = o_p(n^{-1})$.

Proof 2 *Consider the difference in remainder terms from the von Mises expansion in (6.4), for the plug-in estimators $\boldsymbol{\Psi}(\tilde{P}_n)$ and $\boldsymbol{\Psi}(\hat{P}_n^*)$*

$$\begin{aligned}\boldsymbol{R}_n &\equiv \left\{\boldsymbol{\Psi}(\tilde{P}_n) - \boldsymbol{\Psi}(P_0) + \boldsymbol{U}_n(\tilde{P}_n) - \boldsymbol{U}_n(P_0)\right\} - \left\{\boldsymbol{\Psi}(\hat{P}_n^*) - \boldsymbol{\Psi}(P_0) + \boldsymbol{U}_n(\hat{P}_n^*) - \boldsymbol{U}_n(P_0)\right\} \\ &= \boldsymbol{\Psi}(\hat{P}_n^*) - \boldsymbol{\Psi}(\tilde{P}_n) - \boldsymbol{U}_n(\tilde{P}_n)\end{aligned} \tag{6.15}$$

where, to obtain the second line, we use the fact that $\boldsymbol{U}_n(\hat{P}_n^*) = 0$. Since $\boldsymbol{\Psi}(\hat{P}_n^*)$ and $\boldsymbol{\Psi}(\tilde{P}_n) + \boldsymbol{U}_n(\tilde{P}_n)$ are RAL, $\boldsymbol{R}_n = o_p(n^{-1/2})$. By algebraic manipulation

$$\begin{aligned}W_n^*(\boldsymbol{\Psi}(\tilde{P}_n)) &= \{\boldsymbol{U}_n(\tilde{P}_n) + \boldsymbol{R}_n\}^\top I_n^{-1}(\hat{P}_n^*)\{\boldsymbol{U}_n(\tilde{P}_n) + \boldsymbol{R}_n\} \\ &= M_n(\tilde{P}_n) + \boldsymbol{U}_n^\top(\tilde{P}_n)\{I_n^{-1}(\hat{P}_n^*) - I_n^{-1}(\tilde{P}_n)\}\boldsymbol{U}_n(\tilde{P}_n) + 2\boldsymbol{R}_n^\top I_n^{-1}(\hat{P}_n^*)\boldsymbol{U}_n(\tilde{P}_n) + \boldsymbol{R}_n^\top I_n^{-1}(\hat{P}_n^*)\boldsymbol{R}_n \\ &= M_n(\tilde{P}_n) + o_p(n^{-1})\end{aligned}$$

which completes the proof. Note for the final equality that consistency of the covariance estimators implies $I_n^{-1}(\hat{P}_n^*) - I_n^{-1}(\tilde{P}_n) \xrightarrow{p} 0$.

Corollary 6.1 *Define the score set*

$$\tilde{S}_n \equiv \left\{\boldsymbol{\Psi}(\tilde{P}_n) \text{ such that } M_n(\tilde{P}_n) \leq \frac{c_\alpha^2}{n}\right\}$$

where $\tilde{P}_n$ is a distribution estimator satisfying the requirements of the main Theorem. The Wald set $\hat{W}_n$ asymptotically contains the score set in the sense that $\boldsymbol{\psi} \in \tilde{S}_n \implies \boldsymbol{\psi} \in \hat{W}_n$ (almost surely), and if there exists $\tilde{P}_n$ such that $\boldsymbol{\Psi}(\tilde{P}_n) = \boldsymbol{\psi}$, then $\boldsymbol{\psi} \in \hat{W}_n \implies \boldsymbol{\psi} \in \tilde{S}_n$ (almost surely).

Theorem 6 and Corollary 6.1 asymptotically link the Wald confidence set $\hat{W}_n$ to the score set $\tilde{S}_n$. This is significant since the score set $\tilde{S}_n$ is defined independently of the targeted distribution estimator $\hat{P}_n^*$. In particular, the score set (asymptotically) represents the set of plug-in estimand values which are elements of a Wald set that is centred on any targeted point estimate.

The 'alternative' distribution estimators in Theorem 6 effectively describe a region of distribution space centred on some targeted distribution $\hat{P}_n^*$. The requirement that the $\tilde{P}_n$ is 'close' to $\hat{P}_n^*$, in the sense of describing a RAL point estimator with small plug-in bias, suggests that this region might be interpreted as a ball in estimand space, centred on the targeted distribution estimator, with a 'radius' that decreases with $n^{-1/2}$. In essence, the DIE procedure considers smooth parametric submodels which parameterise this region/ ball. Corollary 6.2 describes how Theorem 6 applies to such parametric submodels. We remark that Corollary 6.2 is agnostic to the exact form of the parametric submodel, as reflected in the fact that the parametric submodels of Corollary 6.2 have an arbitrary parameter dimension q, which is not necessarily the same as the dimension of the estimand d.

Corollary 6.2 (Score-Wald Asymptotic Equivalence for parametric submodels) *Let $\hat{P}_{n,\boldsymbol{\epsilon}}^*$ be a parametric submodel indexed by a parameter $\boldsymbol{\epsilon} \in \mathbb{R}^q$ such that $\hat{P}_{n,0}^* = \hat{P}_n^*$ and, assume that $\boldsymbol{U}_n(\hat{P}_{n,\boldsymbol{\epsilon}}^*)$ is*

differentiable w.r.t. $\boldsymbol{\epsilon}$ *in a neighbourhood of* $\boldsymbol{\epsilon} = 0$ *and* $I_n(\hat{P}^*_{n,\boldsymbol{\epsilon}})$ *is continuous in the same neighbourhood. For* $\boldsymbol{\epsilon}$ *such that* $||\boldsymbol{\epsilon}|| \leq k/\sqrt{n}$*, where* $k > 0$ *is a constant,* $M_n(\hat{P}^*_{\boldsymbol{\epsilon}}) - W^*_n(\boldsymbol{\Psi}(\hat{P}^*_{\boldsymbol{\epsilon}})) = o_p(n^{-1})$*, and Corollary 6.1 applies where* $\tilde{P}_n$ *is replaced with* $\hat{P}^*_{n,\boldsymbol{\epsilon}}$*.*

To see why this Corollary is applicable to DIE score-intervals, despite the condition that $||\boldsymbol{\epsilon}|| \leq k/\sqrt{n}$, we consider the Taylor-expansion of $M_n(\hat{P}^*_{\boldsymbol{\epsilon}})$ about $\boldsymbol{\epsilon} = 0$

$$M_n(\hat{P}^*_{\boldsymbol{\epsilon}}) = \boldsymbol{\epsilon}^\top \boldsymbol{V}_n^\top(\hat{P}^*_{n,0}) I_n^{-1}(\hat{P}^*_{n,0}) \boldsymbol{V}_n(\hat{P}^*_{n,0})\boldsymbol{\epsilon} + O(\boldsymbol{\epsilon}^3) \tag{6.16}$$

where we define the d by q matrix

$$\boldsymbol{V}_n(\hat{P}^*_{n,0}) = \frac{d\boldsymbol{U}_n(\hat{P}^*_{n,\boldsymbol{\epsilon}})}{d\boldsymbol{\epsilon}}\Big|_{\epsilon=0}$$

and $O(\boldsymbol{\epsilon}^3)$ is short hand for higher order terms of the form $\sum_{i=1}^q \sum_{j=1}^q \sum_{k=1}^q a_{ijk}\epsilon_i\epsilon_j\epsilon_k$ for some rank 3 tensor a_{ijk}. The Taylor expansion in (6.16) therefore implies that the inequality $M_n(\hat{P}^*_{\boldsymbol{\epsilon}}) \leq c_\alpha^2/n$ is satisfied only by parameter values $\boldsymbol{\epsilon}$ such that $||\boldsymbol{\epsilon}|| \leq k/\sqrt{n}$.

The requirement that the plug-in bias is differentiable, i.e. $\boldsymbol{V}_n(\hat{P}^*_{n,0})$ exists, is perhaps more restrictive, since it requires additional smoothness of the parametric submodel. It may be difficult to construct such smooth parametric sub-models, for example, when the IC $\boldsymbol{\phi}(Z, P_0)$ is discontinuous, e.g. contains indicator/ step functions. In the DIE examples, which we describe below, however, smoothness is easily achieved, and we believe this will be the case for many estimands of interest. We remark that for parametric submodels where (6.12) holds, the differentiability requirement of Corollary 6.2 is equivalent to a smooth curvature requirement on the loss function

$$\boldsymbol{V}_n(\hat{P}^*_{n,0}) = \frac{d^2 L_n(\hat{P}^*_{n,\boldsymbol{\epsilon}})}{d\boldsymbol{\epsilon}^2}\Big|_{\epsilon=0}.$$

Given that Corollary 6.2 is agnostic to the parametric form of the submodel, $\hat{P}^*_{n,\boldsymbol{\epsilon}}$, we recommend that the main criteria for choosing a parametric submodel should be (i) smoothness, and (ii) the extent to which a large estimand range is covered for small parameter values. The latter property motivates parametric submodels which maximally change the estimand for small changes in the submodel index parameter. Such models are described in the one-step TMLE procedure of van der Laan and Gruber (2016), motivating the parametric submodels in the proposed DIE procedure in Section 6.4.2.

Example 3 continued: DIE for the ATE

Let $\hat{P}^*_n$ denote the targeted distribution estimator derived for the ATE above. This distribution is defined through the functions $\hat{\mu}^*(a,x)$ and $\hat{\beta}^*(a,x)$, which denote $\mu_0(a,x)$ and $\beta_0(a,x)$ evaluated under $\hat{P}^*_n$. Additionally, the marginal covariate distribution under $\hat{P}^*_n$ is equal to the empirical (observed) covariate distribution. Consider a parametric submodel $\hat{P}^*_{n,\epsilon}$ for $\epsilon \in \mathbb{R}$ such that for all ϵ, $\hat{\beta}^*_\epsilon(a,x) = \hat{\beta}^*(a,x)$,

$$\text{Logit}\{\hat{\mu}^*_\epsilon(a,x)\} = \text{Logit}\{\hat{\mu}^*(a,x)\} + \epsilon\hat{\beta}^*(a,x) \tag{6.17}$$

and the marginal covariate distribution under $\hat{P}^*_{n,\epsilon}$ is equal to the empirical distribution. This parametric submodel satisfies the boundary condition that $\hat{P}^*_{n,\epsilon} = \hat{P}^*_n$ and, for the logistic loss function, also satisfies the requirements for both the universal-DIE and local-DIE submodels in (6.12) and (6.13) respectively. The fact that the 'universal' and 'local' submodels coincide is unsurprising, given similar observations for TMLE point estimators of the ATE (van der Laan and Gruber, 2016).

We remark that the DIE submodel $\hat{P}^*_{n,\epsilon}$ is equal to the parametric submodel used for TMLE estimation of the ATE in (6.11) since,

$$\text{Logit}\{\hat{\mu}^*_\epsilon(a,x)\} = \text{Logit}\{\hat{\mu}(a,x)\} + (\epsilon^* + \epsilon)\hat{\beta}(a,x)$$

where ϵ^* is the value obtained during the TMLE estimation algorithm, and we have used the fact that $\hat{\beta}^*(a_i, x_i) = \hat{\beta}(a_i, x_i)$.

The next step of the DIE algorithm is to find all ϵ satisfying the inequality

$$M_n(\hat{P}^*_{n,\epsilon}) = \frac{\theta_n^2(\hat{P}^*_{n,\epsilon})}{\sigma_n^2(\hat{P}^*_{n,\epsilon})} \leq \frac{c_\alpha^2}{n}$$

Since, for small ϵ, $M_n(\hat{P}^*_\epsilon)$ behaves like $O(\epsilon^2)$, the inequality above is bounded by two ϵ values which can be obtained numerically e.g. using Newton-Raphson iteration/ grid search/ bisection method. Letting ϵ_1 and ϵ_2 denote the two boundaries of this inequality, then the resulting CI for the ATE $\Psi(P_0)$ is $(\psi_n(\hat{P}^*_{\epsilon_1}), \psi_n(\hat{P}^*_{\epsilon_2}))$, which is asymptotically equivalent to the Wald interval $\psi_n(\hat{P}^*_0) \pm c_\alpha \sigma_n(\hat{P}^*_0)/\sqrt{n}$.

Example 4: Mean counterfactual outcomes

In this example we consider a $d = 2$ dimensional estimand, which uses the same setup as in Example 3. In particular we consider inference for $\boldsymbol{\Psi}(P_0) = (\Psi_1(P_0), \Psi_2(P_0))$ with components $\Psi_1(P_0) = P_0\{\mu_0(1, X)\}$ and $\Psi_2(P_0) = P_0\{\mu_0(0, X)\}$. These estimands have ICs $\boldsymbol{\phi}(Z, P_0) = (\phi_1(Z, P_0), \phi_2(Z, P_0))$ with

$$\begin{aligned}\phi_1(Z, P_0) &= \mu_0(1, X) - \Psi_1(P_0) + \{Y - \mu_0(A, X)\}\beta_0(A, X)A \\ \phi_2(Z, P_0) &= \mu_0(0, X) - \Psi_2(P_0) + \{Y - \mu_0(A, X)\}\beta_0(A, X)(A - 1)\end{aligned}$$

Using these ICs one can construct $\boldsymbol{U}_n(P)$ and $I_n(P)$ and hence the score statistic for the joint counterfactual means $M_n(P)$. As in Example 3, we consider an initial distribution estimator $\hat{P}_n$. In fact this initial distribution estimator can be the same as the one used in Example 3, since the $M_n(P_0)$ is a function of P_0 only through $\mu_0(a, x)$, $\pi_0(x)$ and the marginal covariate distribution.

As in Example 3, we construct a targeted distribution estimator $\hat{P}^*_n$ such that $\boldsymbol{U}_n(\hat{P}^*_n) = 0$. This can be achieved, for instance, by minimising the logistic-loss over a parametric submodel $\hat{P}_{n,\boldsymbol{\epsilon}}$, for $\boldsymbol{\epsilon} \in \mathbb{R}^2$, which is defined such that, for all $\boldsymbol{\epsilon}$, the marginal covariate distribution is equal to the empirical covariate distribution, $\hat{\beta}_\epsilon(a, x) = \hat{\beta}(a, x)$, and

$$\text{Logit}\{\hat{\mu}_{\boldsymbol{\epsilon}}(a, x)\} = \text{Logit}\{\hat{\mu}(a, x)\} + \{a\epsilon_1 + (a - 1)\epsilon_2\}\hat{\beta}(a, x) \tag{6.18}$$

It is straight forward to verify that this parametric submodel satisfies the boundary condition that $\hat{P}_{n,0} = \hat{P}_n$ and the differential equations in both (6.10) and (6.9). We remark that the parametric submodel for the ATE in (6.11) represents a submodel of the (6.18) such that $\epsilon_1 = -\epsilon_2$ is a univariate parameter. The targeted distribution $\hat{P}^*_n$ is obtained as the element of the parametric submodel which minimises the logistic-loss $L_n(\hat{P}_{n,\boldsymbol{\epsilon}})$ over $\boldsymbol{\epsilon}$.

Using the targeted distribution estimator $\hat{P}^*_n$ one can now construct DIE intervals. As with the DIE intervals for the ATE, our local-DIE and universal-DIE intervals will coincide. Consider a parametric submodel $\hat{P}^*_{n,\boldsymbol{\epsilon}}$, for $\boldsymbol{\epsilon} \in \mathbb{R}^2$, which is defined such that, for all $\boldsymbol{\epsilon}$, the marginal covariate distribution is equal to the empirical covariate distribution, $\hat{\beta}^*_\epsilon(a, x) = \hat{\beta}^*(a, x)$, and

$$\text{Logit}\{\hat{\mu}^*_{\boldsymbol{\epsilon}}(a, x)\} = \text{Logit}\{\hat{\mu}^*(a, x)\} + \{a\epsilon_1 + (a - 1)\epsilon_2\}\hat{\beta}^*(a, x) \tag{6.19}$$

Again this is very similar to parametric submodel used for TMLE, i.e. (6.18), but centred on the targeted distribution estimator, rather than the initial distribution estimator. As for the ATE, one can show that (6.19) and (6.18) are in fact the same parametric submodel, however this not a requirement of the DIE proposal.

Finally, the DIE confidence set of mean counterfactual outcomes is the set

$$\hat{S}_n = \left\{ \boldsymbol{\Psi}(\hat{P}^*_{n,\boldsymbol{\epsilon}}) \text{ for } \boldsymbol{\epsilon} \in \mathbb{R}^2 \text{ such that } M_n(\hat{P}^*_{n,\boldsymbol{\epsilon}}) \leq \frac{c_\alpha^2}{n} \right\}.$$

This could be estimated by numerical methods, or else using a grid search. Since $\hat{P}^*_{n,\boldsymbol{\epsilon}}$ is a $d = 2$ dimensional estimand, this set will correspond to a region on the plain, which is visualised example of which is provided in Figure 6.2, which relates to the simulation study in Section 6.5.

Example 5: Variance of treatment effect (VTE)

Consider the same setup as in Example 3 and define the conditional average treatment effect (CATE) by $E(Y_1 - Y_0|X = x)$. Under the same causal assumptions, this is identified by $\tau_0(x) = \mu_0(1, x) - \mu_0(0, x)$. Suppose that interest is in obtaining a CI for the VTE $\Psi(P_0) = \text{var}\{\tau_0(X)\} = P_0\{\tau_0^2(X)\} - P_0\{\tau_0(X)\}^2$, with IC

$$\phi(Z, P_0) = \{\tau_0(X) - \Theta(P_0)\}^2 - \Psi(P_0) + 2\{Y - \mu_0(A, X)\}\{\tau_0(X) - \Theta(P_0)\}\beta_0(A, X)$$

where $\Theta(P_0) = P_0\{\tau_0(X)\}$ denotes the ATE. Using these ICs one can construct $U_n(P)$ and $I_n(P)$ and hence the score statistic for the VTE $M_n(P)$. We remark that when $\tau_0(X)$ is constant, i.e. under treatment effect homogeneity, then $\phi(Z, P_0) = 0$ and $I_0 = 0$. When this is the case then $nM_n(P_0)$ does not converge to a χ_1^2 distributed random variable. Inference in such a setting remains generally an open problem that has motivated the study of higher-order pathwise derivatives of the estimands, e.g. Carone et al. (2018). For this reason, we consider inference under the assumption that $\Psi(P_0) > 0$.

Consider the same initial distribution estimator $\hat{P}_n$ from Examples 3 and 4. We obtain a targeted distribution estimator from $\hat{P}_n$ using a one-step TMLE algorithm, similar to the one described by Levy et al. (2021), which we outline here. Consider the parametric submodel $\hat{P}_{n,\epsilon}$ for $\epsilon \in \mathbb{R}$ such that, for all ϵ, the marginal covariate distribution under $\hat{P}_{n,\epsilon}$ is equal to the empirical covariate distribution, $\hat{\beta}_\epsilon(a, x) = \hat{\beta}(a, x)$ and $\hat{\tau}_\epsilon(x) = \hat{\mu}_\epsilon(1, x) - \hat{\mu}_\epsilon(0, x)$ where, letting $\hat{\theta}_t = \Theta(\hat{P}_{n,\epsilon})$ denote the plug-in estimator for the ATE under $\hat{P}_{n,\epsilon}$, the conditional mean $\hat{\mu}_\epsilon(a, x)$ is defined through the differential equation

$$\frac{d}{d\epsilon}\text{Logit}\{\hat{\mu}_\epsilon(a, x)\} = 2\hat{\beta}(a, x)\{\hat{\tau}_\epsilon(x) - \hat{\theta}_\epsilon\} \tag{6.20}$$

$$\implies \text{Logit}\{\hat{\mu}_\epsilon(a, x)\} = \text{Logit}\{\hat{\mu}(a, x)\} + 2\hat{\beta}(a, x)\int_0^\epsilon \{\hat{\tau}_t(x) - \hat{\theta}_t\}dt \tag{6.21}$$

To obtain the integral expression above, the boundary condition $\hat{P}_{n,0} = \hat{P}_n$ is applied. One can verify that this parametric submodel satisfies (6.8) for the logistic loss. The nonlinearity (with respect to ϵ) of the differential equation in (6.20), however makes this parametric submodel difficult to work with in practice. Unlike the conditional mean models in (6.11) and (6.18), the implicit expression for $\hat{\mu}_\epsilon(a, x)$ in (6.21) cannot easily be evaluated for given ϵ. To construct a numerical approximation to $\hat{\mu}_\epsilon(a, x)$, we let $\epsilon = (m + 1)\delta$ where m is an integer and δ represents a small step, and we replace the differential equation in (6.20) with the finite step approximation

$$\text{Logit}\{\hat{\mu}_{(m+1)\delta}(a, x)\} = \text{Logit}\{\hat{\mu}_{m\delta}(a, x)\} + 2\delta\hat{\beta}(a, x)\{\hat{\tau}_{m\delta}(x) - \hat{\theta}_{m\delta}\}. \tag{6.22}$$

Hence one may approximate $\hat{\mu}_\epsilon(a, x)$ recursively, starting from the known value $\hat{\mu}_0(a, x) = \hat{\mu}(a, x)$. To find the parameter value ϵ^* such that $U_n(\hat{P}_{n,\epsilon^*}) = 0$, one could consider search algorithms where $U_n(\hat{P}_{n,\epsilon})$ is approximated for a discrete set of values with a small (possibly dynamically chosen) step size. This procedure results in a targeted distribution $\hat{P}_n^*$ and associated TMLE point estimator

$$\Psi(\hat{P}_n^*) = n^{-1}\sum_{i=1}^n \left\{\hat{\tau}^*(x_i) - \Theta(\hat{P}_n^*)\right\}^2$$

$$\Theta(\hat{P}_n^*) = n^{-1}\sum_{i=1}^n \hat{\tau}^*(x_i)$$

where $\hat{\tau}^*(x) = \hat{\mu}^*(1, x) - \hat{\mu}^*(0, x)$ denotes the CATE $\tau_0(x)$ evaluated under $\hat{P}_n^*$.

To construct DIE intervals for the VTE, we start by building a parametric submodel $\hat{P}_{n,\epsilon}^*$ around the targeted distribution $\hat{P}_n^*$, with univariate parameter ϵ. For the local-DIE procedure it is sufficient to use a submodel where the marginal covariate distribution follows the empirical covariate distribution, $\hat{\beta}_\epsilon^*(a, x) = \hat{\beta}^*(a, x)$ and

$$\text{Logit}\{\hat{\mu}_\epsilon^*(a, x)\} = \text{Logit}\{\hat{\mu}^*(a, x)\} + 2\epsilon\hat{\beta}^*(a, x)\{\hat{\tau}^*(x) - \hat{\theta}^*\} \tag{6.23}$$

where $\hat{\theta}^* = \Theta(\hat{P}_n^*)$. This submodel is seen to satisfy the boundary condition that $\hat{P}_{n,0}^* = \hat{P}_n^*$ and (6.13) for the logistic loss function. For the universal-DIE procedure, a more complicated submodel is required, which is similar in spirit to (6.21). In particular, consider replacing (6.23) with

$$\text{Logit}\{\hat{\mu}_\epsilon^*(a,x)\} = \text{Logit}\{\hat{\mu}^*(a,x)\} + 2\hat{\beta}^*(a,x)\int_0^\epsilon \{\hat{\tau}_t^*(x) - \hat{\theta}_t^*\}dt$$

which satisfies (6.13) for the logistic loss function.

Regardless of which DIE procedure is used, the final step is to find all ϵ satisfying $M_n(\hat{P}_{n,\epsilon}^*) \leq \frac{c_\alpha^2}{n}$. For small ϵ, $M_n(\hat{P}_\epsilon^*)$ behaves like $O(\epsilon^2)$, regardless of whether the local-DIE or universal-DIE submodel is used. Hence, the inequality is bounded by two ϵ values which can be obtained numerically. Letting ϵ_1 and ϵ_2 denote the two boundaries of this inequality, the resulting CI for the VTE $\Psi(P_0)$ is $(\Psi(\hat{P}_{\epsilon_1}^*), \Psi(\hat{P}_{\epsilon_2}^*))$, which is asymptotically equivalent to the Wald interval $\Psi(\hat{P}^*) \pm c_\alpha I_n(\hat{P}^*)/\sqrt{n}$. One important difference between the two methods, however, is that the DIE interval is guaranteed to contain only positive VTE values.

6.5 Simulation study

6.5.1 Population median

For this simulation study 10^6 iid datasets of size $n \in \{10, 20, ..., 250\}$ were generated from

$$\begin{aligned} Y_1 &\sim \mathcal{N}(0,1) \\ Y_2 &\sim \text{Gamma}(1,1) \\ Y_3 &\sim \text{Beta}(5,1) \end{aligned}$$

and the target estimand was the population median of Y for each of Y_1, Y_2, Y_3. 95% CIs for the population median were estimated using the proposed score based approach in Example 2. In particular, the boundaries of the interval

$$Q_n\left(\frac{1}{2} \pm \frac{c_\alpha}{2\sqrt{n}}\right)$$

were estimated using linear interpolation of the empirical CDF. This was compared for three different values of c_α^2 according to the discussion in Appendix E.2. We compared the proposed score CIs against naive Wald CIs which used linear interpolation of the empirical CDF to obtain a point estimate of the median, and a kernel density estimator with a gaussian kernel and using 'Silverman's rule-of-thumb' to estimate the probability density at the median. The results of this simulation study can be seen in Figure 6.1.

These results suggest that the score based inference of the population median gives improved CI coverage in finite samples compared with Wald based inference. Additionally, we see that the threshold values $c_\alpha^2 = q_{\alpha,d}$ and $c_\alpha^2 = nB_{\alpha,d,n}$ give similar results, with both demonstrating an improvement over the value $c_\alpha^2 = \tilde{q}_{\alpha,d,n}$, which is based on the mean-corrected score statistic.

6.5.2 Causal effects

For this simulation study 10^3 iid datasets of size $n \in \{500, 1000, ...5000\}$ were generated from the structural equation model

$$\begin{aligned} X_1, X_2, X_3 &\sim \text{Uniform}(-1,1) \\ A &\sim \text{Bernoulli}\{\text{Expit}(-0.4X_1 + 0.1X_3 + 0.1X_1X_2)\} \\ Y &\sim \text{Bernoulli}\left\{\text{Expit}\left(-2 - X_1 + X_1X_2 + 2X_2^2 + A(X_1^2(X_1+1.4) + 2.8X_2^2)\right)\right\} \end{aligned}$$

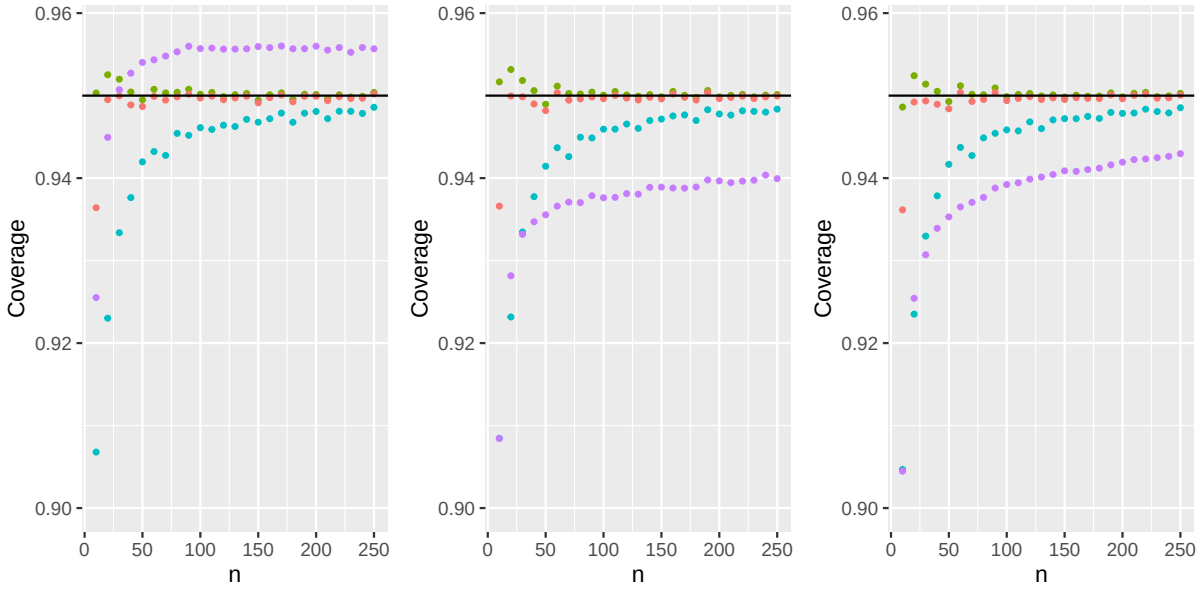


Figure 6.1: Coverage for a 95% CI of the population median plotted against sample size. Each plot represents a different outcome $Y = Y_1, Y_2, Y_3$ respectively. Green, blue and red points represent score based intervals respectively using the threshold values $q_{\alpha,d}, \tilde{q}_{\alpha,d,n}$ and $nB_{\alpha,d,n}$ as discussed in Appendix E.2, with $d = 1$ and $\alpha = 0.05$. Purple points represent coverage of the corresponding Wald based interval.

where Expit denotes the inverse logistic function. We consider construction of the 95% CI for the ATE, and VTE, as in Examples 3 and 5, and construction of the 95% confidence set for the vector of mean potential outcomes as in Example 4. For the latter a visual representation of a $d = 2$ dimensional confidence set for a single dataset is provided in Figure 6.2. The true estimand values for the ATE, VTE and mean potential outcomes were $0.25, 0.035$ and $(0.49, 0.24)$ respectively. This VTE value implies a root-VTE of 0.19, which is similar in magnitude to the ATE.

Two different learning methods were considered to obtain initial estimates of $\mu_0(a, x)$ and $\pi_0(x)$. The first used generalised additive models (GAMs), implemented in the `mgcv` package in R, and the other used gradient boosting trees (GBTs), implemented in the `xgboost` package in R. Initial machine learning fits were 'targeted' towards the relevant estimands using either one-step TMLE, or one-step cross validated-TMLE (CV-TMLE) with $k = 5$ folds. The latter uses a cross-fitting strategy to control biases related to over-fitting. Typically 10-20 folds is recommended in practice, but 5 folds were chosen for this illustration to reduce computation time. The cross validation strategy of CV-TMLE is designed to avoid overfitting of the working models since, e.g. $\mu_0(a_i, x_i)$ is estimated by $\hat{\mu}^{(-i)}(a_i, x_i)$ where $\hat{\mu}^{(-i)}(.,.)$ is trained on a dataset which does not include the ith observation, with similar for $\pi^{(-i)}(x_i)$.

Figures 6.3 and 6.4 show confidence set coverage against samples size and median CI width against sample size for the intervals constructed around the TMLE and CV-TMLE targeted estimators respectively. For the ATE and mean potential outcome vector interval estimators we observe similar results. In particular, the intervals constructed using GAM learning achieve close to nominal (95%) coverage, with and without cross validation and regardless of the interval construction method used (score vs. Wald). For the GBT based learners, the score intervals exceed nominal coverage in the absence of cross validation and are below nominal coverage when cross validated functional estimators are used, whereas the Wald intervals display the opposite behaviour with regard to the effect of cross validation.

For the VTE both score interval types (local and universal) perform similarly, with insufficient evidence to recommend one type over another. For the GAM based intervals, the proposed score methodology achieves significantly improved CI coverage over Wald type intervals for all sample sizes. For GBT based intervals, we observe poor CI coverage, though this is slightly improved in small samples when cross validation is used. The poor coverage of GBT learner based intervals may be caused by the estimator of

$\mu_0(a, x)$ failing to converge at $n^{1/4}$ rate, which is required for the TMLE estimator of the VTE to be RAL.

The improved coverage of the score type CI over the Wald type CI for the VTE could be explained by the fact that, by construction the VTE has a finite support (VTE ≥ 0) hence normality of the TMLE (and CV-TMLE) estimators may be hard to achieve in finite samples, despite asymptotic normality being guaranteed in the asymptotic limit. Conversely, the asymptotic distribution of the score test statistic is unaffected by the finite support of the estimand.

6.6 Conclusion

We have proposed a new method constructing confidence intervals for nonparametric estimands based on score testing. Our framework is theoretically appealing since it is based on test statistics that are invariant to differentiable reparameterisations of the target estimand and do not require estimation of scaling constants which appear in the IC. In simple cases, such as the weighted mean and the population quantile in Examples 1 and 2, our statistic is a function of the unknown data generating distribution only through the estimand of interest. As such, we derive simple confidence intervals which are seen to be asymptotically equivalent to their Wald counterparts, except with certain small sample size corrections. Moreover, we present, to our knowledge, a novel confidence interval estimator for population quantiles, which performs well (in terms of coverage in finite samples) in simulations.

In more complicated settings, such as when the target estimand is the ATE, joint counterfactual mean, or VTE as in Examples 3, 4, and 5 respectively, our score test statistic also depends on infinite dimensional parameters (i.e. functions) of the unknown data generating mechanism. In such settings, we demonstrate that it is sufficient to restrict the space of considered distributions to a parametric submodel centred on a TMLE point estimator. We call our proposal detarged interval estimation (DIE) since it constructs an interval by 'detargeting' (i.e. making worse) a TMLE point estimator, until a null hypothesis test (based on our score statistic) is rejected.

In a simulation studies, our DIE interval estimators show improved coverage compared to Wald type intervals, in settings where the target estimand has a bounded support, e.g. the VTE. We reason that this behaviour occurs because approximate normality of the point estimator is rarely achieved in finite samples when the true value of the target estimand is close to the boundary of the support. Our DIE interval estimators also perform reasonably for the ATE and joint counterfactual mean, where the bounded support of the estimand is less of a concern. In particular, DIE intervals tend to be narrower than their Wald counterparts, for a modest decrease in coverage.

We have also identified two future lines of enquiry which could show promise. Firstly, just as TMLE follows likelihood based inference once a parametric submodel has been constructed, our proposed DIE interval estimators follow GMM based inference once a parametric submodel has been constructed around a TMLE distribution estimator. Connecting nonparametric inference and TMLE to the GMM is potentially significant since it is possible that other GMM techniques could be applied to nonparametric inference problems, e.g. empirical likelihoods and exponentially tilted GMM estimators (Owen, 1988; Qin and Lawless, 1994; Kitamura and Stutzer, 1997; Imbens, 1997; Corcoran, 1998; Imbens, 2002; Newey and Smith, 2004).

Secondly, unlike Wald statistics, it is possible to evaluate the proposed nonparametric score statistics for a given distribution estimator, without having to additionally provide a targeted distribution estimator. This opens the possibility for using score statistics to be used to enhance targeting during the distribution estimation itself, e.g. by comparing (possibly on an independent/ validation sample) two candidate distribution estimators for the initial distribution and selecting the one which requires the least targeting (in the sense of having the smaller p-value under a score test).

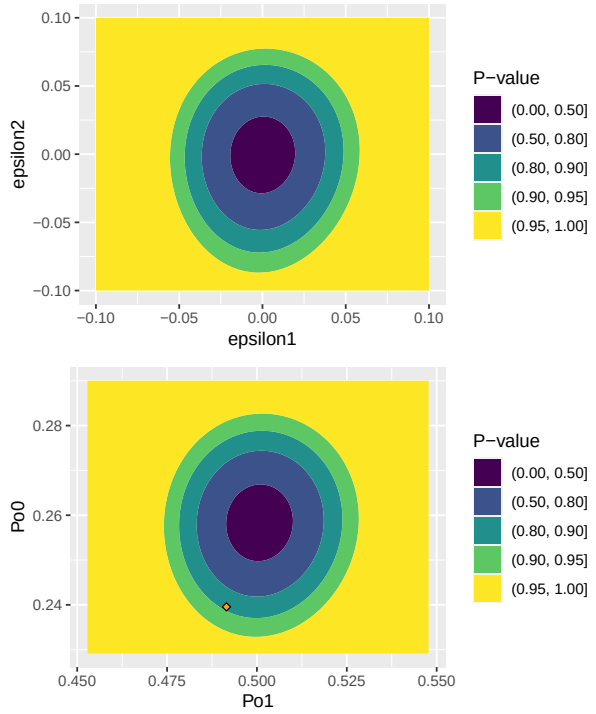


Figure 6.2: Graphical representation of $d = 2$ dimensional confidence set as in Example 4. This example used one dataset of size $n = 2000$ and is centred on a CV-TMLE ($k = 5$) initial targeted fit, with requisite models obtained using GBTs. The upper plot shows the confidence set over the parameter $\epsilon = (\epsilon_1, \epsilon_2)$, where the origin $\epsilon = (0, 0)$ represents the CV-TMLE estimate of the mean potential outcome vector $\boldsymbol{\Psi}(P_0) \equiv (P_0\{\mu_0(1, X)\}, P_0\{\mu_0(0, X)\})$. Colours indicate the P-value of the corresponding score statistic $nM_n(\hat{P}^*_\epsilon)$, according to an asymptotic χ^2_2 distribution. The lower plot shows the same confidence set represented on the estimand scale, i.e $\boldsymbol{\Psi}(\hat{P}^*_\epsilon)$, with the true estimand value $\boldsymbol{\Psi}(P_0)$ marked by the orange point.

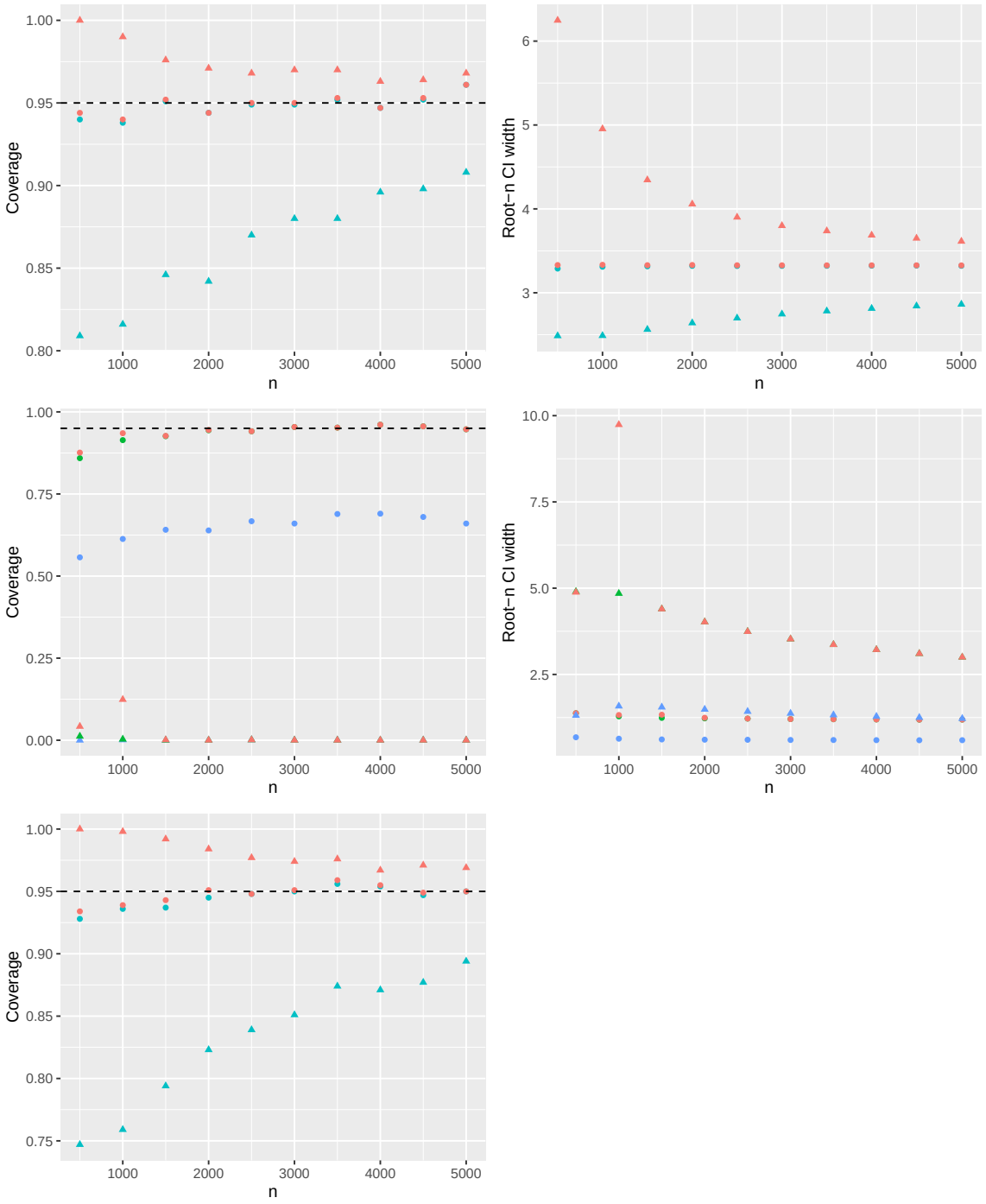


Figure 6.3: Coverage (left column) and median CI width (right column) vs. sample size n for 95% score and Wald based confidence sets for the ATE (top row), VTE (second row), and mean potential outcome vector (bottom row). For the latter estimand no CI width plot is shown since the width of a $d = 2$ dimensional confidence set is not well defined. CI widths have been scaled by $n^{1/2}$. Blue and red points represent Wald and score based intervals respectively. In the second row of plots, red and green points represent the score based interval using a 'local' and 'universal' parametric submodel respectively. All parametric submodels are centred on a TMLE initial targeted fit, with requisite models obtained using GAMs (circular points) and GBTs (triangular points).

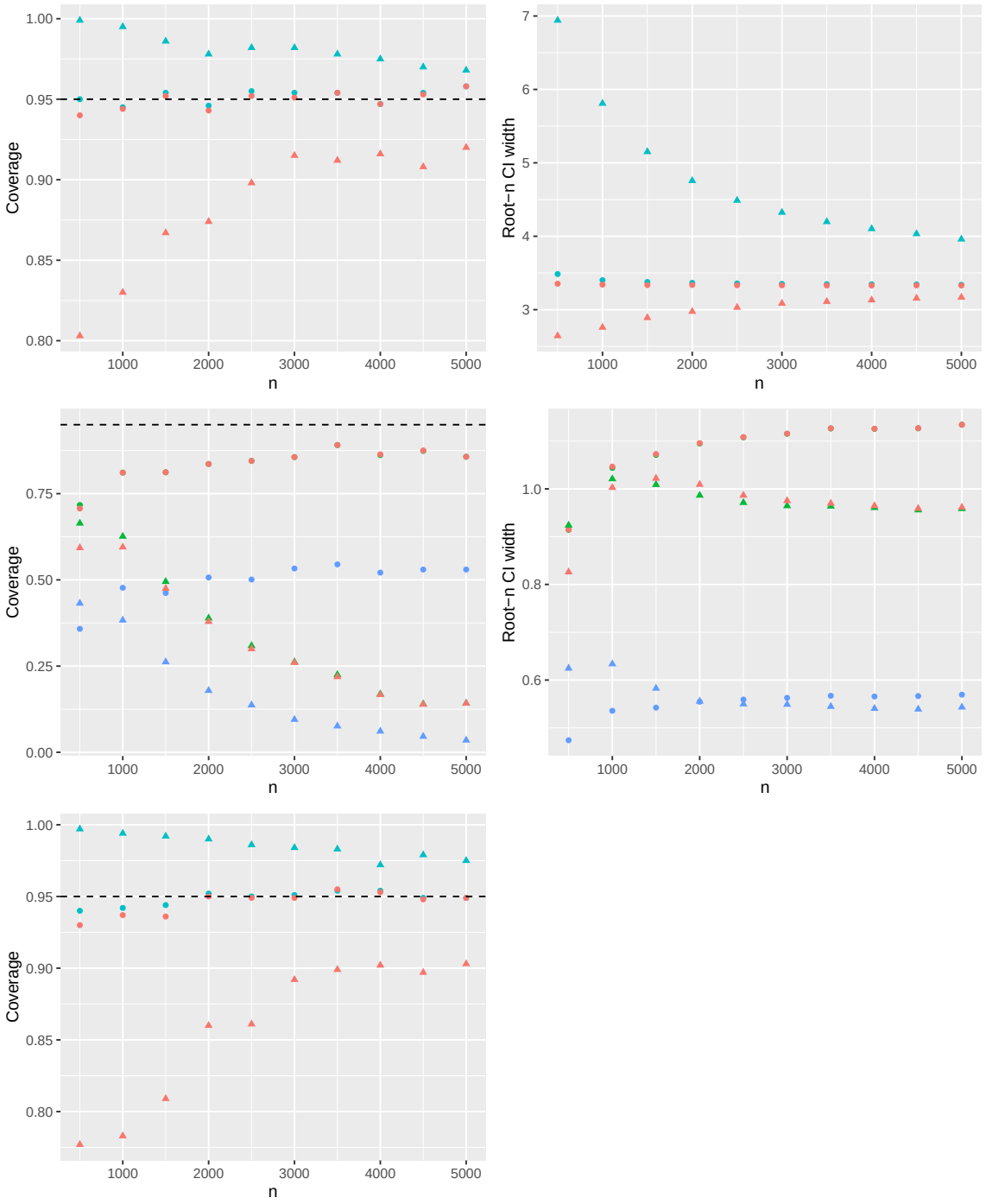


Figure 6.4: Coverage (left column) and median CI width (right column) vs. sample size n for 95% score and Wald based confidence sets for the ATE (top row), VTE (second row), and mean potential outcome vector (bottom row). For the latter estimand no CI width plot is shown since the width of a $d = 2$ dimensional confidence set is not well defined. CI widths have been scaled by $n^{1/2}$. Blue and red points represent Wald and score based intervals respectively. In the second row of plots, red and green points represent the score based interval using a 'local' and 'universal' parametric submodel respectively. All parametric submodels are centred on a CV-TMLE ($k = 5$) initial targeted fit, with requisite models obtained using GAMs (circular points) and GBTs (triangular points).

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT

T: +44 (0)20 7299 4646
F: +44 (0)20 7299 4656
www.lshtm.ac.uk

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	1804954	**Title**	Mr.
First Name(s)	Oliver		
Surname/Family Name	Hines		
Thesis Title	Assumption-Lean Inference for Causal and Statistical Questions in the Era of Machine Learning		
Primary Supervisor	Karla Diaz-Ordaz		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?			
When was the work published?			
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Choose an item.	Was the work subject to academic peer review?	Choose an item.

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	Not decided yet
Please list the paper's authors in the intended authorship order:	Oliver Hines, Karla Diaz-Ordaz, Stijn Vansteelandt
Stage of publication	**Not yet submitted**

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	This study was conceived by all authors. I carried out the mathematical research, computational simulations and writing of the manuscript under the supervision of the other authors.

SECTION E

Student Signature	Ohines
Date	14 Decmeber 2022

Supervisor Signature	Karla Diaz-Ordaz
Date	14 December 2022

Chapter 7

Optimally weighted average derivative effects

7.1 Introduction

Weighted average derivative effects (ADEs), also called average partial effects, were originally motivated for the estimation of parameters in index models (Härdle and Stoker, 1989; Powell et al., 1989; Newey and Stoker, 1993), a problem of substantial practical interest in econometrics, with additional uses in assessing the law of total demand in economics (Härdle et al., 1991) and in policy learning (Athey and Wager, 2021). ADEs are often estimated under a parametric model for the conditional expectation function $\mu(\boldsymbol{x}) = E(Y|\boldsymbol{X} = \boldsymbol{x})$, where Y is an outcome and $\boldsymbol{X}$ is a covariate vector. Using the parametrised model, the derivative, $d\mu(\boldsymbol{x})/d\boldsymbol{x}$, may be easily computed. This forms the basis for several estimators of the weighted ADE vector, $\boldsymbol{\theta} = E\{w(\boldsymbol{X})d\mu(\boldsymbol{X})/d\boldsymbol{X}\}$, where $w(\boldsymbol{X})$ is a weight function (Wooldridge and Zhu, 2020; Hirshberg and Wager, 2018).

The validity of parametric estimators, however, relies on correct specification of functional forms for $\mu(\boldsymbol{x})$, which may be hard in practice. For this reason, nonparametric approaches were developed based on the observation that, under standard assumptions, integration by parts yields, $\boldsymbol{\theta} = E\{\boldsymbol{l}(\boldsymbol{X})\mu(\boldsymbol{X})\}$, where

$$\boldsymbol{l}(\boldsymbol{x}) = -\frac{dw(\boldsymbol{x})}{d\boldsymbol{x}} - \frac{w(\boldsymbol{x})}{f(\boldsymbol{x})}\frac{df(\boldsymbol{x})}{d\boldsymbol{x}} \tag{7.1}$$

and $f(\boldsymbol{x})$ is the joint density of $\boldsymbol{X}$. This result is well studied in the literature and is used to obtain plug-in estimators for $\boldsymbol{\theta}$ where $f(\boldsymbol{x})$ is replaced with a kernel density estimate (Härdle and Stoker, 1989; Powell et al., 1989; Newey and Stoker, 1993; Cattaneo et al., 2010, 2013). The reliance on kernel methods, however, introduces complicated biases as the dimension of Z increases, due to the curse of dimensionality (Cattaneo et al., 2013).

Aside from the unitary weight $w(\boldsymbol{x}) = 1$, which implies $\boldsymbol{l}(\boldsymbol{x}) = -\{f(\boldsymbol{x})\}^{-1}df(\boldsymbol{x})/d\boldsymbol{x}$, the so-called density weight $w(\boldsymbol{x}) = f(\boldsymbol{x})$ is also popular (Powell et al., 1989; Cattaneo et al., 2010). These weights are designed to avoid inverse density weighting, since, for this choice, $\boldsymbol{l}(\boldsymbol{x}) = -2df(\boldsymbol{x})/d\boldsymbol{x}$.

In the current paper we extend the idea of selecting a weight function to facilitate inference. Rather than focusing on inference of the weighted ADE vector $\boldsymbol{\theta}$, we consider optimal weighting strategies to infer a single component, $\theta = \theta_j$. This problem is of particular interest since many practical analyses are interested in the main effect of a single continuous exposure, $A = X_j$, (e.g. dose, duration, frequency), whilst accounting for other covariates (i.e. excluding the jth) $\boldsymbol{Z} = \boldsymbol{X}_{-j}$, which may or may not be continuous. We illustrate such a setting in Section 7.8, with an applied example...

Our proposal shifts the emphasis away from specifying a weight function $w(\boldsymbol{x})$, towards specifying a function $l(\boldsymbol{x}) = l_j(\boldsymbol{x})$ that implies a well-defined weighted ADE under standard conditions. We derive functions $l(\boldsymbol{x})$ which are optimally efficient in the sense of delivering an estimand θ with minimal efficiency bound under the nonparametric model in the class of weighted ADE estimands (Newey and Stoker, 1993).

Our efficiency arguments build on similar optimal weighting strategies for weighted average treatment effects (ATEs) (Crump et al., 2006, 2009). Specifically, we show that two such weighted ADE estimands, which we call 'least squares estimands', are,

$$\psi = E\left\{\frac{\text{cov}(A, Y|\boldsymbol{Z})}{\text{var}(A|\boldsymbol{Z})}\right\} \tag{7.2}$$

and

$$\Psi = \frac{E\left\{\text{cov}(A, Y|\boldsymbol{Z})\right\}}{E\left\{\text{var}(A|\boldsymbol{Z})\right\}}, \tag{7.3}$$

when A is continuous and $d\mu(\boldsymbol{X})/dA$ exists; unlike weighted ADEs, these estimands remain well defined even when A is discrete or $\mu(\boldsymbol{x})$ is not differentiable. We motivate Ψ as an optimally efficient weighted ADE, in a sense described in Section 7.5. In Section 7.6, estimators for ψ and Ψ are derived which attain the efficiency bound under the nonparametric model. These estimators do not require estimation of $f(\boldsymbol{x})$, thus alleviating the aforementioned concerns regarding kernel density estimation.

The fact that ψ and Ψ are weighted ADEs is a surprising and novel contribution of our work. Both estimands have been studied in various other contexts and in Section 7.4 we illustrate their connection to non-parametric model projections (Chambaz et al., 2012; Buja et al., 2019). Additionally: Ψ appears in the context of partially-linear model estimators (Vansteelandt and Dukes, 2022; Newey and Robins, 2018); the numerator of Ψ is the 'generalised covariance measure' for conditional independence testing (Shah and Peters, 2018); ψ has been used to estimate the ADE under conditionally linear modelling assumptions (Hirshberg and Wager, 2018).

When $A \in \{0, 1\}$ is binary, then ψ and Ψ respectively identify the ATE and the propensity overlap weighted effect of A on Y when $\boldsymbol{Z}$ is sufficient to adjust for confounding (Crump et al., 2006, 2009; Robins et al., 2008; Li et al., 2018; Kallus, 2020). Overlap weights (also known as variance weights) are motivated for their utility in policy learning and for addressing limited overlap between the populations exposed to $A = 1$ and $A = 0$. Inspired by the binary setting, we propose estimators for ψ, based on the R-learner of the conditional ATE (Nie and Wager, 2021; Robinson, 1988).

7.2 Preliminaries

Suppose we have n iid observations, $(\boldsymbol{o}_1, ..., \boldsymbol{o}_n)$ of a random variable $\boldsymbol{O}$ distributed according to an unknown distribution P, such that $\boldsymbol{O}$ consists of $(Y, A, \boldsymbol{Z})$, where $Y \in \mathbb{R}$ is an outcome, $A \in \mathbb{R}$ is a continuous covariate of interest which we call an 'exposure' and $\boldsymbol{Z} \in \mathbb{R}^p$ is a p-dimensional vector of covariates. Define the weighted ADE, $\theta_w = E\{w(A, \boldsymbol{Z})\mu'(A, \boldsymbol{Z})\}$, where $\mu(A, \boldsymbol{Z}) \equiv E(Y|A, \boldsymbol{Z})$ has derivative w.r.t. A, denoted $\mu'(A, \boldsymbol{Z})$, and $w(A, \boldsymbol{Z}) \geq 0$ is a weight such that $k \equiv E\{w(A, \boldsymbol{Z})\}$ is positive and finite. We say the weight is 'normalised' when $k = 1$.

Define $w(\boldsymbol{Z}) \equiv E\{w(A, \boldsymbol{Z})|\boldsymbol{Z}\}$, which implies the existence of an 'exposure weight' $w(A|\boldsymbol{Z}) \geq 0$ such that $w(A, \boldsymbol{Z}) = w(A|\boldsymbol{Z})w(\boldsymbol{Z})$ and $E\{w(A|\boldsymbol{Z})|\boldsymbol{Z}\} = 1$. In this way, the contribution of the exposure to the weight $w(A, \boldsymbol{Z})$ can be considered separately. Also, by definition, $E\{w(\boldsymbol{Z})\} = k$.

Invoking regularity conditions, Powell et al. (1989) showed that the weighted ADE can be rewritten using integration by parts (see Appendix F). These conditions require that A is a continuous random variable and thus has a conditional density function, $f(a|\boldsymbol{z})$, given $\boldsymbol{Z} = \boldsymbol{z}$. We also require (C1) that the derivative of $w(a|\boldsymbol{z})f(a|\boldsymbol{z})$ w.r.t. a exists, (C2) that $w(a|\boldsymbol{z})f(a|\boldsymbol{z}) = 0$ for a on the boundary of the support of A, and (C3) that $f(a|\boldsymbol{z}) = 0$ implies $w(a|\boldsymbol{z}) = 0$. Under these conditions, $\theta_w = E\{w(\boldsymbol{Z})l(A|\boldsymbol{Z})Y\}$, where,

$$l(a|\boldsymbol{z}) = -w'(a|\boldsymbol{z}) - w(a|\boldsymbol{z})\frac{f'(a|\boldsymbol{z})}{f(a|\boldsymbol{z})} \tag{7.4}$$

and superscript prime denotes the derivative w.r.t. a. Just as the exposure weight $w(a|\boldsymbol{z})$ separates the contribution of a from the weight $w(a, \boldsymbol{z})$, the function $l(a|\boldsymbol{z})$ separates the contribution of a from (7.1), in the sense that $w(\boldsymbol{z})l(a|\boldsymbol{z})$ refers to a single component of (7.1).

7.3 Contrast functions

A key contribution of the current paper is an inversion of (7.4), providing an expression for the exposure weight $w(a|\boldsymbol{z})$ associated with certain functions $l(a|\boldsymbol{z})$, which we call contrast functions. The implication of Theorem 7 below is a duality between the contrast function and the exposure weight, allowing new weighted ADEs to be specified by their contrast functions rather than their weight functions. We define a contrast function as an arbitrary function $l(a|\boldsymbol{z})$ such that $E\{l(A|\boldsymbol{Z})|\boldsymbol{Z}\} = 0$ and $E\{l(A|\boldsymbol{Z})A|\boldsymbol{Z}\} = 1$. The function in (7.4) satisfies these two conditions under assumptions (C1), (C2) and (C3).

Theorem 7 *Let $l(a|\boldsymbol{z})$ be a contrast function and let $F(a|\boldsymbol{z})$ be the distribution function of A given $\boldsymbol{Z} = \boldsymbol{z}$. Assume that $f(a|\boldsymbol{x}) > 0$ for a on the convex support of A. Define,*

$$w(a|\boldsymbol{z}) = -\frac{E\{l(A|\boldsymbol{Z})|A \leq a, \boldsymbol{Z} = \boldsymbol{z}\}F(a|\boldsymbol{z})}{f(a|\boldsymbol{z})}. \tag{7.5}$$

For all differentiable functions $g(a, \boldsymbol{z})$,

$$E\{l(A|\boldsymbol{Z})g(A, \boldsymbol{Z})|\boldsymbol{Z}\} = E\{w(A|\boldsymbol{Z})g'(A, \boldsymbol{Z})|\boldsymbol{Z}\}$$

almost surely. Proof in Appendix F.

Corollary 7.1 *Let $w(\boldsymbol{Z}) \geq 0$ be a weight such that $k = E\{w(\boldsymbol{Z})\} \in (0, \infty)$. The case $g(A, \boldsymbol{Z}) = \mu(A, \boldsymbol{Z})$ implies,*

$$E\{w(\boldsymbol{Z})l(A|\boldsymbol{Z})Y\} = E\{w(A, \boldsymbol{Z})\mu'(A, \boldsymbol{Z})\} \tag{7.6}$$

where $w(a, \boldsymbol{z}) = w(\boldsymbol{z})w(a|\boldsymbol{z})$ with $w(a|\boldsymbol{z})$ given in Theorem 7.

Corollary 7.1 is particularly significant, since it allows weighted ADEs to be defined by the left hand side of (7.6), given limited restrictions on the function $l(A|\boldsymbol{Z})$. The exposure weight implied in (7.5) however, is not necessarily non-negative for an arbitrary contrast function. This is addressed in Lemma 4 which guarantees non-negativity when the contrast function is constructed from a monotonic function. Consider that by centring and scaling some function, $v(a, \boldsymbol{z})$, one may construct the contrast function,

$$l(a|\boldsymbol{z}) = \frac{v(a, \boldsymbol{z}) - E\{v(A, \boldsymbol{Z})|\boldsymbol{Z} = \boldsymbol{z}\}}{\text{cov}\{v(A, \boldsymbol{Z}), A|\boldsymbol{Z} = \boldsymbol{z}\}} \tag{7.7}$$

provided that $\text{cov}\{v(A, \boldsymbol{Z}), A|\boldsymbol{Z}\} \neq 0$ almost surely. It is easy to verify that this is a contrast function in the sense that $E\{l(A|\boldsymbol{Z})|\boldsymbol{Z}\} = 0$ and $E\{l(A|\boldsymbol{Z})A|\boldsymbol{Z}\} = 1$. The weighted ADE associated with this contrast function, according to (7.6), is,

$$\theta_w = E\left\{w(\boldsymbol{Z})\frac{\text{cov}\{v(A, \boldsymbol{Z}), Y|\boldsymbol{Z}\}}{\text{cov}\{v(A, \boldsymbol{Z}), A|\boldsymbol{Z}\}}\right\}. \tag{7.8}$$

In Section 7.4 we motivate estimands of this type where $v(a, \boldsymbol{z}) = a$. Trivially, when $v(a, \boldsymbol{z})$ is itself a contrast function then this expression recovers $\theta_w = E\{w(\boldsymbol{Z})v(A, \boldsymbol{Z})Y\}$ as on the left hand side of (7.6).

Lemma 4 (Sufficiency Condition for Weight Non-negativity) *Let $v(a, \boldsymbol{z})$ be a function which is monotonically increasing or decreasing in a (but is not everywhere constant), for a on the support of A. Then the contrast function in* (7.7) *implies a non-negative exposure weight as defined in* (7.5)*. Proof in Appendix F.*

We illustrate the connection between the contrast function and the exposure weight in three examples. Example 3 makes use of Corollary 7.1 and Lemma 4, to recover ψ and Ψ in (7.2) and (7.3). Both estimands have interesting connections to existing literature (see Section 7.4), however, here we illustrate that both estimands are weighted ADEs when A is continuous. This observation is a surprising and novel contribution of our work.

Example 1 (Average derivative effect (ADE)) *The ADE with* $w(A|\boldsymbol{Z}) = w(\boldsymbol{Z}) = 1$ *was originally proposed by Härdle and Stoker (1989). This results in the ADE,* $E\{\mu'(A, \boldsymbol{Z})\} = E\{l(A|\boldsymbol{Z})Y\}$*, where*

$$l(a|\boldsymbol{z}) = -\frac{f'(a|\boldsymbol{z})}{f(a|\boldsymbol{z})} = -\frac{d \log f(a|\boldsymbol{z})}{da} \tag{7.9}$$

is a contrast function. This estimand is normalised since $E\{w(A, \boldsymbol{Z})\} = 1$*.*

Example 2 (Density weighted ADE) *Originally proposed by Powell et al. (1989), the density weighted ADE sets the weight to the joint density of* $(A, \boldsymbol{Z})$*, i.e.* $w(a, \boldsymbol{z}) = f(a, \boldsymbol{z})$*. This results in the weight* $w(\boldsymbol{z}) = f(\boldsymbol{z})E\{f(A|\boldsymbol{Z})|\boldsymbol{Z} = \boldsymbol{z}\}$ *and exposure weight* $w(a|\boldsymbol{z}) = f(a|\boldsymbol{z})/E\{f(A|\boldsymbol{Z})|\boldsymbol{Z} = \boldsymbol{z}\}$*, where* $f(\boldsymbol{z})$ *denotes the density of* $\boldsymbol{Z}$*. The density weighted ADE is written* $E\{w(A, \boldsymbol{Z})\mu'(A, \boldsymbol{Z})\} = E\{w(\boldsymbol{Z})l(A|\boldsymbol{Z})Y\}$*, where*

$$l(a|\boldsymbol{z}) = -2\frac{f'(a|\boldsymbol{z})}{E\{f(A|\boldsymbol{Z})|\boldsymbol{Z} = \boldsymbol{z}\}}$$

is a contrast function. This estimand is normalised to $k = E\{f(A, \boldsymbol{Z})\}$*, hence the normalised density weighted ADE is obtained by rescaling* $w(a, \boldsymbol{z})$ *to* $f(a, \boldsymbol{z})/E\{f(A, \boldsymbol{Z})\}$ *in which case* $w(\boldsymbol{z}) = f(\boldsymbol{z})E\{f(A|\boldsymbol{Z})|\boldsymbol{Z} = \boldsymbol{z}\}/E\{f(A, \boldsymbol{Z})\}$ *and* $w(a|\boldsymbol{z})$ *and* $l(a|\boldsymbol{z})$ *are unchanged. This gives the normalised estimand,*

$$\frac{E\{f(A, \boldsymbol{Z})\mu'(A, \boldsymbol{Z})\}}{E\{f(A, \boldsymbol{Z})\}} = -2\frac{E\{f'(A, \boldsymbol{Z})Y\}}{E\{f(A, \boldsymbol{Z})\}}$$

Example 3 (Least Squares Estimands) *The main effect estimands* ψ *and* Ψ *in* (7.2) *and* (7.3) *are both weighted ADEs of the same type, in the sense that they share the same contrast function. We call these 'least squares estimands' due to their connection with the least squares problem (Section 7.4). Consider the construction in* (7.8)*, where* $v(a, \boldsymbol{z}) = a$*, implying the contrast function and estimand,*

$$l(a|\boldsymbol{z}) = \frac{a - E(A|\boldsymbol{Z} = \boldsymbol{z})}{\text{var}(A|\boldsymbol{Z} = \boldsymbol{z})}$$
$$\theta_w = E\left\{w(\boldsymbol{Z})\frac{\text{cov}(A, Y|\boldsymbol{Z})}{\text{var}(A|\boldsymbol{Z})}\right\}$$

where $w(\boldsymbol{Z})$ *is a non-negative weight. By Theorem 7, and using the total law of expectation, this contrast function implies the exposure weight,* $w(a|\boldsymbol{z}) = 0$ *if* $f(a|\boldsymbol{z}) = 0$ *and*

$$w(a|\boldsymbol{z}) = \frac{F(a|\boldsymbol{z})\{1 - F(a|\boldsymbol{z})\}}{f(a|\boldsymbol{z})\text{var}(A|\boldsymbol{Z} = \boldsymbol{z})}\{E(A|A > a, \boldsymbol{Z} = \boldsymbol{z}) - E(A|A \leq a, \boldsymbol{Z} = \boldsymbol{z})\}$$

otherwise. This exposure weight is non-negative by Lemma 4. The estimand $\theta_w = \psi$ *is recovered by setting* $w(\boldsymbol{z}) = 1$*. Setting* $w(\boldsymbol{z}) = \text{var}(A|\boldsymbol{Z} = \boldsymbol{z})$ *gives the unnormalised estimand* $\theta_w = E\{\text{cov}(A, Y|\boldsymbol{Z})\}$*, which is normalised by setting* $w(\boldsymbol{z}) = \text{var}(A|\boldsymbol{Z} = \boldsymbol{z})/E\{\text{var}(A|\boldsymbol{Z})\}$*, i.e. the variance weight (Robins et al., 2008), which recovers the estimand* $\theta_w = \Psi$*.*

7.4 Related literature

Here we describe how least squares estimands in Example 3 are connected to least squares projection, and discuss other related observations. Consider a semi-parametric partially linear model, of the type studied by Robinson (1988), where the model, $\mathcal{M}_a$ is the set of functions of the form $\omega(\boldsymbol{z}) + \beta a$, indexed by the infinite dimensional parameter (β, ω), where $\omega : \mathbb{R}^p \mapsto \mathbb{R}$ is a function and $\beta \in \mathbb{R}$ is a constant.

Our goal is to find the model projection $\tilde{\mu}_{(a)} \in \mathcal{M}_a$ that is 'nearest' to the unknown regression function $\mu(a, \boldsymbol{z})$ in the sense of minimising the mean squared remainder, $E[\{\mu(A, \boldsymbol{Z}) - \tilde{\mu}(A, \boldsymbol{Z})\}^2]$, where we assume that $E\{\mu(A, \boldsymbol{Z})^2\} < \infty$. This notion of model projection is considered by Neugebauer and van der Laan (2007) and Chambaz et al. (2012), who propose projections on to similar linear working models, and also

by Buja et al. (2019) who consider likelihood based projections. Projecting the regression function on to $\mathcal{M}_a$ gives,

$$\begin{aligned}\tilde{\mu}_{(a)}(A, \boldsymbol{Z}) &\equiv \underset{g \in \mathcal{M}_a}{\arg\min}\, E\left[\{\mu(A, \boldsymbol{Z}) - g(A, \boldsymbol{Z})\}^2\right] \\ &= \mu(\boldsymbol{Z}) + \Psi\{A - \pi(\boldsymbol{Z})\}\end{aligned}$$

where $\mu(\boldsymbol{z}) \equiv E(Y|\boldsymbol{Z} = \boldsymbol{z})$, and $\pi(\boldsymbol{z}) \equiv E(A|\boldsymbol{Z} = \boldsymbol{z})$. Hence we say the estimand Ψ is a 'least squares estimand' as it is the coefficient in a partially linear projection model which minimises the mean squared remainder. Crucially the model $\mathcal{M}_a$ is used to interpret the nonparametrically defined estimand Ψ, but we do not assume that the model is 'true', in the sense that we de not require that $\mu(a, \boldsymbol{z}) \in \mathcal{M}_a$.

The projection view of least squares estimands is further extended by considering the (more flexible) conditionally linear working model $\mathcal{M}_b \supseteq \mathcal{M}_a$, which is the set of functions of the form $\omega(\boldsymbol{z}) + \nu(\boldsymbol{z})a$, indexed by the infinite dimensional parameter (ν, ω), where $\nu : \mathbb{R}^p \mapsto \mathbb{R}$ is a function. Projecting the regression function on to $\mathcal{M}_b$, as above gives,

$$\begin{aligned}\tilde{\mu}_{(b)}(A, \boldsymbol{Z}) &\equiv \underset{g \in \mathcal{M}_b}{\arg\min}\, E\left[\{\mu(A, \boldsymbol{Z}) - g(A, \boldsymbol{Z})\}^2\right] \\ &= \mu(\boldsymbol{Z}) + \lambda(\boldsymbol{z})\{A - \pi(\boldsymbol{Z})\} \\ \lambda(\boldsymbol{z}) &\equiv \frac{\text{cov}(A, Y|\boldsymbol{Z} = \boldsymbol{z})}{\text{var}(A|\boldsymbol{Z} = \boldsymbol{z})}\end{aligned}$$

Hence the effects described in Example 3 are 'least squares estimands' since they represent weighted averages over the conditional least squares function $\lambda(\boldsymbol{z})$, i.e. they are of the form $\theta_w = E\{w(\boldsymbol{Z})\lambda(\boldsymbol{Z})\}$. This function has particular relevance to the setting where $A \in \{0, 1\}$ is a binary treatment indicator, since, in that setting it is generally true that $\mu(a, \boldsymbol{z}) \in \mathcal{M}_b$, and $\lambda(\boldsymbol{z}) = \mu(1, \boldsymbol{z}) - \mu(0, \boldsymbol{z})$ identifies the conditional ATE.

The estimand Ψ also appears in Vansteelandt and Dukes (2022) who consider inference for the constant term indexing $h(\mu(a, \boldsymbol{z})) \in \mathcal{M}_a$ where $h(.)$ represents a canonical link function. Rather than consider model projection explicitly, they set out desirable properites of an estimand under model mispecification, defining a nonparametric estimand which reduces to Ψ, in the case of an identity link. Similarly Ψ appears elsewhere in the partially linear model setting without reference to projection (Newey and Robins, 2018; Robins et al., 2008).

The fact that least squares estimands are weighted ADEs is a novel contribution of this work, however, relates closely to three observations in the literature. The first, by Banerjee (2007), is that an estimator of the vector ADE may be constructed by partitioning the support of $\boldsymbol{X}$ into disjoint bins, and applying a linear regression to each bin. An ADE estimate is obtained by taking the average of these regression coefficients, weighted by the number of observations in each bin. The second observation, by Buja et al. (2019) is that the ordinary least squares (OLS) coefficient may be interpreted as a weighted sum of 'slopes' between pairs of observations, without invoking differentiability. Thirdly, Hirshberg and Wager (2018) show that when the response function is conditionally linear, i.e. $\mu(a, \boldsymbol{z}) \in \mathcal{M}_b$, then ψ recovers the ADE. The key difference between Hirshberg and Wager (2018) and the current work is that our interpretation does not rely on any functional form for $\mu(a, \boldsymbol{z})$ beyond differentiability, rather we interpret ψ as an ADE with a certain kind of weighting.

7.5 Efficiency optimisation

In this Section we consider choosing weights $w(A, \boldsymbol{Z})$ to optimise estimation of $\theta_w = E\{w(A, \boldsymbol{Z})\mu'(A, \boldsymbol{Z})\}$. We draw heavily on inference methods that are based on efficient influence curves (ICs) under the nonparametric model, and recommend two recent tutorial papers for an introduction to these ideas (Hines et al., 2022; Fisher and Kennedy, 2020). In brief, an IC is a model-free, mean zero, functional of the true data distribution, which characterizes the sensitivity of a 'pathwise differentiable' estimand to small changes in the data distribution. As such, ICs are useful for constructing efficient estimators and for

understanding their asymptotic efficiency bounds. This efficiency bound is a property of the estimand itself and is given by the variance of the IC, which is finite.

According to Newey and Stoker (1993), when the weight function, $w(a, \boldsymbol{z}) = w(a|\boldsymbol{z})w(\boldsymbol{z})$, is known and (C1) and (C2) (Section 7.2) are assumed, the IC of θ_w is

$$\phi_{\theta,w}(\boldsymbol{o}) = w(\boldsymbol{z})l(a|\boldsymbol{z})\{y - \mu(a, \boldsymbol{z})\} + w(a, \boldsymbol{z})\mu'(a, \boldsymbol{z}) - \theta_w \tag{7.10}$$

where $l(a|\boldsymbol{z})$ is the contrast function in (7.4) and $\boldsymbol{o} = (y, a, \boldsymbol{z})$. In Examples 2 and 3 the exposure weight function is an unknown functional. However, the IC above, where the weight is known, offers some insight into optimal weight selection. Our approach to efficiency optimisation is analogous to that described in Crump et al. (2006, 2009) where optimal weights for the ATE are derived, and weights are assumed to be known. When the outcome is homoscedastic, they show that variance weights are optimal.

We derive a similar result here. Specifically we minimize the efficiency bound of an efficient estimator, $\hat{\theta}_w$, of the sample analogue of θ_w,

$$\theta_{w,S} = n^{-1}\sum_{i=1}^{n} w(a_i, \boldsymbol{z}_i)\mu'(a_i, \boldsymbol{z}_i)$$
$$n^{1/2}(\hat{\theta}_w - \theta_{w,S}) \xrightarrow{d} \mathcal{N}(0, V)$$
$$V = E\{w^2(\boldsymbol{Z})l^2(A|\boldsymbol{Z})\sigma^2(A, \boldsymbol{Z})\}$$

where $\sigma^2(A, \boldsymbol{Z}) = \text{var}(Y|A, \boldsymbol{Z})$. The efficiency bound with respect to $\theta_{w,S}$, rather than θ_w, is chosen so that the final two terms in (7.10) may be disregarded. Not only does this simplify the subsequent analysis, but these terms capture the difference between the ADE conditional on the sample distribution and that of the population as a whole, which depends on the unknown value of θ_w. I.e.,

$$n^{1/2}(\hat{\theta}_w - \theta_w) \xrightarrow{d} \mathcal{N}(0, V + U)$$
$$U = E[\{w(A, \boldsymbol{Z})\mu'(A, \boldsymbol{Z}) - \theta_w\}^2]$$

thus selecting weights to minimise $V + U$ is conceptually problematic as θ_w is itself the target estimand. Theorem 7, offers constraints on the contrast function under which appropriate weights are obtained. Our goal, therefore, is to minimise V subject to $E\{l(A|\boldsymbol{Z})|\boldsymbol{Z}\} = 0$, $E\{l(A|\boldsymbol{Z})A|\boldsymbol{Z}\} = 1$, and $E\{w(\boldsymbol{Z})\} = 1$, with the final constraint ensuring that the resulting estimands are normalised. Unnormalised estimands can be obtained by multiplying normalised estimands by a positive constant.

The optimal solution is given in general by Theorem 8, however there is no guarantee that the optimal exposure weight is non-negative. When Y is conditionally homoscedastic, i.e. $\sigma^2(a, \boldsymbol{z})$ does not depend on a, then the optimal exposure weight is non-negative, and when Y is homoscedastic, i.e. $\sigma^2(a, \boldsymbol{z})$ is constant, then the optimal solution recovers Ψ.

Theorem 8 (Optimally weighted ADE) *Minimizing the efficiency bound $V = n\text{var}\{\hat{\theta}_w - \theta_{w,S}\}$, subject to the constraints, $E\{l(A|\boldsymbol{Z})|\boldsymbol{Z}\} = 0$, $E\{l(A|\boldsymbol{Z})A|\boldsymbol{Z}\} = 1$, and $E\{w(\boldsymbol{Z})\} = 1$, has the solution*

$$l(a|\boldsymbol{z}) = \frac{b_1 - b_0 a}{(b_1^2 - b_0 b_2)\sigma^2(a, \boldsymbol{z})}$$
$$w(\boldsymbol{z}) = E\left(\frac{(b_1^2 - b_0 b_2)^2}{b_1^2 - 2b_0 b_1 c_1 + b_0^2 c_2}\right)^{-1} \frac{(b_1^2 - b_0 b_2)^2}{b_1^2 - 2b_0 b_1 c_1 + b_0^2 c_2}$$

where $b_n = b_n(\boldsymbol{z}) = E\{A^n\sigma^{-2}(A, \boldsymbol{Z})|\boldsymbol{Z} = \boldsymbol{z}\}$ and $c_n = c_n(\boldsymbol{z}) = E\{A^n|\boldsymbol{Z} = \boldsymbol{z}\}$. Proof in Appendix F.

Corollary 8.1 (Optimally weighted ADE under conditional homoscedasticity) *When Y is homoscedastic conditional on $\boldsymbol{Z}$, i.e. $\sigma^2(a, \boldsymbol{z}) = \sigma^2(\boldsymbol{z})$, where $\sigma^2(\boldsymbol{z}) = \text{var}(Y|\boldsymbol{Z} = \boldsymbol{z})$ then the estimand implied by Theorem 8 is*

$$\frac{E\{\text{cov}(A, Y|\boldsymbol{Z})/\sigma^2(\boldsymbol{Z})\}}{E\{\text{var}(A|\boldsymbol{Z})/\sigma^2(\boldsymbol{Z})\}}$$

For proof, observe that under conditional homoscedasticity, $b_n = c_n\sigma^{-2}(\boldsymbol{z})$. Furthermore, when Y is homoscedastic, i.e. $\sigma^2(a, \boldsymbol{z})$ is constant, then this optimal estimand recovers Ψ.

In practice, an estimate of the main effect of A on Y may be used to refute the null hypothesis that $Y \perp\!\!\!\perp A|\boldsymbol{Z}$. This hypothesis is hard to test, since any valid test has no power against any alternative (Shah and Peters, 2018). Under the null, one is in the setting of Corollary 8.1. There is no reason, however, to prefer a test based on the main effect of A on Y rather than the main effect of Y on A. The 'Generalized Covariance Measure' proposed by Shah and Peters (2018) uses $E\{\text{cov}(A, Y|\boldsymbol{Z})\}$ as a proxy to test for independence. This is also the optimal solution we propose when $\sigma^2(a, \boldsymbol{z})$ is constant and has the appealing property that it is invariant to swapping the roles of A and Y. The hardness of conditional independence testing arises since it is possible that A and Y are not independent, but $E\{\text{cov}(A, Y|\boldsymbol{Z})\} = 0$. These tests therefore have no power to test against such alternatives.

7.6 Estimation

7.6.1 Efficient estimators

Here we focus on efficient estimation of ψ and Ψ as in (7.2) and (7.3). Both are derivative effects of the type described in Example 3 and share the same contrast function. The ICs of ψ and Ψ respectively are,

$$\phi_\psi(\boldsymbol{o}) = \frac{\{a - \pi(\boldsymbol{z})\}}{\beta(\boldsymbol{z})}\left[y - \mu(\boldsymbol{z}) - \lambda(\boldsymbol{z})\{a - \pi(\boldsymbol{z})\}\right] + \lambda(\boldsymbol{z}) - \psi$$
$$\phi_\Psi(\boldsymbol{o}) = \frac{\{a - \pi(\boldsymbol{z})\}}{E\{\beta(\boldsymbol{Z})\}}\left[y - \mu(\boldsymbol{z}) - \Psi\{a - \pi(\boldsymbol{z})\}\right]$$

where $\beta(\boldsymbol{z}) = \text{var}(A|\boldsymbol{Z} = \boldsymbol{z})$. These ICs may be used to construct efficient estimating equation estimators of ψ and Ψ by setting (an estimate of) the sample mean IC to zero. In the current setting, this strategy is equivalent to the so-called one-step correction which we outline in Appendix F. For ψ and Ψ, we thus obtain the estimators

$$\hat{\psi} = n^{-1}\sum_{i=1}^{n}\frac{\{a_i - \hat{\pi}(\boldsymbol{z}_i)\}}{\hat{\beta}(\boldsymbol{z}_i)}[y_i - \hat{\mu}(\boldsymbol{z}_i) - \hat{\lambda}(\boldsymbol{z}_i)\{a_i - \hat{\pi}(\boldsymbol{z}_i)\}] + \hat{\lambda}(\boldsymbol{z}_i)$$
$$\hat{\Psi} = \frac{\sum_{i=1}^{n}\{a_i - \hat{\pi}(\boldsymbol{z}_i)\}\{y_i - \hat{\mu}(\boldsymbol{z}_i)\}}{\sum_{i=1}^{n}\{a_i - \hat{\pi}(\boldsymbol{z}_i)\}^2}.$$

where superscript hat denotes fitted models obtained from an independent sample. In practice, a cross-fitting approach may be used to obtain the fitted models and evaluate the estimators using a single sample (Chernozhukov et al., 2018; Zheng and van der Laan, 2011). Theorem 9 below demonstrates that $\hat{\psi}$ is regular asymptotically linear when

(A1) The propensity score error, $||\pi - \hat{\pi}||$ is $o_P(n^{-1/4-\delta})$ for some $\delta \geq 0$.

(A2) The outcome error, $||\mu - \hat{\mu}||$ is $o_P(n^{-1/4+\delta})$.

(A3) The product of $||\lambda - \hat{\lambda}||$ and $||\beta - \hat{\beta}||$ is $o_P(n^{-1/2})$.

where, for some function $f(\boldsymbol{z})$, we denote the $L_2(P)$ norm $||f|| \equiv E\{f^2(\boldsymbol{Z})\}^{1/2}$. Similarly, Theorem 10 demonstrates that $\hat{\Psi}$ is regular asymptotically linear under (A1) and (A2).

Theorem 9 *Under (A1), (A2), (A3), and regularity assumptions given in Appendix F, $\hat{\psi}$ is regular asymptotically linear with IC, $\phi_\psi(\boldsymbol{O})$, and hence $\hat{\psi}$ converges to ψ in probability, and $n^{1/2}(\hat{\psi} - \psi)$ converges in distribution to a mean-zero normal random variable with variance $E\{\phi_\psi^2(\boldsymbol{O})\}$.*

Theorem 10 *Under (A1), (A2), and regularity assumptions given in Appendix F, $\hat{\Psi}$ is regular asymptotically linear with IC, $\phi_\Psi(\boldsymbol{O})$, and hence $\hat{\Psi}$ converges to Ψ in probability, and $n^{1/2}(\hat{\Psi} - \Psi)$ converges in distribution to a mean-zero normal random variable with variance $E\{\phi_\Psi^2(\boldsymbol{O})\}$.*

We remark that the estimator $\hat{\psi}$ requires modelling the functions $\beta(.)$ and $\lambda(.)$, whereas the estimator $\hat{\Psi}$ does not, with Theorem 9 requiring (A3) to control the error in estimating $\beta(.)$ and $\lambda(.)$. This distinction makes Ψ generally more straightforward to efficiently estimate than ψ. Assumption (A3) also demonstrates that $\hat{\psi}$ is 'rate double robust', in the sense that one may trade-off accuracy in $\hat{\lambda}(.)$ and $\hat{\beta}(.)$. In other words, the $\lambda(.)$ estimator can converge slowly, as long as the $\beta(.)$ estimator converges sufficiently quickly, and vice-versa.

Similar rate double robustness has been demonstrated previously for example in the augmented inverse probability weighted (AIPW) estimator of the ATE (Robins, 1994), where one can trade-off accuracy in the propensity score estimator and outcome estimator. With regards to (A1) and (A2), a similar robustness is observed, since the $\mu(.)$ estimator can converge slowly, as long as the $\pi(.)$ estimator converges sufficiently quickly. The converse, however, is not true, since (A1) requires that $\hat{\pi}(.)$ converges to $\pi(.)$ at least at $n^{1/4}$ rate.

In the setting where $A \in \{0, 1\}$ is a binary exposure then $\hat{\psi}$ reduces to the well-known augmented inverse probability weighted (AIPW) estimator of the ATE (Robins, 1994), since $\lambda(\boldsymbol{z}) = \mu(1, \boldsymbol{z}) - \mu(0, \boldsymbol{z})$ is the conditional ATE, $\beta(\boldsymbol{z}) = \pi(\boldsymbol{z})\{1 - \pi(\boldsymbol{z})\}$, and $\mu(\boldsymbol{z}) + \lambda(\boldsymbol{z})\{a - \pi(\boldsymbol{z})\} = \mu(a, \boldsymbol{z})$, hence one obtains the AIPW estimator,

$$\hat{\psi} = n^{-1} \sum_{i=1}^{n} \frac{\{a_i - \hat{\pi}(\boldsymbol{z}_i)\}}{\hat{\pi}(\boldsymbol{z}_i)\{1 - \hat{\pi}(\boldsymbol{z}_i)\}} \{y_i - \hat{\mu}(a_i, \boldsymbol{z}_i)\} + \hat{\mu}(1, \boldsymbol{z}_i) - \hat{\mu}(0, \boldsymbol{z}_i)$$

It follows that $\hat{\psi}$, and its asymptotic distribution in Theorem 9, represent a generalisation of the AIPW estimator to the setting of continuous exposures. Indeed the estimation approaches which we consider in Section 7.6.3 estimate the ATE when A is replaced with a binary exposure. The estimator $\hat{\Psi}$ has been studied before in the context of the 'partialling out' estimator of Robinson (1988) (see e.g. Newey and Robins (2018); Vansteelandt and Dukes (2022)).

7.6.2 Nuisance function estimators

The estimator $\hat{\Psi}$ is indexed by the choice of estimator for $\hat{\mu}(.)$ and $\hat{\pi}(.)$, with the estimator $\hat{\psi}$ additionally indexed by the choice of estimator for $\hat{\lambda}(.)$ and $\hat{\beta}(.)$. Generally, we are not constrained to any particular learning method, making these estimators amenable to data adaptive/ machine learning estimation of these working models.

Data adaptive regression algorithms are well developed for the regularised regression of an observed variable on to a set of explanatory variables, e.g. for the functions $\mu(.)$, and $\pi(.)$ in the present context, which can be estimated by respectively regressing Y and A on $\boldsymbol{Z}$. For $\lambda(.)$ and $\beta(.)$, however, estimation methods are less well developed, and we propose so-called meta-learning approaches, which estimate $\lambda(.)$ and $\beta(.)$ by solving a series of regression problems.

In the setting where $A \in \{0, 1\}$ is binary, $\lambda(.)$ represents the conditional ATE, estimation of which is a highly active area of research, with an emphasis on flexible machine learning methods (Abrevaya et al., 2015; Athey and Imbens, 2016; Nie and Wager, 2021; Kallus et al., 2018; Wager and Athey, 2018; Künzel et al., 2019; Kennedy, 2020). Estimation of the variance function $\beta(.)$, is also of interest in the literature with applications in constructing confidence intervals for the mean function $\pi(.)$ and for estimating signal to noise ratios (Shen et al., 2020; Wang et al., 2008; Cai et al., 2009; Verzelen and Gassiat, 2018). We consider two approaches to estimating $\lambda(.)$ and $\beta(.)$.

The first approach, which we shall refer to as the direct learning approach, involves decomposing $\lambda(.)$ and $\beta(.)$ into functions of conditional expectations, each of which can be estimated using standard regression methods, with the estimates combined to produce $\hat{\lambda}(.)$ and $\hat{\beta}(.)$. Specifically, letting $\hat{E}\{YA|\boldsymbol{Z} = \boldsymbol{z}\}$ and $\hat{E}\{A^2|\boldsymbol{Z} = \boldsymbol{z}\}$ denote estimates obtained by respectively regressing YA and A^2 on $\boldsymbol{Z}$, we define nuisance estimators

$$\hat{\lambda}(\boldsymbol{z}) = \frac{\hat{E}\{YA|\boldsymbol{Z} = \boldsymbol{z}\} - \hat{\mu}(\boldsymbol{z})\hat{\pi}(\boldsymbol{z})}{\hat{E}\{A^2|\boldsymbol{Z} = \boldsymbol{z}\} - \hat{\pi}^2(\boldsymbol{z})} \tag{7.11}$$

$$\hat{\beta}(\boldsymbol{z}) = \hat{E}\{A^2|\boldsymbol{Z} = \boldsymbol{z}\} - \hat{\pi}^2(\boldsymbol{z}) \tag{7.12}$$

The issue with this direct approach, however, is that whilst regularization methods can be used to control the smoothness of each individual regression function, there is no guarantee on the smoothness of $\hat{\lambda}(.)$ and $\hat{\beta}(.)$. In practice these may be erratic functions due to artefacts of the regularization of the individual regression functions. Additionally, there is no guarantee that $\hat{\beta}(.)$, which also represents the denominator of $\hat{\lambda}(.)$, is greater than zero. This motivates an alternative approach where the complexities of $\hat{\lambda}(.)$ and $\hat{\beta}(.)$ can be controlled directly, and one can ensure that $\hat{\beta}(\boldsymbol{z}) > 0$.

The second approach, which we shall refer to as the quasi-oracle learning approach, is a meta-learning method based on the R-learner of the conditional ATE (Nie and Wager, 2021; Robinson, 1988). In our description we make use of the following Lemma.

Lemma 5 *Let $\boldsymbol{O} = (\boldsymbol{U}, V, W)$ be a random variable consisting of $\boldsymbol{U} \in \mathbb{R}^d$, $V \in \mathbb{R}$, and $W \in \mathbb{R}$ with $W > 0$ almost surely. Let $\mathcal{F}$ denote the set of functions $g : \mathbb{R}^d \mapsto \mathbb{R}$. Then*

$$\frac{E(V|\boldsymbol{U} = \boldsymbol{u})}{E(W|\boldsymbol{U} = \boldsymbol{u})} = \underset{g(\boldsymbol{u})\in\mathcal{F}}{\arg\min}\, E\left[W\left\{\frac{V}{W} - g(\boldsymbol{U})\right\}^2\right] \tag{7.13}$$

where we say the part in the square brackets is to equal 0 when $W = 0$ and requisite moments of $\boldsymbol{O}$ are assumed to be finite. See Appendix F for proof.

This Lemma connects the problem of estimating a ratio of conditional expectations, with minimisation of a weighted mean squared error. For example, in the setting where $W = 1$, this Lemma then the right hand side of (7.13) reduces to the familiar mean squared error. Similarly, the left hand side of (7.13) recovers $\lambda(\boldsymbol{z})$ in the setting where $\boldsymbol{U} = \boldsymbol{Z}$, $W = \{A - \pi(\boldsymbol{Z})\}^2$ and $V = \{A - \pi(\boldsymbol{Z})\}\{Y - \mu(\boldsymbol{Z})\}$.

This suggests that an estimator for $\lambda(.)$ is obtained by regressing $\{Y - \mu(\boldsymbol{Z})\}/\{A - \pi(\boldsymbol{Z})\}$ on $\boldsymbol{Z}$ with weights $\{A - \pi(\boldsymbol{Z})\}^2$. We call this an 'oracle' estimator for $\lambda(.)$, since it is the regression problem that we would like to solve if these outcomes and weights were known. Instead, the R-learner of the conditional ATE essentially mimics the oracle learner by first estimating $\mu(.)$ and $\pi(.)$ using an independent sample, then using these to estimate the unobserved outcomes and weights. This method is referred to as 'quasi-oracle' since it the error bound for the $\lambda(.)$ estimator may decay faster those of the $\mu(.)$ and $\pi(.)$ estimators (Nie and Wager, 2021).

We propose a similar approach to learning $\beta(.)$, which appears in our target estimator $\hat{\psi}$ as an inverse weight. Such inverse weighting may be problematic when $\beta(\boldsymbol{z}_i)$ is in truth small, since small errors in $\hat{\beta}(\boldsymbol{z}_i)$ could result in large differences in the value of $1/\hat{\beta}(\boldsymbol{z}_i)$. This extreme weighting problem is well documented in the context of inverse probability weighting estimators of the ATE (Kang and Schafer, 2007). Concerns regarding extreme weights, however, could be mitigated by regularizing the function $1/\hat{\beta}(.)$ rather than $\hat{\beta}(.)$ itself. For this reason we consider that the left hand side of (7.13) recovers $1/\beta(\boldsymbol{z})$ in the setting where $\boldsymbol{U} = \boldsymbol{Z}$, $V = 1$ and $W = \{A - \pi(\boldsymbol{Z})\}^2$.

This suggests that an oracle estimator for $1/\beta(.)$ is obtained by regressing $\{A - \pi(\boldsymbol{Z})\}^{-2}$ on $\boldsymbol{Z}$ with weights $\{A - \pi(\boldsymbol{Z})\}^2$. Like the R-learner, we propose a quasi-oracle learner which mimics this oracle learner by first estimating $\pi(.)$ using an independent sample, then estimating the oracle outcomes and weights.

7.6.3 Proposed algorithms

The proposed working function estimators are implemented in Algorithms 1 and 2 below. The latter uses a cross fitting regime to ensure that $\hat{\mu}(\boldsymbol{z}_i), \hat{\pi}(\boldsymbol{z}_i), \hat{\lambda}(\boldsymbol{z}_i)$, and $\hat{\beta}(\boldsymbol{z}_i)$ are obtained using working models which are constructed from a dataset that does not include the ith observation. This is useful in controlling the so-called empirical process term (Chernozhukov et al., 2018; Zheng and van der Laan, 2011).

Algorithms 1 and 2 return the estimates $\{\hat{\pi}_i\}_{i=1}^n$, $\{\hat{\mu}_i\}_{i=1}^n$, $\{\hat{\lambda}_i\}_{i=1}^n$, and $\{\hat{\beta}_i\}_{i=1}^n$, which can be used to

obtain

$$\hat{\psi} = n^{-1} \sum_{i=1}^{n} \frac{(a_i - \hat{\pi}_i)}{\hat{\beta}_i} \{y_i - \hat{\mu}_i - \hat{\lambda}_i(a_i - \hat{\pi}_i)\} + \hat{\lambda}_i$$
$$\hat{\Psi} = \frac{\sum_{i=1}^{n}(a_i - \hat{\pi}_i)(y_i - \hat{\mu}_i)}{\sum_{i=1}^{n}(a_i - \hat{\pi}_i)^2},$$

with variances respectively estimated by $n^{-2}\sum_{i=1}^{n}\phi_{\psi,i}^2$ and $n^{-2}\sum_{i=1}^{n}\phi_{\Psi,i}^2$ where

$$\phi_{\psi,i} = \frac{(a_i - \hat{\pi}_i)}{\hat{\beta}_i} \{y_i - \hat{\mu}_i - \hat{\lambda}_i(a_i - \hat{\pi}_i)\} + \hat{\lambda}_i - \hat{\psi}$$
$$\phi_{\Psi,i} = \frac{(a_i - \hat{\pi}_i)}{\hat{\eta}} \{y_i - \hat{\mu}_i - \hat{\Psi}(a_i - \hat{\pi}_i)\}$$

and $\hat{\eta} \equiv n^{-1}\sum_{i=1}^{n}(a_i - \hat{\pi}_i)^2$ is an estimate of $E\{\beta(\boldsymbol{Z})\}$. We note that where the algorithms require regression estimates to be 'fitted', any suitable regression/ machine learning method can be used.

Both algorithms are also indexed by the choice of learner for $\hat{\lambda}(.)$ and $\hat{\beta}(.)$ in steps 2 and 3 of each algorithm respectively, with the substeps marked (A) and (B) referring to the direct, and quasi-oracle approaches. Note that the quasi-oracle methods do not themselves use sample splitting to learn the unobserved outcomes and weights, due to the impracticality of excessive sample splitting in finite samples. These steps do not need to be carried out for inference of $\hat{\Psi}$ only.

For estimators such as $\hat{\Psi}$, it has been suggested that faster convergence rates may be achieved through additional sample splitting to ensure that $\hat{\mu}(\boldsymbol{z}_i)$ and $\hat{\pi}(\boldsymbol{z}_i)$ are obtained from two different and independent datasets, both of which do not contain the ith observation (Newey and Robins, 2018). We do not consider such 'double cross fitting' here, since extensions, to estimate ψ, would require significant additional sample splitting to estimate $\{\hat{\lambda}_i\}_{i=1}^n$ and $\{\hat{\beta}_i\}_{i=1}^n$, which may be impractical in finite samples.

Algorithm 1 - Without sample splitting

(1) Fit $\hat{\mu}(\boldsymbol{z})$ and $\hat{\pi}(\boldsymbol{z})$. Use these fitted models to obtain $\hat{\mu}_i \equiv \hat{\mu}(\boldsymbol{z}_i)$ and $\hat{\pi}_i \equiv \hat{\pi}(\boldsymbol{z}_i)$.

(2) (A) Fit $\hat{E}\{YA|\boldsymbol{Z} = \boldsymbol{z}\}$ and $\hat{E}\{A^2|\boldsymbol{Z} = \boldsymbol{z}\}$ and use these to construct $\hat{\lambda}(\boldsymbol{z})$ and $\hat{\beta}(\boldsymbol{z})$ as in (7.11) and (7.12). Or (B) obtain $\hat{\lambda}(\boldsymbol{z})$ and $1/\hat{\beta}(\boldsymbol{z})$ respectively by regressing $\{Y - \hat{\mu}(\boldsymbol{Z})\}/\{X - \hat{\pi}(\boldsymbol{Z})\}$ and $\{X - \hat{\pi}(\boldsymbol{Z})\}^{-2}$ on $\boldsymbol{Z}$ with weights $\{X - \hat{\pi}(\boldsymbol{Z})\}^2$ using all the data. After doing (A) or (B), use the fitted models to obtain $\hat{\lambda}_i \equiv \hat{\lambda}(\boldsymbol{z}_i)$ and $\hat{\beta}_i \equiv \hat{\beta}(\boldsymbol{z}_i)$.

Algorithm 2 - With sample splitting

(1) Split the data into K folds.

(2) **For** each fold k: Fit $\hat{\mu}(\boldsymbol{z})$ and $\hat{\pi}(\boldsymbol{z})$ using the data set excluding fold k. Use these fitted models to obtain $\hat{\mu}_i \equiv \hat{\mu}(\boldsymbol{z}_i)$ and $\hat{\pi}_i \equiv \hat{\pi}(\boldsymbol{z}_i)$ for i in fold k.

(3) (A) Fit $\hat{E}\{YA|\boldsymbol{Z} = \boldsymbol{z}\}$ and $\hat{E}\{A^2|\boldsymbol{Z} = \boldsymbol{z}\}$ using the data set excluding fold k. and use these to construct $\hat{\lambda}(\boldsymbol{z})$ and $\hat{\beta}(\boldsymbol{z})$ as in (7.11) and (7.12). Or (B) obtain $\hat{\lambda}(\boldsymbol{z})$ and $1/\hat{\beta}(\boldsymbol{z})$ respectively by regressing $\{Y - \hat{\mu}(\boldsymbol{Z})\}/\{X - \hat{\pi}(\boldsymbol{Z})\}$ and $\{X - \hat{\pi}(\boldsymbol{Z})\}^{-2}$ on $\boldsymbol{Z}$ with weights $\{X - \hat{\pi}(\boldsymbol{Z})\}^2$ using the data set excluding fold k. After doing (A) or (B), use the fitted models to obtain $\hat{\lambda}_i \equiv \hat{\lambda}(\boldsymbol{z}_i)$ and $\hat{\beta}_i \equiv \hat{\beta}(\boldsymbol{z}_i)$ for i in fold k. **End for.**

7.7 Simulation study

In our simulation study we compared Algorithms 1 and 2 for estimating Ψ and Algorithms 1A,1B, 2A and 2B for estimating ψ on generated data in finite samples, using $K = 5$ fold sample splitting. We generated

1000 datasets of size $n \in \{500, 1000, ..., 4000\}$ from the following structural equation model

$$\begin{aligned} Z_1, Z_2, Z_3 &\sim \text{Uniform}(-1, 1) \\ \epsilon_1, \epsilon_2 &\sim \mathcal{N}(0, 1) \\ A &= Z_1 + 0.5Z_1^3 - 2Z_2^2 + Z_1^2 Z_2 + (1 + Z_1^2)\epsilon_1 \\ Y &= A(1 + Z_1 - Z_1^2 - 0.5Z_2^2) - Z_1^2 Z_2 + Z_2 Z_3 + \epsilon_2 \end{aligned}$$

with the least squares estimands taking the true values $\psi = 0.5$ and $\Psi = 107/294 \approx 0.36$.

For each dataset, $\hat{\psi}$ and $\hat{\Psi}$ were estimated along with their variance and associated Wald based (95%) confidence intervals. Two regression model approaches were considered, the first used generalised additive models, as implemented through the `mgcv` package in R (Wood et al., 2016). These models use flexible spline smoothing including pairwise interaction terms. The second regression modelling approach used random forest learners available through the `ranger` package in R (Wright and Ziegler, 2017).

Figure 7.1 shows empirical estimates of the empirical bias and empirical variance of $\hat{\psi}$ and $\hat{\Psi}$ scaled by $n^{1/2}$ and n respectively, as well as the empirical coverage probability of a Wald based 95% confidence-interval. Comparing Algorithms 1 and 2 (i.e. no sample splitting vs sample splitting) for the estimation of $\hat{\psi}$, we notice that sample splitting generally improves confidence interval coverage.

Additionally, for estimation of ψ the quasi-oracle approach (Algorithm B) outperforms the direct approach (Algorithm A) in terms of reduced bias, variance and improved CI coverage. This is achieved since the quasi-oracle approach controls the smoothess of $\hat{\lambda}(.)$ and the inverse weights $1/\hat{\beta}(.)$, whereas Algorithm A does not, leading to the possiblity of extreme inverse weighting in the estimator. On the basis of these results, we recommend Algorithm 2B for estimation of ψ and Algorithm 2 for estimation of Ψ.

7.8 Warfarin dose example

We illustrate the proposed estimators using the International Warfarin Pharmacogenetics Consortium (2009) dataset, which has also been reanalysed several times in literature on dynamic treatment rule estimation (Schulz and Moodie, 2021; Wallace et al., 2018; Chen et al., 2016). The data consists of $n = 1732$ patients receiving Warfarin therapy, which is a commonly prescribed anticoagulant used to treat thrombosis and thromboembolism. We consider least squares estimands for the effect of Warfarin dose (A) on international normalised ratio (INR) (Y), which is a measure of blood clotting function, given 13 other patient characteristics ($\boldsymbol{Z}$), including genetic data, as described in International Warfarin Pharmacogenetics Consortium (2009).

Fitted models were obtained using the Super Learner (van der Laan et al., 2007), an ensemble learning method, implemented in the `SuperLearner` package in R. This used 20 cross validation folds, and a 'learner library' containing various routines (`glm`, `glmnet`, `gam`, `xgboost`, `ranger`). Additional results which use the 'discrete' Super Learner for model fitting are presented in Appendix F. The discrete Super Learner selects the regression algorithm in the learner library which minimises a cross validated estimate of e.g. the mean squared error loss, whereas the Super Learner minmizes the same loss by taking a convex combination of learners. For the sample splitting algorithms (Algorithm 2), $K = 20$ folds were chosen (between 10 to 20 folds is typical for cross-fitting procedures).

The results, presented in Table 7.1, suggest that increased Warfarin does is associated with an increase in INR. We see that the estimators for Ψ tend to give results with narrower confidence intervals, and commensurately smaller Wald based p-values for the estimand null, as expected from the efficiency arguments presented in Section 7.5. Additionally, the estimators for ψ, which use the R-learner for conditional effect estimation (Algorithms 1B and 2B) give more credible estimates than those that use the direct approach (Algorithms 1A and 2A), in the sense that they are of a similar order of magnitude to the Ψ estimates. Moreover, we see that sample splitting leads to more credible estimates, compared with no sample splitting, as evident in Algorithms 2A versus 1A. This difference is because sample splitting helps to control for overfitting of the functional estimators.

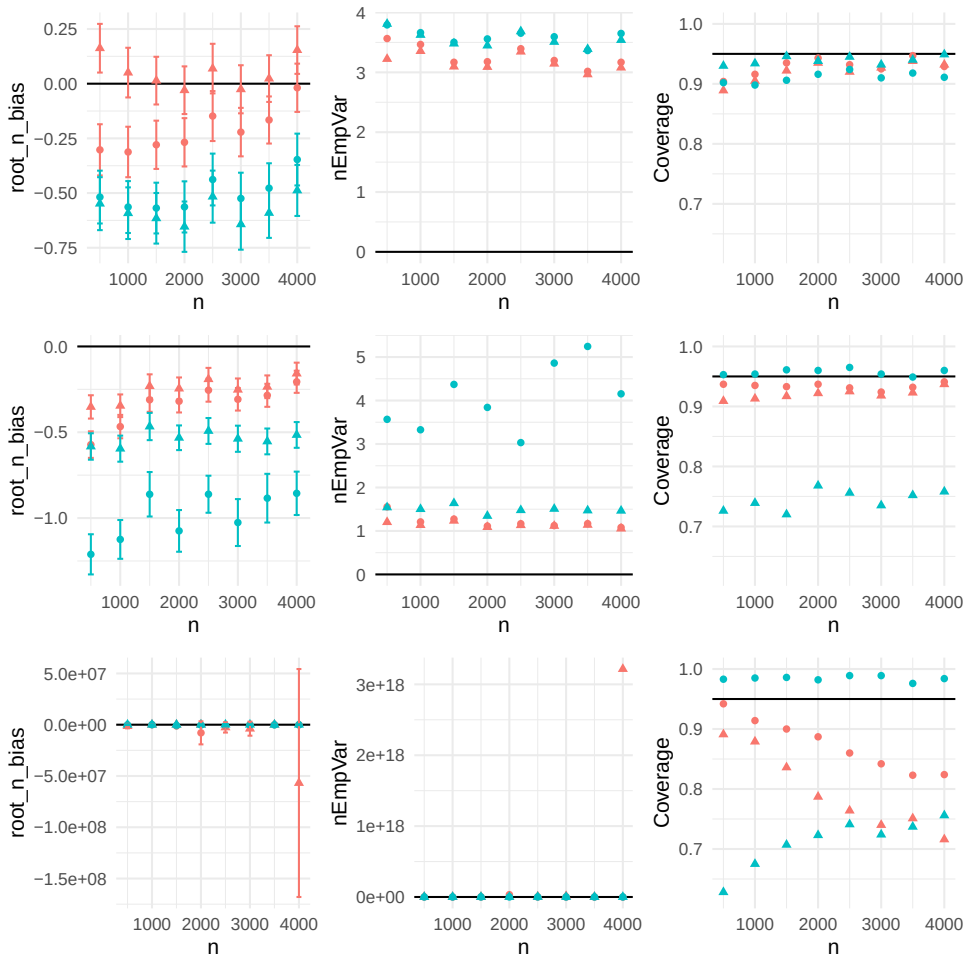


Figure 7.1: Sample size against Bias, variance and 95% Wald CI coverage for $\hat{\Psi}$ (top row), $\hat{\psi}$ using the quasi-oracle approach (middle row), and $\hat{\psi}$ using the direct approach (bottom row). Red and blue points indicate that working models are fitted using generalised additive modelling and random forests respectively. Circular and triangular points indicate that the estimators were fitted with and without sample splitting respectively (i.e. Algorithm 2 vs. Algorithm 1). We highlight that the y-axis limits are not the same for each row of the bias and variance plots.

Table 7.1: Least squares estimands applied to IWPC data. Values indicate point estimates, given in INR/(mg/week), with 95% Wald confidence intervals given in parentheses. P-values represent those obtained from a Wald based test of the null hypothesis that the estimand is 0.

Estimand	Algorithm	Result
Ψ	1	2.01×10^{-3} (0.731×10^{-3},3.28×10^{-3}) p=0.002
Ψ	2	1.91×10^{-3} (0.70×10^{-3},3.12×10^{-3}) p=0.002
ψ	1A	-6.39 (-19.4,6.65) p=0.33
ψ	2A	0.611 (-0.664,1.89) p=0.35
ψ	1B	1.59×10^{-3} (-0.0660×10^{-3},3.25×10^{-3}) p=0.06
ψ	2B	1.55×10^{-3} (-0.329×10^{-3},3.42×10^{-3}) p=0.11

7.9 Extensions

We showed that least squares estimands are weighted ADEs with non-negative exposure weights, by using Theorem 7 to connect contrast functions to exposure weights and using Lemma 4 to demonstrate non-negativity when the contrast function has certain monotonicity. Here we apply these results again to propose weighted ADEs that reduce to the ADE (Example 1) when A follows a specific parametric distribution given $\boldsymbol{Z}$.

We consider the form of the ADE contrast function (7.9) when A follows a known parametric distribution conditional on $\boldsymbol{Z} = \boldsymbol{z}$. Consider the normal, gamma ($A > 0$) and asymmetric Laplace distributions (Yu and Zhang, 2005) with the respective density functions

$$\begin{aligned} f_1(a|\mu,\sigma) &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(a-\mu)^2}{2\sigma^2}\right\} \\ f_2(a|\alpha,\beta) &= \frac{\beta}{\Gamma(\alpha)}a^{\alpha-1}\exp(-\beta a) \\ f_3(a|p,\sigma,x_0) &= \frac{p(1-p)}{\sigma}\exp\left\{-(a-a_0)\frac{\Theta(a-a_0)-p}{\sigma}\right\} \end{aligned}$$

where a_0 is a known value and the parameters $\mu, \sigma > 0, \alpha > 0, \beta > 0, p \in (0,1)$ are all constant given $\boldsymbol{Z} = \boldsymbol{z}$. Also, $\Gamma(.)$ represents the gamma function and $\Theta(u)$ is a step function which takes the value 1 when $u > 0$ and 0 otherwise. Plugging these density functions, in turn, into (7.9) gives the contrast functions,

$$\begin{aligned} -\frac{d\log f_1(a|\mu,\sigma^2)}{da} &= \frac{a-\mu}{\sigma^2} \\ -\frac{d\log f_2(a|\alpha,\beta)}{da} &= (1-\alpha)a^{-1}+\beta \\ -\frac{d\log f_3(a|p,\sigma,x_0)}{da} &= \frac{\Theta(a-a_0)-p}{\sigma}. \end{aligned}$$

Technically the third equation above describes the derivative for $a \neq a_0$, since $f_3(a|p,\sigma,a_0)$ is not differentiable at this point. This is not problematic, however, since these expressions are used only to inspire well defined contrast functions. It is readily seen that all three are of the form of a known function of a up to centring and scaling by parameters which are constant given $\boldsymbol{z}$. In particular, they are of the form in (7.7) where $v(a,\boldsymbol{z})$ is replaced with the known functions, a, a^{-1} and $\Theta(a-a_0)$ respectively. According to (7.8), therefore, these contrast functions imply the weighted ADEs,

$$\begin{aligned} \theta_{w,1} &= E\left\{w(\boldsymbol{Z})\frac{\mathrm{cov}(A,Y|\boldsymbol{Z})}{\mathrm{var}(A|\boldsymbol{Z})}\right\} \\ \theta_{w,2} &= E\left\{w(\boldsymbol{Z})\frac{\mathrm{cov}(A^{-1},Y|\boldsymbol{Z})}{\mathrm{cov}(A^{-1},A|\boldsymbol{Z})}\right\} \\ \theta_{w,3} &= E\left\{w(\boldsymbol{Z})\frac{E(Y|A>a_0,\boldsymbol{Z})-E(Y|A\leq a_0,\boldsymbol{Z})}{E(A|A>a_0,\boldsymbol{Z})-E(A|A\leq a_0,\boldsymbol{Z})}\right\} \end{aligned}$$

which, for $w(\boldsymbol{Z}) = 1$, reduce to the ADE when A respectively follows the normal, gamma, and asymmetric Laplace distribution conditional on $\boldsymbol{Z}$. The estimand $\theta_{w,1}$ is the least squares estimand studied in this paper, and is thus further motivated by its connection to normally distributed exposures. These estimands, however, are nonparametrically well defined, even when the exposure does not follow the associated parametric distribution or indeed when it is not continuous or $\mu'(a, \boldsymbol{z})$ does not exist. This raises the questions of the extent to which $\theta_{w,2}$ and $\theta_{w,3}$ are useful and interpretable estimands in their own right, in what contexts one might use them, how $w(\boldsymbol{Z})$ should be chosen, and how best they should be estimated. Such questions are beyond the scope of the current work. We remark, however, that $\theta_{w,3}$ with the weight $w(\boldsymbol{Z}) = E(A|A > a_0, \boldsymbol{Z}) - E(A|A \leq a_0, \boldsymbol{Z})$ reduces to

$$E\left\{E(Y|A > a_0, \boldsymbol{Z}) - E(Y|A \leq a_0, \boldsymbol{Z})\right\},$$

which identifies the ATE of a dichotomised exposure on outcome.

7.10 Discussion

The current work makes several contributions to the literature on weighted ADEs for a single covariate (exposure). We decompose the weight into an exposure weight and a subgroup weight, and demonstrate that the former is equivalently represented by a 'contrast function'. This is a function $l(a|\boldsymbol{z})$ such that $E\{l(A|\boldsymbol{Z})|\boldsymbol{Z}\} = 0$ and $E\{l(A|\boldsymbol{Z})A|\boldsymbol{Z}\} = 1$, which ensures that the quantity $E\{l(A|\boldsymbol{Z})Y|\boldsymbol{Z}\}$ quantifies the effect of A on Y.

We show that least squares estimands, which are estimands connected to partially linear model projections, are in fact weighted ADEs with a particular choice of contrast function. We further motivate least squares estimands by considering the weighted ADE that minimises the nonparametric efficiency bound when the weight (i.e. the exposure distribution) is known and the outcome is homoscedastic. Our efficiency analysis extends the methods of Crump et al. (2006) to the setting of a continuous exposures.

We further use the ICs of the proposed least squares estimands to derive efficient one-step estimators, $\hat{\Psi}$ and $\hat{\psi}$, the latter of which generalises the AIPW to the setting of a continuous exposure. To estimate the working models we recommend a quasi-oracle approach based on the R-learner (Nie and Wager, 2021). Our proposal involves a novel quasi-oracle learner for the inverse variance, $1/\text{var}(A|\boldsymbol{Z})$, which is designed to mitigate extreme weighting in the estimator.

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT

T: +44 (0)20 7299 4646
F: +44 (0)20 7299 4656
www.lshtm.ac.uk

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	1804954	**Title**	Mr.
First Name(s)	Oliver		
Surname/Family Name	Hines		
Thesis Title	Assumption-Lean Inference for Causal and Statistical Questions in the Era of Machine Learning		
Primary Supervisor	Karla Diaz-Ordaz		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?			
When was the work published?			
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Choose an item.	Was the work subject to academic peer review?	Choose an item.

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	Not yet decided
Please list the paper's authors in the intended authorship order:	Oliver Hines, Karla Diaz-Ordaz, Stijn Vansteelandt
Stage of publication	**Not yet submitted**

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	This study was conceived by all authors. I carried out the mathematical research, computational simulations and writing of the manuscript under the supervision of the other authors.

SECTION E

Student Signature	Ohines
Date	14 December 2022

Supervisor Signature	Karla Diaz-Ordaz
Date	14 December 2022

Chapter 8

Causal derivative effects for continuous exposures

8.1 Introduction

One of the central goals of statistical methods in epidemiology is to establish the main effect of an exposure A on an outcome Y and to determine its magnitude and direction. When the exposure is continuous (e.g. dose, duration, frequency), nonparametrically defining main effect estimands is not so straightforward. Traditional parametric regression methods define main effects through parameters indexing a model for the conditional mean outcome given exposure and a sufficient collection of confounders, $\boldsymbol{X}$, an approach which is problematic for several reasons.

The approach relies entirely on correct specification of the outcome model. However, in practice, regression models tend to be chosen for their mathematical convenience, rather than because of any a priori knowledge of the data generating mechanism. This problem persists when data adaptive variable selection methods are used to choose a working model from a set of candidate models, in which case the uncertainty in choosing the right model tends to be systematically ignored. Similar concerns apply to methods which generalise propensity score modelling to the continuous exposure setting, using models for the exposure density given confounders (Hirano and Imbens, 2005; Imai and Van Dyk, 2004; Galvao and Wang, 2015). Additionally, when the outcome model itself is complicated, for example when machine learning methods are used to model interaction terms and non-linearities, it may be difficult to define parsimonious scalar summary statistics, thereby encouraging researchers to ignore these complexities (Breiman, 2001b).

When the exposure is binary (coded 0,1), one can draw on a rich literature of causal effects to nonparametrically define main effect estimands of interest, for example through the average treatment effect (ATE) (Robins, 1994), $E(Y^1 - Y^0)$, where Y^a denotes the outcome that would be observed if exposure had taken the value $A = a$. The ATE is motivated by contrasting the mean outcome in two counterfactual worlds, where all units are assigned the exposure levels 1 and 0 respectively (Rubin, 1974). Under standard identifiability assumptions, regular asymptotically linear estimators for the ATE (and weighted variations) may be constructed which permit valid inference even when data adaptive methods are used to fit working models (Zheng and van der Laan, 2011; Chernozhukov et al., 2018), such as for the conditional response surface $\mu(a, \boldsymbol{x}) \equiv E(Y|A = a, \boldsymbol{X} = \boldsymbol{x})$, and the propensity score $\pi(\boldsymbol{x}) \equiv E(A|\boldsymbol{X} = \boldsymbol{x})$.

The dose-response curve is perhaps the most common generalization of the ATE to continuous exposures. It imagines a counterfactual world where an intervention assigns the same exposure level for all units (Robins and Rotnitzky, 2001; Kennedy et al., 2017). The result is a function of the intervention level, $\varphi(a) \equiv E(Y^a)$. This curve, however, is usually also problematic for the following reasons.

Firstly, interventions that set the exposure to the same level are often unrealistic and therefore scientifically less interesting. This is even more so in settings where confounders are strong predictors of exposure (e.g. diet and physical activity level are strong predictors of exposure human body mass index),

thus assigning the value $A = a^*$ may represent an unrealistic intervention for treatment units in some confounder subgroups, even when for others it may be reasonable. We would argue that the dose-response curve is therefore practically uninformative for answering many scientific questions of interest, not least in the exploratory stage of an analysis where there may not be a particular intervention in mind.

Secondly, in the setting of poor overlap between confounder subgroups, estimation of $\varphi(a^*)$, under standard identifiability assumptions, may also require significant extrapolation, for example to obtain an estimate of $\mu(a^*, \boldsymbol{x}_0)$ when there are few observations of $A \approx a^*$ for a particular confounder group, $\boldsymbol{x}_0$. Additional concerns relate to the fact that the dose-response curve is an infinite dimensional parameter (i.e. a function) rather than a scalar summary statistic. This inherently makes estimation more difficult and also means there is no clear way to summarize the resulting curve once it has been obtained, see e.g. Kennedy et al. (2017); Neugebauer and van der Laan (2007) for estimation strategies.

In view of these concerns we make an alternative proposal, which is to imagine a counterfactual world where the exposure distribution for all treatment units is shifted by an infinitesimal amount. This proposal has the advantage that for all units we consider only realistic exposure values. It also relates to an existing literature on ADEs (sometimes called weighted average derivatives or average partial effects), popular in econometrics (Härdle and Stoker, 1989; Powell et al., 1989; Newey and Stoker, 1993). These were originally motivated by semi-parametric index models, under which ADEs are proportional to indexing parameters. In our developments we rely instead on a causal interpretation by considering the 'counter factual derivative'

$$\delta(a, \boldsymbol{x}) \equiv \lim_{\epsilon \to 0} \epsilon^{-1} E(Y^{a+\epsilon} - Y^a | \boldsymbol{X} = \boldsymbol{x}), \tag{8.1}$$

which we assume exists. One such effect estimand is the average derivative effect (ADE), $E\{\delta(A, \boldsymbol{X})\}$, which considers the effect of shifting each individual's observed exposure and acts as a continuous analogue of the ATE. We note that the ADE is not the same as the average derivative of the dose-response curve, which we discuss in Section 8.2.4.

Using the identification result $\delta(a, \boldsymbol{x}) = \mu'(a, \boldsymbol{x})$, where $\mu'(a, \boldsymbol{x})$ denotes the derivative of $\mu(a, \boldsymbol{x})$ w.r.t. a, then the ADE is identifed by $E\{\mu'(A, \boldsymbol{X})\}$. Inference for $E\{\mu'(A, \boldsymbol{X})\}$ usually requires estimation of (a) the conditional density of exposure given confounders and (b) the derivative of (a) w.r.t. exposure (Härdle and Stoker, 1989; Newey and Stoker, 1993; Cattaneo et al., 2013). Estimates of (a) and (b) are typically obtained by nonparametric kernel based methods, however, the reliance on kernel methods introduces complicated biases as the dimension of $\boldsymbol{X}$ increases, due to the curse of dimensionality (Cattaneo et al., 2013).

Alternative estimation strategies, when $\boldsymbol{X}$ is high dimensional, rely on parametric (usually single-index) models for the conditional mean outcome (Wooldridge and Zhu, 2020; Hirshberg and Wager, 2018). These assume that the outcome model is known and apriori specified, which is problematic for the reasons mentioned above. Instead we advocate changing the focus to weighted ADE estimands which are amenable to data adaptive estimation of $\mu(a, \boldsymbol{x})$. In Chapter 7, it is shown that two such weighted ADE estimands, are,

$$\psi = E\left\{ \frac{\text{cov}(A, Y|\boldsymbol{X})}{\text{var}(A|\boldsymbol{X})} \right\} \tag{8.2}$$

and

$$\Psi = \frac{E\left\{\text{cov}(A, Y|\boldsymbol{X})\right\}}{E\left\{\text{var}(A|\boldsymbol{X})\right\}}, \tag{8.3}$$

when A is continuous and $\mu'(a, \boldsymbol{x})$ exists, although they remain well defined even when A is discrete or $\mu(a, \boldsymbol{x})$ is not differentiable. In Chapter 7, we show that Ψ is optimally efficient in the the class of weighted ADEs, under heteroskedasticity of the outcome. We refer to ψ and Ψ as least squares estimands due to their connection to to linear model projections Vansteelandt and Dukes (2022); Robins et al. (2008); Newey and Robins (2018), and note that the numerator of Ψ has been proposed as a nonparametric estimand in the context of conditional independence testing (Shah and Peters, 2018).

Moreover, when A is binary, ψ identifies the ATE and Ψ identifies the propensity overlap weighted effect (Crump et al., 2006), however, in the current paper we focus on the setting where A is a continuous random variable. In Section 8.2 we ascribe a causal interpretation to ψ and Ψ using the counter factual derivative, which is the main contribution of the current work. Our interpretation relates least squares estimands to weighted derivative effects under specific stochastic interventions (Díaz and van der Laan, 2012; Kennedy, 2019).

In Section 8.3, estimators for ψ and Ψ are discussed which attain the efficiency bound under the nonparametric model. These estimators do not contain contributions from (a) or (b), thus alleviating the aforementioned concerns regarding estimation of the ADE.

8.2 Methodology

8.2.1 Causal derivative estimands

Suppose we have n iid observations, $(\boldsymbol{z}_1, ..., \boldsymbol{z}_n)$ of a random variable $\boldsymbol{Z}$ distributed according to an unknown distribution P_0, such that $\boldsymbol{Z}$ consists of $(Y, A, \boldsymbol{X})$, where $Y \in \mathbb{R}$ is an outcome, $A \in \mathbb{R}$ is a continuous covariate of interest which we call an 'exposure' and $\boldsymbol{X} \in \mathbb{R}^p$ is a p-dimensional vector of covariates. Also let $f(a|\boldsymbol{x})$ denote the density of A given X under P_0. Assuming such a limit exists, we define the counterfactual derivative,

$$\delta(a, \boldsymbol{x}) \equiv \lim_{\epsilon \to 0} \epsilon^{-1} E(Y^{a+\epsilon} - Y^a | \boldsymbol{X} = \boldsymbol{x})$$

which we use to define the conditional ADE,

$$\begin{aligned} \lambda(\boldsymbol{x}) &\equiv E\{\delta(A, \boldsymbol{X}) | \boldsymbol{X} = \boldsymbol{x}\} \\ &= \lim_{\epsilon \to 0} \epsilon^{-1} \int E(Y^{a+\epsilon} - Y^a | \boldsymbol{X} = \boldsymbol{x}) f(a|\boldsymbol{x}) da \end{aligned}$$

and the ADE

$$\theta \equiv E\{\delta(A, \boldsymbol{X})\} = E\{\lambda(\boldsymbol{X})\}$$

In Section 8.2.4 these definitions are extended by introducing nonnegative weight functions, and in Section 8.2.5 identification assumptions are discussed.

8.2.2 Motivating example

We motivate the counterfactual derivative, conditional ADE, and ADE in the following short example. Consider that A represents the time in days between the first and second doses of a vaccine, and Y represents a measurement of the immune response taken some prescribed time (e.g. 60 days) after the first dose, with $\boldsymbol{X}$ representing observed patient characteristics at baseline, see e.g. Gilbert et al. (2021) for a similar set up. Suppose we are presented with the ith patient, with covariates $\boldsymbol{x}_i$, and exposure a_i. The value $\delta(a_i, \boldsymbol{x}_i)$ quantitatively answers the causal question, "How would would one expect the outcome to change under an intervention where the patient receives an exposure which is slightly (i.e. infinitesimally) shifted away from the one which they actually received?".

In particular, supposing an exposure time of $a_i = 29$ days for the ith individual, then $\delta(29, \boldsymbol{x}_i)$ represents the change in outcome for small shifts in the exposure (in the sense of a derivative) around the true exposure value for that individual. Whilst this is retrospectively interesting after an exposure level has been assigned, for the purposes of advising treatment policy we may wish so consider how this measure might look before the exposure level is assigned.

Consider, for example, that the individual receives their first vaccination dose then is asked to return to a walk-in clinic to receive a second vaccination dose, 4 to 8 weeks after the first, with walk-in behaviour strongly predicted by $\boldsymbol{x}_i$. In this instance it may be of interest to pose the causal question: "Given the

characteristics of the patient, would they benefit from returning to the clinic slightly sooner or later than they otherwise would?".

We argue that the answer to this question is scientifically interesting at an early stage of analysis, to obtain a general understanding of effect direction when no particular treatment intervention is planned. The conditional ADE answers this question by averaging the causal derivative over realistic exposure values for the each individual. Specifically, for the ith individual, let $a_{i,j}$ denote one of m draws from the distribution of A given $\boldsymbol{X} = \boldsymbol{x}_i$. The conditional ADE $\lambda(\boldsymbol{x}_i)$ represents the probability limit of $m^{-1}\sum_{j=1}^{m}\delta(a_{i,j}, \boldsymbol{x}_i)$, as the number of draws $m \to \infty$.

Finally, the ADE $\theta = E\{\lambda(\boldsymbol{X})\}$, provides an answer to a similar question, asked of the population as a whole, "Do members of the patient population benefit from returning to the clinic slightly sooner or later than they otherwise would?". We argue that this causal question represents the effect of modest realistic policy interventions that have small effects on patient behaviour, e.g. should the vaccinator generally emphasise to patients the importance of returning promptly, or should they advise patients that there is no rush to return. In comparison, the dose-response function $\varphi(a) \equiv E(Y^a)$ seeks to answer a more ambitious question, "what would be the mean outcome if every patient returned after exactly a days?". We illustrate how the two causal questions could suggest different conclusions in the following numerical illustration.

8.2.3 Numerical Illustration

Consider a single binary confounder $X \in \{0, 1\}$, with both values equally prevalent in the population, i.e. $P(X = 1) = 0.5$. Let $E(Y^a|X = 0) = -(a - 30)^2 + 300$ and $E(Y^a|X = 1) = (a - 30)^2 + 100$, so that the dose-response curve takes the value $\varphi(a) \equiv E(Y^a) = 200$ for all exposure values. The counterfactual derivative takes the form, $\delta(a, 0) = 2(30 - a)$ and $\delta(a, 1) = 2(a - 30)$.

Assume that A is conditionally normally distributed in each subgroup, i.e. $A|X = 0 \sim \mathcal{N}(25, \sigma)$ and $A|X = 1 \sim \mathcal{N}(35, \sigma)$, where $\mathcal{N}(\mu, \sigma)$ represents a normal distribution with mean μ and variance σ^2. In this case one obtains the conditional ADEs, $\lambda(0) = \lambda(1) = 10$, hence the ADE is $\theta = 10$. This set up is shown in the plots in Figure 8.1 and 8.2.

In this illustration, the fact that the dose-response curve is constant suggests that no exposure value should be preferred over any other when applied to the population as a whole. The ADE, however, is positive, which suggests that individuals in the population would, on average, benefit from increasing their exposure level. Whilst counterintuitive, these two conclusions are not incompatible, and we would argue that in this instance, the ADE captures the effect of a modest change in the exposure distribution which is missed by dose-response curve modelling.

8.2.4 Interventional derivative estimands

Here we extend the causal derivative effect definitions in Section 8.2 by considering a stochastic intervention distribution (Díaz and van der Laan, 2012; Kennedy, 2019). Let $\tilde{P}$ denote a distribution over A conditional on $\boldsymbol{X}$ such that $\tilde{P}$ is absolutely continuous w.r.t. P_0. Letting $\tilde{f}(a|\boldsymbol{x})$ denote the density of A given X under $\tilde{P}$, we define the 'exposure weight' $w(A|\boldsymbol{X}) \equiv \tilde{f}(a|\boldsymbol{x})/f(a|\boldsymbol{x})$. It follows that the exposure weight is non-negative and normalised such that $E\{w(A|\boldsymbol{X})|\boldsymbol{X}\} = 1$ almost surely. We define the conditional interventional ADE (IADE) as

$$\begin{aligned}\lambda_w(\boldsymbol{x}) &\equiv E\{w(A|\boldsymbol{X})\delta(A, \boldsymbol{X})|\boldsymbol{X} = \boldsymbol{x}\} = E_{\tilde{P}}\{\delta(A, \boldsymbol{X})|\boldsymbol{X} = \boldsymbol{x}\} \\ &= \lim_{\epsilon \to 0} \epsilon^{-1} \int E(Y^{a+\epsilon} - Y^a|\boldsymbol{X} = \boldsymbol{x})\tilde{f}(a|\boldsymbol{x})da\end{aligned}$$

which we interpret as the conditional derivative effect in a world where the exposure follows the interventional distribution $\tilde{P}$ given $\boldsymbol{X}$. We note that setting the intervention distribution to the true distribution (i.e. $\tilde{P} = P_0$) recovers the conditional ADE $\lambda_w(\boldsymbol{x}) = \lambda(\boldsymbol{x})$. Next, consider a 'subgroup weight' $w(\boldsymbol{X})$ which is non-negative and normalised such that $E\{w(\boldsymbol{X})\} = 1$. Using this subgroup weight, we define the

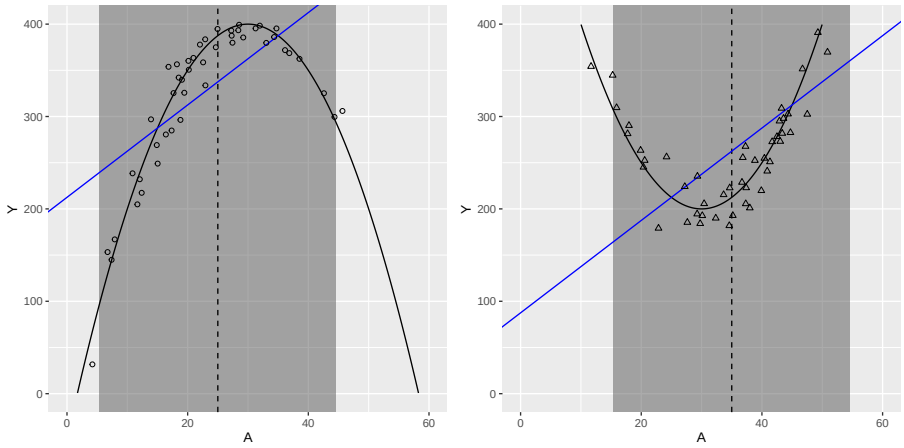


Figure 8.1: Exposure A plotted against outcome Y for the $X = 0$ subpopulation (left plot, circular points) and the $X = 1$ subpopulation (right plot, triangular points). In each plot, the vertical dashed line represents the mean exposure value, with dark grey bands denoting the region between the $(0.05, 0.95)$ quantiles of the exposure distribution. The gradient of the blue line denotes the conditional ADEs $\lambda(0)$ and $\lambda(1)$ respectively.

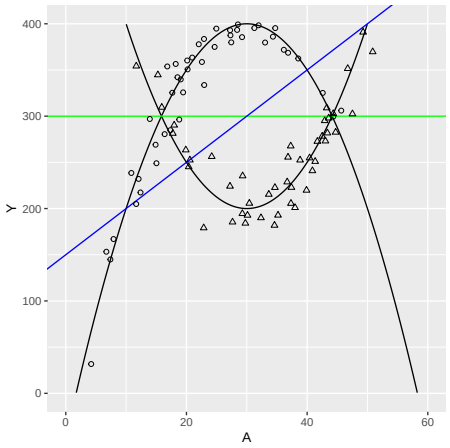


Figure 8.2: Exposure A plotted against outcome Y for the $X = 0$ subpopulation (circular points) and the $X = 1$ subpopulation (triangular points). The gradient of the blue line denotes the ADE θ, and the green line denotes the dose-response function $\varphi(a)$.

weighted IADE as

$$\theta_w \equiv E\{w(A, \boldsymbol{X})\delta(A, \boldsymbol{X})\} = E\{w(\boldsymbol{X})\lambda_w(\boldsymbol{X})\}$$

where $w(A, \boldsymbol{X}) \equiv w(A|\boldsymbol{X})w(\boldsymbol{X})$ is the product of the exposure and subgroup weights. We interpret the weighted IADE in the following way. First we imagine a counterfactual world, where we intervene on the distribution of exposure given confounders, replacing the true distribution with the intervention distribution, $\tilde{P}$. Next we imagine a second counterfactual world where the exposure is distributed according to the intervention distribution, but with an infinitesimal perturbation for all treatment units. The proposed weighted IADE estimand is the difference in (subgroup weighted) mean outcome between these two worlds, rescaled by the perturbation size in the sense of the derivative.

We now demonstrate two examples of weighted IADEs. To highlight the difference between the ADE and dose-response curve modelling, we compare the ADE (Example 4) with the average derivative of the dose-response curve (Example 5), which is an IADE. The latter quantifies the ADE in a counterfactual world where, for all individuals in the population, the exposure is distributed according to the true marginal exposure distribution.

Example 4 (Average derivative effect) *Setting the intervention distribution to the true distribution ($\tilde{P} = P_0$), i.e. $\tilde{f}(a|\boldsymbol{x}) = f(a|\boldsymbol{x})$, implies a unitary exposure weight, $w(A|\boldsymbol{X}) = 1$. Letting the subgroup weight $w(\boldsymbol{X}) = 1$ results in the conditional and marginal derivative effects*

$$\lambda(\boldsymbol{x}) = E\{\delta(A, \boldsymbol{X})|\boldsymbol{X} = \boldsymbol{x}\}$$
$$\theta = E\left\{\delta(A, \boldsymbol{X})\right\},$$

as in Section 8.2.1.

Example 5 (Average dose-response derivative) *Consider the dose-response curve, $\varphi(a) = E(Y^a)$. The mean derivative of this curve, $\theta_w = E\{\varphi'(\boldsymbol{A})\}$ is an IADE with subgroup weight $w(\boldsymbol{X}) = 1$ and intervention distribution $\tilde{P}$ that has density $\tilde{f}(a|\boldsymbol{x}) = f(a)$, where $f(a)$ is the marginal density of A under P_0. This distribution implies the exposure weight $w(A|\boldsymbol{X}) = f(A)/f(A|\boldsymbol{X})$, where we make a positivity assumption such that $f(a) \neq 0 \implies f(a|\boldsymbol{x}) \neq 0$ for all $\boldsymbol{x}$. This intervention distribution implies the conditional and weighted IADEs*

$$\lambda_w(\boldsymbol{x}) = E\left\{\left[\frac{f(A)}{f(A|\boldsymbol{X})}\right]\delta(A, \boldsymbol{X})|\boldsymbol{X} = \boldsymbol{x}\right\} = E\left\{\delta(A, \boldsymbol{x})\right\}$$
$$\theta_w = E\left\{\left[\frac{f(A)}{f(A|\boldsymbol{X})}\right]\delta(A, \boldsymbol{X})\right\} = E\{\lambda_w(\boldsymbol{X})\} = E\{\varphi'(\boldsymbol{A})\}.$$

To reiterate, we interpret the IADE above in the same way as the ADE, except in a counterfactual world where the exposure distribution is the same for all individuals, and equal to the true marginal exposure distribution. Intuitively, the value for the average dose-response derivative could be very different from that of the ADE when the conditional and marginal exposure distributions differ significantly, i.e. when $\boldsymbol{X}$ is a strong predictor of A.

8.2.5 Identifiability

Under standard assumptions of consistency ($A = a \implies Y = Y^a$) and ignorability ($Y^a \perp\!\!\!\perp A|\boldsymbol{X}$ for all a), one obtains the identification result,

$$E(Y^a|\boldsymbol{X} = \boldsymbol{x}) = E(Y^a|A = a, \boldsymbol{X} = \boldsymbol{x}) = E(Y|A = a, \boldsymbol{X} = \boldsymbol{x}) = \mu(a, \boldsymbol{x}).$$

Hence, under these assumptions we write the dose-response curve as

$$\varphi(a) = E(Y^a) = E\{E(Y^a|\boldsymbol{X})\} = E\{\mu(a, \boldsymbol{X})\}$$

and the counterfactual derivative as

$$\begin{aligned}\delta(a,\boldsymbol{x}) &= \lim_{\epsilon\to 0}\epsilon^{-1}E(Y^{a+\epsilon}-Y^a|\boldsymbol{X}=\boldsymbol{x})\\ &= \lim_{\epsilon\to 0}\epsilon^{-1}\left\{E(Y^{a+\epsilon}|\boldsymbol{X}=\boldsymbol{x})-E(Y^a|\boldsymbol{X}=\boldsymbol{x})\right\}\\ &= \lim_{\epsilon\to 0}\epsilon^{-1}\left\{\mu(a+\epsilon,\boldsymbol{x})-\mu(a,\boldsymbol{x})\right\}\equiv\mu'(a,\boldsymbol{x})\end{aligned}$$

It follows that the conditional IADE and weighted IADE are respectively identified by

$$\lambda_w(\boldsymbol{x}) = E\{w(A|\boldsymbol{X})\mu'(A,\boldsymbol{X})|\boldsymbol{X}=\boldsymbol{x}\} \tag{8.4}$$

$$\theta_w = E\{w(A,\boldsymbol{X})\mu'(A,\boldsymbol{X})\} \tag{8.5}$$

The latter expression is exactly the expression for the weighted ADE, popular in econometrics literature (Härdle and Stoker, 1989; Powell et al., 1989; Newey and Stoker, 1993). The weighted ADE is conventionally defined through a nonnegative weight function $w(A,\boldsymbol{X})$, rather than through an intervention distribution $\tilde{P}$ and an exposure weight $w(\boldsymbol{X})$, as in the current work. The two definitions, however are seen to be equivalent, since, for a given weight function $w(A,\boldsymbol{X})$, one can define a subgroup weight as $w(\boldsymbol{X})=E\{w(A,\boldsymbol{X})|\boldsymbol{X}\}$, which implies the existance of a nonnegative exposure weight $w(A,\boldsymbol{X}) = w(A|\boldsymbol{X})w(\boldsymbol{X})$ such that $E\{w(A|\boldsymbol{X})|\boldsymbol{X}\}=1$ almost surely. We argue that by introducing an intervention distribution $\tilde{P}$, weighted ADE estimands can be interpreted in terms of the causal IADE estimands proposed in the current paper.

8.2.6 Contrast representation

It is a standard result that weighted ADEs can be rewritten using integration by parts under limited additional assumptions (Powell et al., 1989) (Chapter 7). Here we restate this result in terms of intervention distributions, and illustrate how least squares estimands are weighted IADEs based on specific intervention distributions.

Assume that (A1) $\tilde{f}(a|\boldsymbol{x})$ has a derivative w.r.t. a that we denote $\tilde{f}'(a|\boldsymbol{x})$, assume (A2) $\tilde{f}(a|\boldsymbol{x})=0$ for a on the boundary of the support of A, and (A3) $f(a|\boldsymbol{x})=0$ implies $\tilde{f}(a|\boldsymbol{x})=0$, i.e. $\tilde{P}$ is absolutely continuous w.r.t. P_0. Under (A1), (A2) and (A3) it follows from integration by parts that (8.4) and (8.5) can be written as

$$\lambda_w(\boldsymbol{x}) = E\{l(A|\boldsymbol{X})Y|\boldsymbol{X}=\boldsymbol{x}\} \tag{8.6}$$

$$\theta_w = E\{w(\boldsymbol{X})l(A|\boldsymbol{X})Y\} \tag{8.7}$$

where

$$l(a|\boldsymbol{x}) = -\frac{\tilde{f}'(a|\boldsymbol{x})}{f(a|\boldsymbol{x})} \tag{8.8}$$

We refer to the expressions in (8.6) and (8.7) as 'the contrast representation' of the weighted ADE.

In Chapter 7, we showed that weighted ADEs of the form in (8.5) are equivalent expressed in as

$$E\left\{w(\boldsymbol{X})\frac{\mathrm{cov}\{v(A,\boldsymbol{X}),Y|\boldsymbol{X}\}}{\mathrm{cov}\{v(A,\boldsymbol{X}),A|\boldsymbol{X}\}}\right\}. \tag{8.9}$$

where $v(a,\boldsymbol{x})$ is an almost arbitrary function such that $\mathrm{cov}\{v(A,\boldsymbol{X}),A|\boldsymbol{X}\}\neq 0$. For example, setting $v(a,\boldsymbol{x})=l(a|\boldsymbol{x})$ recovers (8.7), since $E\{l(A|\boldsymbol{X})|\boldsymbol{X}\}=0$ and $E\{l(A|\boldsymbol{X})A|\boldsymbol{X}\}=1$.

In Theorem 11 below, we restate the main result of Chapter 7 in terms of a function $\tilde{f}(a|\boldsymbol{x})$. This function is a probability density function, provided that $\tilde{f}(a|\boldsymbol{x})$ is non-negative. Lemma 6 shows that monotonicity of $v(a,\boldsymbol{x})$ is sufficient to guarantees non-negativity of $\tilde{f}(a|\boldsymbol{x})$.

Theorem 11 *Let $v(a,\boldsymbol{x})$ be function such that $\mathrm{cov}\{v(A,\boldsymbol{X}),A|\boldsymbol{X}\}\neq 0$. Let $F(a|\boldsymbol{x})$ be the distribution function of A given $\boldsymbol{X}=\boldsymbol{x}$ under P_0 and assume that $f(a|\boldsymbol{x})>0$ for a on the convex support of A. Define*

$$\tilde{f}(a|\boldsymbol{x}) = \frac{F(a|\boldsymbol{x})\{1-F(a|\boldsymbol{x})\}\left[E\{v(A,\boldsymbol{X})|A>a,\boldsymbol{X}=\boldsymbol{x}\}-E\{v(A,\boldsymbol{X})|A\leq a,\boldsymbol{X}=\boldsymbol{x}\}\right]}{\mathrm{cov}\{v(A,\boldsymbol{X}),A|\boldsymbol{X}=\boldsymbol{x}\}}. \tag{8.10}$$

Then $\tilde{f}(a|\boldsymbol{x})$ satisfies (A1), (A2), (A3),

$$\int \tilde{f}(a|\boldsymbol{x})da = 1,$$

and for the weight function $w(a|\boldsymbol{x}) = \tilde{f}(a|\boldsymbol{x})/f(a|\boldsymbol{x})$,

$$\frac{\text{cov}\{v(A,\boldsymbol{X}),Y|\boldsymbol{X}\}}{\text{cov}\{v(A,\boldsymbol{X}),A|\boldsymbol{X}\}} = E\{w(A|\boldsymbol{X})\mu'(A,\boldsymbol{X})|\boldsymbol{X}\} \quad a.s.$$

$$E\left\{w(\boldsymbol{X})\frac{\text{cov}\{v(A,\boldsymbol{X}),Y|\boldsymbol{X}\}}{\text{cov}\{v(A,\boldsymbol{X}),A|\boldsymbol{X}\}}\right\} = E\{w(A,\boldsymbol{X})\mu'(A,\boldsymbol{X})\} \tag{8.11}$$

where $w(a,\boldsymbol{x}) = w(\boldsymbol{x})w(a|\boldsymbol{x})$ for some exposure weight $w(\boldsymbol{x})$.Proof in Appendix F.

Lemma 6 *Let $v(a,\boldsymbol{x})$ be a function which is monotonically increasing or decreasing in a (but is not everywhere constant), for a on the support of A. Then $\tilde{f}(a|\boldsymbol{x})$ in* (8.10) *is a probability density function. Proof in Appendix F.*

The significance of Theorem 11 and Lemma 6 is that together they imply that any estimand of the type in (8.9) can be interpreted as a weighted ADE. Moreover, provided that $v(a,\boldsymbol{x})$ is monotonic, then (8.9) identifies an IADE with the intervention distribution $\tilde{P}$ given by the density function in (8.10).

We focus in particular on the so-called least squares estimand, which corresponds to the choice $v(a,\boldsymbol{x}) = a$.

$$E\left\{w(\boldsymbol{X})\frac{\text{cov}\{A,Y|\boldsymbol{X}\}}{\text{var}\{A|\boldsymbol{X}\}}\right\} \tag{8.12}$$

with ψ and Ψ, introduced in (8.2) and (8.3) arising as special cases for the exposure weights $w(\boldsymbol{X}) = 1$ and $w(\boldsymbol{X}) = \text{var}(A|\boldsymbol{X})/E\{\text{var}(A|\boldsymbol{X})\}$ respectively. These least squares estimands are well-motivated by considering projection coefficients in linear models, with Ψ additionally representing an optimally efficient weighted ADE, and ψ representing the ADE θ when the exposure A is normally distributed given $\boldsymbol{X}$ (Chapter 7).

In view of Theorem 11 and Lemma 6, we argue that the least squares estimands, ψ and Ψ, identify weighted IADEs corresponding to a specific intervention distribution $\tilde{P}$ and in the next Section, we examine exactly how this intervention distribution relates to the true distribution P_0. To reiterate, in this work we interpret these least squares estimands as the (weighted) mean difference in outcome between two counterfactual worlds, one where the exposure is distributed according to an intervention density, and one where the exposure is distributed according to the same intervention density, but with an infinitesimal shift in exposure.

We remark that one could choose alternative functions $v(a,\boldsymbol{x})$ to construct interesting estimands. In particular, if we let a_0 be a known value and let $v(a,\boldsymbol{x}) = 1$ for $a > a_0$, with $v(a,\boldsymbol{x}) = 0$ otherwise, then one recovers the estimand

$$E\left\{w(\boldsymbol{X})\frac{E(Y|A > a_0,\boldsymbol{X}) - E(Y|A \leq a_0,\boldsymbol{X})}{E(A|A > a_0,\boldsymbol{X}) - E(A|A \leq a_0,\boldsymbol{X})}\right\}.$$

which, like the least squares estimands, identifies a weighted IADE for a specific intervention distribution. This estimand is noteworthy because, under standard assumptions, it identifies the weighted ATE of a dichotomised exposure on outcome, i.e. the treatment effect of the binary variable which is constructed by dichotomising A at a_0. In the interest of brevity we will not examine the intervention distribution of this dichotomisation estimand, and refer to Chapter 4 for some additional discussion.

8.2.7 Least squares intervention distribution

Here we examine the intervention distribution associated with the least squares estimand in (8.12), i.e. the distribution obtained from (8.10) with $v(a,\boldsymbol{x}) = a$. This distribution has the density

$$\tilde{f}(a|\boldsymbol{x}) = \frac{F(a|\boldsymbol{x})\{1 - F(a|\boldsymbol{x})\}}{\text{var}(A|\boldsymbol{X} = \boldsymbol{x})}\{E(A|A > a,\boldsymbol{X} = \boldsymbol{x}) - E(A|A \leq a,\boldsymbol{X} = \boldsymbol{x})\} \tag{8.13}$$

which we call the least squares intervention distribution.

It is informative to consider that the cumulant function of this distribution, $\tilde{K}(t|\boldsymbol{x}) = \log E_{\tilde{P}}(e^{tA}|\boldsymbol{X} = \boldsymbol{x})$, in terms of the cumulant function of the true distribution, $K(t|\boldsymbol{x}) = \log E(e^{tA}|\boldsymbol{X} = \boldsymbol{x})$, and its derivative w.r.t. t, $K'(t|\boldsymbol{x})$, is

$$\tilde{K}(t|\boldsymbol{x}) = K(t|\boldsymbol{x}) + \log\left(\frac{K'(t|\boldsymbol{x}) - \pi(\boldsymbol{x})}{t\beta(\boldsymbol{x})}\right) \tag{8.14}$$

where $\pi(\boldsymbol{x}) = E(A|\boldsymbol{X} = \boldsymbol{x}) = K'(0|\boldsymbol{x})$ and $\beta(\boldsymbol{x}) = \text{var}(A|\boldsymbol{X} = \boldsymbol{x}) = K''(0|\boldsymbol{x})$, see Appendix G for details. The least squares intervention distribution is therefore that of a shifted exposure, $\tilde{A} = A + \delta$, where δ is a random variable, which is conditionally independent of A given $\boldsymbol{X}$, with cumulant function given by the second term on the right hand side of (8.14). Setting this second term equal to zero, and using the boundary condition that $K(0|\boldsymbol{x}) = 0$ gives

$$K(t|\boldsymbol{x}) = \pi(\boldsymbol{x})t + \beta(\boldsymbol{x})\frac{t^2}{2}$$

This is exactly the cumulant function of a normally distributed variable, with mean $\pi(\boldsymbol{x})$ and variance $\beta(\boldsymbol{x})$. Thus, if A is normally distributed (conditional on $\boldsymbol{X}$), then the least squares intervention distribution is the true distribution, and this is the only exposure distribution for which this is the case. It follows that, under conditional normality of the exposure, the least squares estimand with subgroup weight $w(\boldsymbol{X}) = 1$ identifies the ADE $\theta = E\{\delta(A, \boldsymbol{X})\}$.

To consider other exposure distributions, we imagine a transformation, $\mathcal{F}$, which transforms a probability density $f(a|\boldsymbol{x})$ to its associated least squares intervention distribution, $\tilde{f}(a|\boldsymbol{x})$, according to (8.13), i.e. $\mathcal{F}\{f(.|\boldsymbol{x})\}(a) = \tilde{f}(a|\boldsymbol{x})$. This transformation preserves the symmetry of the density function, as formalized in Theorem 12. We apply this transformation to some well-known distributions to obtain the results in Table 8.1, illustrative plots of which are shown in Fig. 8.3. The fact that the distribution families in Table 8.1 are closed under this transformation is, we believe, somewhat surprising.

Theorem 12 *Let $f(x)$ be a distribution, with finite mean μ, and finite variance.*

$$f(\mu + x) = f(\mu - x) \implies \mathcal{F}\{f(.)\}(\mu + x) = \mathcal{F}\{f(.)\}(\mu - x)$$

Proof in Appendix G.

Table 8.1: Least squares intervention distribution associated with some common distributions. For each result, $f(x|.)$ denotes the density function of the given distribution. See Appendix G for details.

Distribution	Parameters	Result
Normal	mean μ, variance σ^2	$\mathcal{F}\{f(.\|\mu,\sigma)\}(x) = f(x\|\mu,\sigma)$
Gamma	shape α, rate β	$\mathcal{F}\{f(.\|\alpha,\beta)\}(x) = f(x\|\alpha+1,\beta)$
Chi-Squared	degrees of freedom k	$\mathcal{F}\{f(.\|k)\}(x) = f(x\|k+2)$
Beta	shape α and β	$\mathcal{F}\{f(.\|\alpha,\beta)\}(x) = f(x\|\alpha+1,\beta+1)$
Beta Prime	shape α and $\beta > 2$	$\mathcal{F}\{f(.\|\alpha,\beta)\}(x) = f(x\|\alpha+1,\beta-2)$

8.3 Inference

Here we sketch an inference procedure for the estimands

$$\psi_v = E\left\{\frac{\text{cov}(v(A), Y|\boldsymbol{X})}{\text{cov}(v(A), A|\boldsymbol{X})}\right\} \tag{8.15}$$

$$\Psi_v = \frac{E\left\{\text{cov}(v(A), Y|\boldsymbol{X})\right\}}{E\left\{\text{cov}(v(A), A|\boldsymbol{X})\right\}}, \tag{8.16}$$

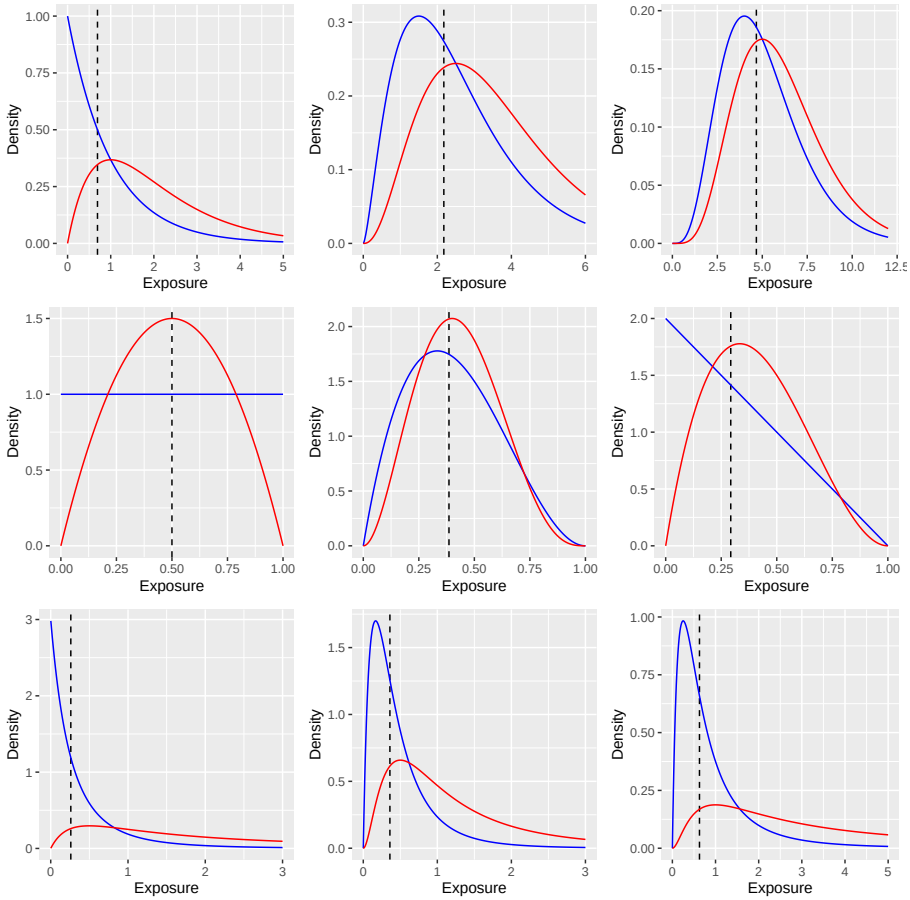


Figure 8.3: Density functions of the least squares intervention distribution (red) corresponding to various true exposure distributions (blue) with the median of the true distribution marked with a dashed line. For the top row of plots, the exposure is in truth gamma distributed with $\beta = 1$ and $\alpha = 1, 2.5, 5$. For the second row, the exposure is in truth beta distributed with $(\alpha, \beta) = (1, 1), (2, 3), (1, 2)$. For the third row the exposure is beta prime distributed with $(\alpha, \beta) = (1, 3), (2, 5), (2, 3)$

which are both indexed by a known function $v(a)$, and which represent special cases of the (8.9) for the exposure weights $w(\boldsymbol{X}) = 1$ and $w(\boldsymbol{X}) \propto \text{cov}(v(A), A|\boldsymbol{X})$. The latter exposure weight is non-negative when $v(a)$ is monotone, which happens to be one of the conditions of Corollary 6. We remark that for the choice $v(a) = a$, the estimands ψ_v and Ψ_v respectively recover the least squares estimands ψ and Ψ in (8.2) and (8.3).

For the least squares estimands, an inference procedure is described in Chapter 7, which uses the nonparametric inference framework based on efficient influence curves, described in Chapter 4. Estimators for the least squares estimands can be generalised by considering the efficient influence curves for ψ_v and Ψ_v

$$\phi_{\psi_v}(\boldsymbol{z}) = \frac{v(a) - \rho(\boldsymbol{x})}{\beta(\boldsymbol{x})}\{y - \mu(\boldsymbol{x}) - \lambda(\boldsymbol{x})(a - \pi(\boldsymbol{x}))\} + \lambda(\boldsymbol{x}) - \psi_v$$
$$\phi_{\Psi_v}(\boldsymbol{z}) = \frac{\{v(a) - \rho(\boldsymbol{x})\}\{y - \mu(\boldsymbol{x}) - \Psi_v(a - \pi(\boldsymbol{x}))\}}{E\{\beta(\boldsymbol{X})\}}$$

where, for convenience we let $\mu(\boldsymbol{x}) = E(Y|\boldsymbol{X} = \boldsymbol{x})$ $\rho(\boldsymbol{x}) = E\{v(A)|\boldsymbol{X} = \boldsymbol{x}\}$, $\pi(\boldsymbol{x}) = E\{A|\boldsymbol{X} = \boldsymbol{x}\}$, $\beta(\boldsymbol{x}) = \text{cov}(v(A), A|\boldsymbol{X} = \boldsymbol{x})$, and $\lambda(\boldsymbol{x}) = \text{cov}(v(A), Y|\boldsymbol{X} = \boldsymbol{x})/\beta(\boldsymbol{x})$.

Using these efficient influence functions, one-step bias correction estimators can be derived, which rely on (possibly data-adaptive/ machine learning) estimators for $\mu(.), \rho(.), \pi(.)$, and for Ψ_v, also $\beta(.)$ and $\lambda(.)$.

$$\hat{\psi}_v = n^{-1}\sum_{i=1}^{n} \frac{v(a_i) - \hat{\rho}(\boldsymbol{x}_i)}{\hat{\beta}(\boldsymbol{x}_i)}\{y_i - \hat{\mu}(\boldsymbol{x}_i) - \hat{\lambda}(\boldsymbol{x}_i)(a_i - \hat{\pi}(\boldsymbol{x}_i))\} + \hat{\lambda}(\boldsymbol{x}_i)$$
$$\hat{\Psi}_v = \frac{\sum_{i=1}^{n}\{v(a_i) - \hat{\rho}(\boldsymbol{x}_i)\}\{y_i - \hat{\mu}(\boldsymbol{x}_i)\}}{\sum_{i=1}^{n}\{v(a_i) - \hat{\rho}(\boldsymbol{x}_i)\}\{a_i - \hat{\pi}(\boldsymbol{x}_i)\}}$$

where superscript hat denotes an estimator that has been trained using data excluding the ith observation. Though we do not formally derive asymptotic results for these estimators, we expect that they will be very similar to those of the the least squares estimands in Chapter 7, since the estimands in that Chapter represent a special case of those above.

Like the least squares estimands, and unlike estimators for e.g. the ADE, the $\hat{\psi}_v$ and $\hat{\Psi}_v$ estimators do not require estimating unknown conditional density functions, thus alleviating concerns regarding unstable density weights in ADE estimators. This comparison is particularly interesting since, under limited causal assumptions, ψ_v and Ψ_v represent ADEs in certain counterfactual worlds, as described in Section 8.2.

8.4 Discussion

The main contribution of the current chapter is to outline a causal interpretation of weighted derivative effects in terms of stochastic interventions. In particular, we connect least squares estimands to the change in mean outcome under small perturbations in a stochastic exposure intervention. The appeal of these estimands is firstly that they focus attention towards modest shifts in exposure around realistic values for each treatment unit. This is advantageous in settings where confounders are strong predictors of exposure, thus the dose-response curve may be both uninformative in answering scientific questions of interest, and too ambitious to estimate due to extrapolation concerns. Derivative effects, however, capture the magnitude and direction of the main effect, which is especially useful in an exploratory analysis, where no specific intervention is planned. They moreover provide a generic effect measure that can be used without needing to choose between many possible shift interventions that could be considered, see e.g. (Rothenhäusler and Yu, 2019).

Additionally, least squares estimands are amenable to data-adaptive estimation of requisite statistical functionals, more so than average derivative effect estimation or dose-response curve modelling. This permits modern inference methods, analogous to those used for average treatment effect inference, to be applied to the setting of continuous exposures. This analogy is unsurprising since the proposed least squares estimands identify the ATE when the exposure is binary, therefore representing a generalisation of the ATE to continuous exposures.

Chapter 9

Conclusion and outlook

9.1 Conclusion

The standard approach to data analysis based on statistical modelling is increasingly being challenged. Despite its undeniable utility over the last century, criticisms have centred around the use of implausible modelling assumptions and, more recently, around the failure of statistical models to outperform algorithmic machine learning models in some prediction tasks. It has been the goal of a growing community within statistics to overcome these limitations and, in the last few decades, a framework based on nonparametric estimands has emerged as an appealing compromise between the interpretability of statistical modelling and the flexibility of algorithmic machine learning. This framework also reduces the burden on the analyst to consider typical statistical modelling questions, which are usually ancillary to the intended investigation, such as whether higher-order terms/ interaction terms/ covariates should be included in the statistical model.

The transition from statistical model based analysis to estimand based analysis arguably represents a paradigm shift toward nonparametric rather than parametric reasoning. This paradigm shift is evidenced by the growing taxonomy of nonparametric estimands (and related statistical/ computational tooling) available to quantitative researchers, with a number of recent proposals related to survival analysis, longitudinal treatments, transportability, etc.[1], as well as those related to mediation, treatment effect heterogeneity, and continuous exposures, which are considered in the current thesis. Mediation estimands in particular exemplify the shift to nonparametric reasoning, where early methods (Sobel, 1982; Baron and Kenny, 1986), which relied heavily on linear models for the outcome and the mediator, have largely been superseded (in the setting of binary treatment at least) by causally defined natural mediation estimands, e.g. those connected to the so-called mediation formula of Pearl (2001).

The initial focus for this thesis project was to investigate applying mediation analysis methods to answer causal questions in the fields of genetics and genetic epidemiology. Genetic cohort data represents an interesting challenge for causal inference and nonparametric estimation, due to (i) the inherently high-dimensional covariate space which SNP measurement data represents, (ii) the natural causal structure implied by accepted rules of genetic inheritance, and (iii) the difficulty in measuring/ modelling environmental variables, which vary during the course of an individual's lifetime. Due to these challenges, existing methods, such as genome wide association studies (GWAS) and Mendelian randomisation (MR) studies rely heavily on both statistical model and causal model assumptions[2], with efforts to apply flexible machine learning estimators to genetic data tending to focus on obtaining 'clinically relevant' genetic risk score estimates, rather than estimating causal effects.

With the underlying goal of applying mediation methods to genetic data, Chapter 3 considers modern 'natural' mediation estimands under semi-parametric models similar to those seen in GWAS and MR,

[1] Assumption-lean Cox regression (Vansteelandt et al., 2022), Longitudinal TMLE (van der Laan and Gruber, 2012), transportability of causal effects (Hernán and VanderWeele, 2011).

[2] GWAS and MR methods are usually based on linear mixed models, with MR requiring causal instrumental variable assumptions, often including e.g. ancestral homogeneity and non-existance of pleiotropic effects.

and close to the original work of Baron and Kenny (1986), which remains widely used, especially in applications to psychological data. Compared with the fully nonparametric theory, we show that 'partially linear' modelling assumptions circumvent the need for 'inverse density weighting', a common source of instability in the efficient estimator of mediation estimands under the nonparametric model. The approach taken in Chapter 3, therefore, represents a compromise between parametric and non-parametric reasoning, where the target of inference is a nonparametrically defined estimand, but semi-parametric models are assumed in order to facilitate inference, and ensure that the proposed robust/ assumption-lean methods are familiar to researchers in a range of fields.

The semi-parametric work presented in Chapter 3, also highlights that causal inference methods are equally applicable to parametric models as they are to nonparametric estimands. It is easy to lose sight of this fact, given that several of the most well known non-parametric estimands arise in the context of causal inference. The key difference between causal inference and nonparametric inference is that the latter solves a purely statistical problem, using methods that are agnostic to the estimand's causal interpretation. In this thesis we contribute to the pedagogy of both fields through the articles on causal methods for genetic data and on efficient influence curve based estimators, which are reproduced in Chapters 2 and 4. Aside from covering different topics, these two articles are also written for two distinct audiences: domain-specific data analysts, and statistical methodologists, both crucial to the adoption of statistical methodology.

In particular, Chapter 4 aims to demystify the technical nuances of deriving efficient estimators for nonparametric estimands. The first step for constructing such estimators, is to obtain the estimand's so-called efficient influence function, and we illustrate how this can be derived using the method of point mass contamination. Moreover, we build on familiar intuition from basic calculus and probability theory, and rederive several known efficient influence functions as examples. By focussing on the setting of nonparametric rather than semi-parametric modelling, we do not need to account for additional efficiency that is gained through the semi-parametric model assumptions, usually resulting in simpler derivations, which do not require manipulation of integral expressions.

One example where the nonparametric inference framework is readily applicable is to the novel treatment effect variable importance measures (TE-VIMs) described in Chapter 5. TE-VIMs are causally motivated estimands, which we propose for answering scientific questions regarding treatment effect heterogeneity. Like mediation analyses, treatment effect heterogeneity analyses provide insight into the the mechanism by which treatment affects outcome. Rather than analysing post-treatment variables, as in mediation analysis, however, the proposed TE-VIM esimands help identify covariates which are important in predicting the treatment effect, thereby helping to identify population strata which benefit most/ least from treatment.

An interesting feature of TE-VIM estimands is that, by definition, they lie on the bounded interval $[0, 1]$. For such estimands, approximate normality of the point estimator is rarely observed in finite samples when the true estimand value is close to one of the boundaries. This problem motivates the study of confidence interval construction methods, which do not rely on asymptomatic normality of the point estimator, unlike standard Wald intervals. Chapter 6 sets out a 'score based' proposal that builds on the generalised method of moments hypothesis testing framework, used in the semi-parametric mediation setting of Chapter 3. For nonparametric estimands that depend on unknown infinite dimensional parameters (functions), such as the average treatment effect, our proposal aligns closely with the principles of targeted learning. Compared with Wald based intervals, score based intervals also have appealing invariance properties, and require knowledge of the target efficient influence curve only up to constants of proportionality.

Returning to the contrast between causal inference and nonparametric inference methods, it is, I believe, often under appreciated that causal inference represents just one possible framework for motivating, deriving, and interpreting nonparametric estimands of scientific interest. This is evident in the work on derivative effect estimands presented in Chapter 7, though these estimands can also be ascribed a causal interpretation, as described in Chapter 8. The so-called least squares estimands, which we set out, are motivated with reference to 'least squares' projection, and through derivative effect efficiency arguments, rather than the potential outcomes etc. of causal inference. Our proposal also contributes to the longstanding problem of generalising to continuous exposures, treatment effect estimands which are now canonical in the setting of binary exposure.

I would like to conclude my thesis by sketching two ideas for future work, which may one day lead to a deeper understanding of 'nonparametric reasoning'. The first vignette builds on the derivative ideas in Chapters 7 and 8, connecting derivative estimands to mediation estimands through a kind of product-rule decomposition. In fact, my interest in derivative effect estimands for continuous exposures originally arose from considering how mediation estimands handle the 'effect' of a continuous mediator on outcome. The second vignette builds on the observation that Taylor expansions are commonly used to derive low-order polynomial approximations of nonlinear functions in the natural sciences. By considering a generalised version of Taylor's expansion, we hope to derive similar polynomial approximations for unknown statistical functionals, and hence derive/ motivate nonparametric estimands.

9.2 Derivative approach to mediation

Consider a random variable $Z = (Y, M, A, X)$ where $A \in \mathbb{R}$ is a continuous treatment, and $M \in \mathbb{R}$ is a continuous mediator, and $Y \in \mathbb{R}$ is an outcome. Letting $f(m|a, x)$ denote the mediator density conditional on (A, X), and $\mu(m, a, x) \equiv E(Y|M = m, A = a, X = x)$ then it follows by the product rule that

$$\begin{aligned}\frac{d}{da}E(Y|A=a,X=x) &= \frac{d}{da}\int \mu(m,a,x)f(m|a,x)dm \\ &= \underbrace{\int \left\{\frac{d\mu(m,a,x)}{da}\right\} f(m|a,x)dm}_{\approx \text{ direct effect}} + \underbrace{\int \mu(m,a,x)\left\{\frac{df(m|a,x)}{da}\right\} dm}_{\approx \text{ indirect effect}} \end{aligned} \tag{9.1}$$

The two terms which result from this decomposition have the feel of a direct and indirect effect respectively, since the former can be interpreted like the derivative change in the conditional mean outcome, when the mediator density is held fixed, and the latter like the conditional mean outcome when a derivative change is applied to the mediator density. The fact that the problem of mediation has the feel of a derivative decomposition has been noted elsewhere (Stolzenberg, 1980; Hayes and Preacher, 2010; Huber et al., 2020), albeit without mention of the decomposition in (9.1). Moreover, the additive nature of (9.1) is reminiscent of mediation proposals by Baron and Kenny (1986) and Pearl (2001), where mediation 'effects' sum to a 'total effect'.

To further connect product rule decompositions to natural mediation estimands, consider replacing the derivative operation above with a finite difference. Let Δ denote a finite difference operator, such that for an arbitrary function $f(u)$, we define $\Delta f(.) = f(1) - f(0)$. Unlike the derivative operator, the finite difference operator does not have a unique product rule

$$\begin{aligned}\Delta f(.)g(.) &= f(1)g(1) - f(0)g(0) \\ &= \{\Delta f(.)\}g(1) + f(0)\{\Delta g(.)\} \\ &= \{\Delta f(.)\}g(0) + f(1)\{\Delta g(.)\}\end{aligned}$$

with the final two expressions both representing equally valid product-rules. To reconcile this ambiguity, consider taking a linear combination of both rules with weights $p \in [0, 1]$ and $1 - p$. Doing so gives a set of product rules indexed by p

$$\Delta f(.)g(.) = \{\Delta f(.)\}\{pg(1) + (1-p)g(0)\} + \{pf(0) + (1-p)f(1)\}\{\Delta g(.)\}$$

In the setting of binary treatment, $A \in \{0, 1\}$, applying this product rule decomposition to $E(Y|A =$

$., X = x)$ gives

$$\begin{aligned}\Delta E(Y|A = ., X = x) &= \Delta \int \mu(m, ., x) f(m|., x) dm \\ &= \underbrace{\int \{\Delta\mu(m, ., x)\} \{pf(m|1, x) + (1 - p)f(m|0, x)\} dm}_{\approx \text{ direct effect}} \\ &+ \underbrace{\int \{p\mu(m, 0, x) + (1 - p)\mu(m, 1, x)\} \{\Delta f(m|., x)\} \, dm}_{\approx \text{ indirect effect}}\end{aligned}$$

Finally, since this decompositions is valid for all $p \in [0, 1]$, one could define p through a weight functions that plays an analogous role to the 'exposure weights' in Chapter 7. Taking the expectation of the decomposition above, and carefully choosing p, gives a generalised version of the mediation formula of Pearl (2001)

$$\begin{aligned}\underbrace{E\{\Delta E(Y|A = ., X = x)\}}_{\text{ATE}} &= \underbrace{\int \{\Delta\mu(m, ., x)\} \, w(a|x) dP(m, a, x)}_{\text{direct effect}} \\ &+ \underbrace{\int \mu(m, 1 - a, x) \left\{ \frac{\Delta f(m|., x)}{f(m|a, x)} \right\} w(a|x) dP(m, a, x)}_{\text{indirect effect}}\end{aligned} \tag{9.2}$$

where $w(a|x)$ is a non-negative weight that is normalised such that $E\{w(A|X)|X\} = 1$ and $dP(m, a, x)$ represents the joint probability measure over (M, A, X). For example, the weight $w(a|x) = a/E(A|X = x)$ recovers the mediation formula. We remark that, unlike the mediation formula, the identity in (9.2) is nonparametrically defined and makes no causal independence assumptions on the distribution of $Z = (Y, M, A, X)$.

9.3 Functional approximations

Consider that an analytic function $f(x)$ can be written as

$$f(x) = \sum_{i=0}^{\infty} a_i k_i(x) \tag{9.3}$$

where, for now, $k_n(x)$ is an undefined sequence of order n polynomials, and a_n represents a sequence of real numbers. Next, imagine that one constructs the polynomials sequence such that $k_0(x) = 1$ and for $n > 0$,

$$\begin{aligned}\frac{d}{dx} k_n(x) &= n k_{n-1}(x) \\ E\{k_n(X)\} &= 0\end{aligned} \tag{9.4}$$

where $X \in \mathbb{R}$ is a random variable with finite moment generating function $M_X(t) = E(e^{tX})$. Given these properties, differentiating (9.3) n times and taking the expectation of the resulting expression at $x = X$, gives $a_n = E\{f^{(n)}(X)\}/n!$ where $f^{(n)}(x)$ denotes the nth derivative of $f(x)$. Hence, the special case $f(x) = e^{tx}$ immediately gives a generating function for $k_n(x)$,

$$\sum_{n=0}^{\infty} \frac{t^n}{n!} k_n(x) = \frac{e^{tx}}{M_X(t)} \tag{9.5}$$

Polynomial sequences with generating functions of this form are known as Appell sequences, as defined by Appell (1880). We refer to Avram and Taqqu (1987); Ta (2015) and Appendix H.3 for properties of these polynomials in a probabilistic setting. The first few terms in the polynomial sequence, $k_n(x)$ are, $k_0(x) = 1$

$$\begin{aligned}
k_1(x) &= x - \kappa_1 \\
k_2(x) &= (x-\kappa_1)^2 - \kappa_2 \\
k_3(x) &= (x-\kappa_1)^3 - 3(x-\kappa_1)\kappa_2 - \kappa_3 \\
k_4(x) &= (x-\kappa_1)^4 - 6(x-\kappa_1)^2\kappa_2 - 4(x-\kappa_1)\kappa_3 + 3\kappa_2^2 - \kappa_4
\end{aligned}$$

where κ_n denotes the nth cumulant of X and we note that κ_1 and κ_2 respectively denote the mean and variance of X. Aside from the sequence $(x - x^*)^n$, which appears in the conventional Taylor series, some well-known Appell sequences are the Hermite, Euler and Bernoulli polynomials with moment generating functions that coincide with those of the standard normal, Bernoulli($p = 1/2$), and Uniform$(0, 1)$ distributions respectively. To progress beyond analytic functions, we consider truncating the sum in (9.3) to obtain functional approximations. The properties of the remainder under such a truncation are provided in Theorem 13 below.

Theorem 13 *Let X be a random variable with moment generating function $M_X(t)$. Then for all measurable functions, $f(x)$, which are $n \geq 1$ times differentiable on the support of X, there exists a function $R_n(x)$ such that*

$$f(x) = \sum_{j=0}^{n} \frac{E\{f^{(j)}(X)\}}{j!} k_j(x) + R_n(x) \tag{9.6}$$

and

$$\lim_{\delta \to 0} \frac{E\{R_n(X+\delta)\}}{\delta^n} = 0$$

where $k_n(x)$ is defined by (9.5)*. Proof in Appendix H.1.*

Theorem 13 has been studied before in the context of operational calculus, e.g. Roman and Gian-Carlo (1978) and Theorem 2 of Bourbaki (2004), however, to our knowledge, it has not yet been described in a probabilistic setting. We view this Theorem as a generalisation of Taylor's Theorem, since the latter represents a special case where $X \sim P$ is a probability point mass, i.e. when $P(X = x^*) = 1$, then the polynomials in (9.5) reduce to $k_n(x) = (x - x^*)^n$ and $E\{f^{(n)}(X)\} = f^{(n)}(x^*)$. The significance of Theorem 13 is that it tells us that the remainder, after truncating (9.3) at the nth term, is of the order δ^n in expectation when evaluated at the shifted random variable, $X + \delta$. In this way we view (9.6) as a Taylor expansion about a random distribution of points X, whereas the conventional Taylor's theorem considers an expansion about a single point, x^*.

For a measurable function $f(x)$, with derivative $f'(x)$, the linear approximation from the proposed expansion is

$$f(x) \approx E\{f(X)\} + E\{f'(X)\}\{x - E(X)\}.$$

This linear approximation may be used, for instance, to understand the derivative effect estimands of Chapters 7 and 8. In particular, consider replacing $f(.)$ above with the conditional response function $\mu(x, z) = E(Y|X = x, Z = z)$, where $(Y, X, Z) \sim P_0$ with $Y, X \in \mathbb{R}$, and $Z \in \mathbb{R}^d$ is a vector of covatiates. Letting $M_X(t, z) = E(e^{tX}|Z = z)$ be a moment function indexed by z, the linear expansion of the conditional response function is

$$\mu(x, z) \approx E\{\mu(X, Z)|Z = z\} + \underbrace{E\{\mu'(X, Z)|Z = z\}}_{\text{CADE}}\{x - E(X|Z = z)\} \tag{9.7}$$

where we recognise the conditional average derivative effect (CADE) on the right-hand side. Consequently we interpret the linear expansion in (9.7) as a formal first-order approximation to the condition response function, with a remainder which is 'small' in the sense described by Theorem 13. Such Taylor-like functional approximations might be extended in future work e.g. to consider other statistical functionals, higher order terms, or expansions about other random variables (such as the stochastic distributions presented in Chapter 8).

Bibliography

Abrevaya, J., Hsu, Y. C., and Lieli, R. P. (2015). Estimating Conditional Average Treatment Effects. *Journal of Business and Economic Statistics*, 33(4):485–505.

Alwin, D. F. and Hauser, R. M. (1975). The Decomposition of Effects in Path Analysis. *American Sociological Review*, 40(1):37.

Anderson, C. A., Pettersson, F. H., Clarke, G. M., Cardon, L. R., Morris, A. P., and Zondervan, K. T. (2010). Data quality control in genetic case-control association studies. *Nature Protocols*, 5(9):1564–1573.

Andrews, G. E. (1984). *The Theory of Partitions*. Cambridge University Press.

Appell, P. (1880). Sur une classe de polynômes. *Annales scientifiques de l'École normale supérieure*, 9:119–144.

Aroian, L. A. (1947). The Probability Function of the Product of Two Normally Distributed Variables. *The Annals of Mathematical Statistics*, 18(2):265–271.

Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences of the United States of America*, 113(27):7353–7360.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *Annals of Statistics*, 47(2):1179–1203.

Athey, S. and Wager, S. (2021). Policy Learning With Observational Data. *Econometrica*, 89(1):133–161.

Avagyan, V. and Vansteelandt, S. (2017). Honest data-adaptive inference for the average treatment effect under model misspecification using penalised bias-reduced double-robust estimation. Forthcoming in Biostatistics and Epidemiology.

Avram, F. and Taqqu, M. S. (1987). Noncentral Limit Theorems and Appell Polynomials. *The Annals of Probability*, 15(2):347–370.

Bahadur, R. R. and Savage, L. J. (1956). The Nonexistence of Certain Statistical Procedures in Nonparametric Problems. *The Annals of Mathematical Statistics*, 27(4):1115–1122.

Banerjee, A. N. (2007). A method of estimating the average derivative. *Journal of Econometrics*, 136(1):65–88.

Bareinboim, E., Tian, J., and Pearl, J. (2014). Recovering from Selection Bias in Causal and Statistical Inference Elias. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pages 2410 – 2416. AAAI Press.

Baron, R. M. and Kenny, D. A. (1986). The moderator mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6):1173–1182.

Benkeser, D. (2020). Nonparametric inference for interventional effects with multiple mediators. pages 1–27.

Berk, R., Buja, A., Brown, L., George, E., Kuchibhotla, A. K., Su, W., and Zhao, L. (2021). Assumption Lean Regression. *American Statistician*, 75(1):76–84.

Bickel, P. J., Klaassen, C. A., Bickel, P. J., Ritov, Y., Klaassen, J., Wellner, J. A., and Ritov, Y. (1993). *Efficient and adaptive estimation for semiparametric models*, volume 4. Johns Hopkins University Press Baltimore.

Bickel, P. J. and Lehmann, E. L. (1975). Descriptive Statistics for Nonparametric Models I. Introduction. *The Annals of Statistics*, 3(5):1038–1044.

Bourbaki, N. (2004). Generalized Taylor expansions Euler-Maclaurin summation formula. In *Elements of Mathematics Functions of a Real Variable*. Springer Berlin Heidelberg. [Translated by Spain, P. from the (1976) French original, *Fonctions d'une variable réelle. Éléments de mathématique.*].

Bowden, J., Smith, G. D., and Burgess, S. (2015). Mendelian randomization with invalid instruments: Effect estimation and bias detection through Egger regression. *International Journal of Epidemiology*, 44(2):512–525.

Brady, H. E. (2013). *Oxford Handbooks Online Causation and Explanation in Social Science 1 Causality*. Number April 2017.

Breiman, L. (2001a). Random Forests. *Machine Learning*, 45:5–32.

Breiman, L. (2001b). Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–215.

Browning, B. L. and Browning, S. R. (2011). A fast, powerful method for detecting identity by descent. *American Journal of Human Genetics*, 88(2):173–182.

Buja, A., Brown, L., Berk, R., George, E., Pitkin, E., Traskin, M., Zhang, K., and Zhao, L. (2019). Models as approximations I: Consequences illustrated with linear regression. *Statistical Science*, 34(4):523–544.

Buniello, A., Macarthur, J. A., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., Suveges, D., Vrousgou, O., Whetzel, P. L., Amode, R., Guillen, J. A., Riat, H. S., Trevanion, S. J., Hall, P., Junkins, H., Flicek, P., Burdett, T., Hindorff, L. A., Cunningham, F., and Parkinson, H. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, 47(D1):D1005–D1012.

Burgess, S. and Thompson, S. G. (2015). *Mendelian Randomization: Methods for Using Genetic Variants in Causal Estimation*. Chapman & Hall/CRC.

Cai, T. T., Levine, M., and Wang, L. (2009). Variance function estimation in multivariate nonparametric regression with fixed design. *Journal of Multivariate Analysis*, 100(1):126–136.

Carone, M., Díaz, I., and van der Laan, M. J. (2018). Higher-Order Targeted Loss-Based Estimation. In van der Laan, M. J. and Rose, S., editors, *Targeted Learning in Data Science*, Springer Series in Statistics, chapter 26. Springer International Publishing, Cham.

Cattaneo, M. D., Crump, R. K., and Jansson, M. (2010). Robust data-driven inference for density-weighted average derivatives. *Journal of the American Statistical Association*, 105(491):1070–1083.

Cattaneo, M. D., Crump, R. K., and Jansson, M. (2013). Generalized jackknife estimators of weighted average derivatives. *Journal of the American Statistical Association*, 108(504):1243–1256.

Chambaz, A., Neuvial, P., and van der Laan, M. J. (2012). Estimation of a non-parametric variable importance measure of a continuous exposure. *Electronic Journal of Statistics*, 6:1059–1099.

Chen, G., Zeng, D., and Kosorok, M. R. (2016). Personalized Dose Finding Using Outcome Weighted Learning. *Journal of the American Statistical Association*, 111(516):1509–1521.

Chen, H., Janizek, J. D., Lundberg, S., and Lee, S.-I. (2020). True to the model or true to the data? *arXiv preprint arXiv:2006.16234*.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., and Newey, W. (2017). Double/Debiased/Neyman Machine Learning of Treatment Effects. *American Economic Review*, 107(5):261–265.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, 21(1):C1–C68.

Chuang, Y. F., Tanaka, T., Beason-Held, L. L., An, Y., Terracciano, A., Sutin, A. R., Kraut, M., Singleton, A. B., Resnick, S. M., and Thambisetty, M. (2015). FTO genotype and aging: Pleiotropic longitudinal effects on adiposity, brain function, impulsivity and diet. *Molecular Psychiatry*, 20(1):133–139.

Corcoran, S. A. (1998). Bartlett adjustment of empirical discrepancy statistics. *Biometrika*, 85(4):967–972.

Coston, A., Kennedy, E. H., and Chouldechova, A. (2020). Counterfactual predictions under runtime confounding. *Advances in Neural Information Processing Systems*, 2020-December(NeurIPS).

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2006). Moving the Goalposts: Addressing Limited Overlap in the Estimation. *National Bureau of Economic Research*.

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199.

Cui, Y., Kosorok, M. R., Sverdrup, E., Wager, S., and Zhu, R. (2020). Estimating heterogeneous treatment effects with right-censored data via causal survival forests. *arXiv (2001.09887)*, pages 1–27.

Curth, A., Alaa, A. M., and van der Schaar, M. (2020). Semiparametric estimation and inference on structural target functions using machine learning and influence functions. *arXiv preprint arXiv:2008.06461*.

Díaz, I. and van der Laan, M. J. (2012). Population Intervention Causal Effects Based on Stochastic Interventions. *Biometrics*, 68(2):541–549.

Didelez, V., Meng, S., and Sheehan, N. A. (2010). Assumptions of IV methods for observational epidemiology. *Statistical Science*, 25(1):22–40.

Didelez, V. and Sheehan, N. (2007). Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research*, 16(4):309–330.

Ding, E. L., Song, Y., Manson, J. E., Hunter, D. J., Lee, C. C., Rifai, N., Buring, J. E., Gaziano, J. M., and Liu, S. (2009). Sex Hormone–Binding Globulin and Risk of Type 2 Diabetes in Women and Men. *New England Journal of Medicine*, 361(12):1152–1163.

Ding, P., Feller, A., and Miratrix, L. (2016). Randomization inference for treatment effect variation. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 78(3):655–671.

Drton, M. and Xiao, H. (2016). Wald tests of singular hypotheses. *Bernoulli*, 22(1):38–59.

Dufour, J.-M., Renault, E., and Zinde-Walsh, V. (2013). Wald tests when restrictions are locally singular.

Dufour, J. M., Trognon, A., and Tuvaandorj, P. (2017). Invariant tests based on M-estimators, estimating functions, and the generalized method of moments. *Econometric Reviews*, 36(1-3):182–204.

Dukes, O. and Vansteelandt, S. (2019). Uniformly valid confidence intervals for conditional treatment effects in misspecified high-dimensional models.

Feller, W. (1966). *An Introduction to Probability Theory and Its Applications*. John Wiley and Sons., New York, NY, 2 edition.

Fisher, A. and Kennedy, E. H. (2020). Visually Communicating and Teaching Intuition for Influence Functions. *The American Statistician*, pages 1–11.

Frey, J. and Zhang, Y. (2017). What Do Interpolated Nonparametric Confidence Intervals for Population Quantiles Guarantee? *American Statistician*, 71(4):305–309.

Fritz, M. S., Taylor, A. B., and MacKinnon, D. P. (2012). Explanation of Two Anomalous Results in Statistical Mediation Analysis. *Multivariate Behavioral Research*, 47(1):61–87.

Galvao, A. F. and Wang, L. (2015). Uniformly Semiparametric Efficient Estimation of Treatment Effects With a Continuous Treatment. *Journal of the American Statistical Association*, 110(512):1528–1542.

Ganeff, I. M. M., Bos, M. M., van Heemst, D., and Noordam, R. (2019). BMI-associated gene variants in FTO and cardiometabolic and brain disease: obesity or pleiotropy? *Physiological Genomics*, 51(8):311–322.

Giersbergen, N. (2014). Inference about the indirect effect: a likelihood approach. UvA-Econometrics Working Papers 14-10, Universiteit van Amsterdam, Dept. of Econometrics.

Gilbert, P. B., Montefiori, D. C., McDermott, A., Fong, Y., Benkeser, D., Deng, W., Zhou, H., Houchens, C. R., Martins, K., Jayashankar, L., Castellino, F., Flach, B., Lin, B. C., O'Connell, S., McDanal, C., Eaton, A., Sarzotti-Kelsoe, M., Lu, Y., Yu, C., Borate, B., van der Laan, L. W. P., Hejazi, N., Huynh, C., Miller, J., El Sahly, H. M., Baden, L. R., Baron, M., De La Cruz, L., Gay, C., Kalams, S., Kelley, C. F., Kutner, M., Andrasik, M. P., Kublin, J. G., Corey, L., Neuzil, K. M., Carpp, L. N., Pajon, R., Follmann, D., Donis, R. O., and Koup, R. A. (2021). Immune Correlates Analysis of the mRNA-1273 COVID-19 Vaccine Efficacy Trial. *medRxiv : the preprint server for health sciences*.

Glonek, G. F. V. (1993). On the Behaviour of Wald Statistics for the Disjunction of Two Regular Hypotheses. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(3):749–755.

Glymour, M. M. (2006). Using causal diagrams to understand common problems in social epidemiology. In *In Methods in Social Epidemiology*, pages 393–428. John Wiley and Sons.

Glymour, M. M. and Spiegelman, D. (2017). Evaluating public health interventions: 5. Causal inference in public health research-do sex, race, and biological factors cause health outcomes? *American Journal of Public Health*, 107(1):81–85.

Glymour, M. M., Tchetgen Tchetgen, E. J., and Robins, J. M. (2012). Credible mendelian randomization studies: Approaches for evaluating the instrumental variable assumptions. *American Journal of Epidemiology*, 175(4):332–339.

Grömping, U. (2009). Variable importance assessment in regression: Linear regression versus random forest. *American Statistician*, 63(4):308–319.

Gudbjartsson, D. F., Walters, G. B., Thorleifsson, G., Stefansson, H., Halldorsson, B. V., Zusmanovich, P., Sulem, P., Thorlacius, S., Gylfason, A., Steinberg, S., Helgadottir, A., Ingason, A., Steinthorsdottir, V., Olafsdottir, E. J., Olafsdottir, G. H., Jonsson, T., Borch-Johnsen, K., Hansen, T., Andersen, G., Jorgensen, T., Pedersen, O., Aben, K. K., Witjes, J. A., Swinkels, D. W., Heijer, M. D., Franke, B., Verbeek, A. L., Becker, D. M., Yanek, L. R., Becker, L. C., Tryggvadottir, L., Rafnar, T., Gulcher, J., Kiemeney, L. A., Kong, A., Thorsteinsdottir, U., and Stefansson, K. (2008). Many sequence variants affecting diversity of adult human height. *Nature Genetics*, 40(5):609–615.

Hammer, S. M., Katzenstein, D. A., Hughes, M. D., Gundacker, H., Schooley, R. T., Haubrich, R. H., Henry, W. K., Lederman, M. M., Phair, J. P., Niu, M., Hirsch, M. S., and Merigan, T. C. (1996). Therapy in Hiv-Infected Adults With Cd4 Cell Counts. *The New England Journal of Medicine*, 335:1081–1090.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393.

Hansen, L. P. (1982). Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica*, 50(4):1029.

Hansen, L. P., Heaton, J., and Yaron, A. (1996). Finite-sample properties of some alternative GMM estimators. *Journal of Business and Economic Statistics*, 14(3):262–280.

Härdle, W., Hildenbrand, W., and Jerison, M. (1991). Empirical Evidence on the Law of Demand. *Econometrica*, 59(6):1525.

Härdle, W. and Stoker, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association*, 84(408):986–995.

Hayes, A. F. (2018). *Introduction to mediation, moderation, and conditional process analysis: a regression-based approach*. Guilford Press, New York, NY, 2 edition.

Hayes, A. F. and Preacher, K. J. (2010). Quantifying and testing indirect effects in simple mediation models when the constituent paths are nonlinear. *Multivariate Behavioral Research*, 45(4):627–660.

Heckman, J. J., Smith, J., and Clements, N. (1997). Making the Most out of Programme Evaluations and Social Experiments : Accounting for Heterogeneity in Programme Impacts. *Review of Economic Studies*, 64(4):487–535.

Hemani, G., Zheng, J., Elsworth, B., Wade, K. H., Haberland, V., Baird, D., Laurin, C., Burgess, S., Bowden, J., Langdon, R., Tan, V. Y., Yarmolinsky, J., Shihab, H. A., Timpson, N. J., Evans, D. M., Relton, C., Martin, R. M., Davey Smith, G., Gaunt, T. R., and Haycock, P. C. (2018). The mr-base platform supports systematic causal inference across the human phenome. *eLife*, 7:e34408.

Hernán, M. A. and Robins, J. M. (2006). Estimating causal effects from epidemiological data. *Journal of Epidemiology & Community Health*, 60(7):578–586.

Hernán, M. A. and Robins, J. M. (2020). *Causal Inference: What If*. Boca Raton: Chapman and Hall/CRC.

Hernán, M. A. and VanderWeele, T. J. (2011). Compound treatments and transportability of causal inference. *Epidemiology*, 22(3):368–377.

Hines, O., Diaz-Ordaz, K., and Vansteelandt, S. (2021a). Parameterising the effect of a continuous exposure using average derivative effects. *arXiv (2109.13124)*, pages 1–25.

Hines, O., Dukes, O., Diaz-Ordaz, K., and Vansteelandt, S. (2022). Demystifying statistical learning based on efficient influence functions. *The American Statistician*, 0(0):1–48.

Hines, O., Vansteelandt, S., and Diaz-Ordaz, K. (2021b). Robust Inference for Mediated Effects in Partially Linear Models. *Psychometrika*, 86(2):595–618.

Hirano, K. and Imbens, G. W. (2005). The Propensity Score with Continuous Treatments. *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An Essential Journey with Donald Rubin's Statistical Family*, (2001):73–84.

Hirshberg, D. A. and Wager, S. (2018). Debiased Inference of Average Partial Effects in Single-Index Models. (1):1–10.

Hoffman, G. E. (2013). Correcting for Population Structure and Kinship Using the Linear Mixed Model: Theory and Extensions. *PLoS ONE*, 8(10):e75707.

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, 81(396):945–960.

Hotelling, H. (1931). The Generalization of Student's Ratio. *The Annals of Mathematical Statistics*, 2(3):360–378.

Huber, M., Hsu, Y. C., Lee, Y. Y., and Lettry, L. (2020). Direct and indirect effects of continuous treatments based on generalized propensity score weighting. *Journal of Applied Econometrics*, (May 2019):1–27.

Hutson, A. D. (1999). Calculating nonparametric confidence intervals for quantiles using fractional order statistics. *Journal of Applied Statistics*, 26(3):343–353.

Ichimura, H. and Newey, W. K. (2022). The influence function of semiparametric estimators. *Quantitative Economics*, 13(1):29–61.

Imai, K., Keele, L., and Tingley, D. (2010a). A General Approach to Causal Mediation Analysis. *Psychological Methods*, 15(4):309–334.

Imai, K., Keele, L., and Yamamoto, T. (2010b). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, 25(1):51–71.

Imai, K. and Van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, 99(467):854–866.

Imbens, G. W. (1997). One-Step Method of Estimators Moments for Models Generalized. *Review of Economic Studies*, 383(64):359–383.

Imbens, G. W. (2002). Generalized method of moments and empirical likelihood. *Journal of Business and Economic Statistics*, 20(4):493–506.

Imbens, G. W. (2019). Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *NBER Working Paper No.w26104*.

International Warfarin Pharmacogenetics Consortium (2009). Estimation of the Warfarin Dose with Clinical and Pharmacogenetic Data. *New England Journal of Medicine*, 360(8):753–764.

Janzing, D., Minorics, L., and Blöbaum, P. (2020). Feature relevance quantification in explainable ai: A causal problem. In *International Conference on Artificial Intelligence and Statistics*, pages 2907–2916. PMLR.

Kallus, N. (2020). More Efficient Policy Learning via Optimal Retargeting. *Journal of the American Statistical Association*, 0(0):1–34.

Kallus, N., Mao, X., and Zhou, A. (2018). Interval estimation of individual-level causal effects under unobserved confounding. *arXiv*, pages 1–32.

Kang, J. D. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22(4):523–539.

Kennedy, E. H. (2015). Semiparametric theory and empirical processes in causal inference.

Kennedy, E. H. (2019). Nonparametric Causal Effects Based on Incremental Propensity Score Interventions. *Journal of the American Statistical Association*, 114(526):645–656.

Kennedy, E. H. (2020). Optimal doubly robust estimation of heterogeneous causal effects. *arXiv (2004.14497)*, pages 1–35.

Kennedy, E. H. (2022). Semiparametric doubly robust targeted double machine learning: a review. *arXiv (2203.06469)*, pages 1–37.

Kennedy, E. H., Ma, Z., McHugh, M. D., and Small, D. S. (2017). Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 79(4):1229–1245.

Kenny, D. A. and Judd, C. M. (2014). Power Anomalies in Testing Mediation. *Psychological Science*, 25(2):334–339.

Kitamura, Y. and Stutzer, M. (1997). An Information-Theoretic Alternative to Generalized Method of Moments Estimation. *Econometrica*, 65(4):861–874.

Kivimáki, M., Lawlor, D. A., Smith, G. D., Kumari, M., Donald, A., Britton, A., Casas, J. P., Shah, T., Brunner, E., Timpson, N. J., Halcox, J. P., Miller, M. A., Humphries, S. E., Deanfield, J., Marmot, M. G., and Hingorani, A. D. (2008). Does high C-reactive protein concentration increase atherosclerosis? The Whitehall II study. *PLoS ONE*, 3(8):1–8.

Knaus, M. C., Lechner, M., and Strittmatter, A. (2021). Machine learning estimation of heterogeneous causal effects: Empirical Monte Carlo evidence. *Econometrics Journal*, 24(1):134–161.

Koh, P. W. and Liang, P. (2017). Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, pages 1885–1894. PMLR.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences of the United States of America*, 116(10):4156–4165.

Lanke, J. (1974). Interval Estimation of a Median. *Scandinavian Journal of Statistics*, 1(1):28–32.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-Free Predictive Inference for Regression. *Journal of the American Statistical Association*, 113(523):1094–1111.

Levy, J. (2019). Tutorial: Deriving The Efficient Influence Curve for Large Models.

Levy, J. and van der Laan, M. J. (2018). Kernel Smoothing of the Treatment Effect CDF. *arXiv*.

Levy, J., van der Laan, M. J., Hubbard, A., and Pirracchio, R. (2021). A fundamental measure of treatment effect heterogeneity. *Journal of Causal Inference*, 9(1):83–108.

Li, F., Morgan, K. L., and Zaslavsky, A. M. (2018). Balancing Covariates via Propensity Score Weighting. *Journal of the American Statistical Association*, 113(521):390–400.

Lin, D. Y. and Zeng, D. (2009). Proper analysis of secondary phenotype data in case-control association studies. *Genetic Epidemiology*, 33(3):256–265.

Lippert, C., Quon, G., Kang, E. Y., Kadie, C. M., Listgarten, J., and Heckerman, D. (2013). The benefits of selecting phenotype-specific variants for applications of mixed models in genomics. *Scientific Reports*, 3:1815.

Liu, L., Shahn, Z., Robins, J. M., and Rotnitzky, A. (2021). Efficient estimation of optimal regimes under a no direct effect assumption. *Journal of the American Statistical Association*, 116(533):224–239.

Loos, R. J., Lindgren, C. M., Li, S., Wheeler, E., Hua Zhao, J., Prokopenko, I., Inouye, M., Freathy, R. M., Attwood, A. P., Beckmann, J. S., Berndt, S. I., Bergmann, S., Bennett, A. J., Bingham, S. A., Bochud, M., Brown, M., Cauchi, S., Connell, J. M., Cooper, C., Davey Smith, G., Day, I., Dina, C., De, S., Dermitzakis, E. T., Doney, A. S., Elliott, K. S., Elliott, P., Evans, D. M., Sadaf Farooqi, I., Froguel, P., Ghori, J., Groves, C. J., Gwilliam, R., Hadley, D., Hall, A. S., Hattersley, A. T., Hebebrand, J., Heid, I. M., Herrera, B., Hinney, A., Hunt, S. E., Jarvelin, M. R., Johnson, T., Jolley, J. D., Karpe, F., Keniry, A., Khaw, K. T., Luben, R. N., Mangino, M., Marchini, J., McArdle, W. L., McGinnis, R., Meyre, D., Munroe, P. B., Morris, A. D., Ness, A. R., Neville, M. J., Nica, A. C., Ong, K. K., O'Rahilly, S., Owen, K. R., Palmer, C. N., Papadakis, K., Potter, S., Pouta, A., Qi, L., Randall, J. C., Rayner, N. W., Ring, S. M., Sandhu, M. S., Scherag, A., Sims, M. A., Song, K., Soranzo, N., Speliotes, E. K., Syddall, H. E., Teichmann, S. A., Timpson, N. J., Tobias, J. H., Uda, M., Ganz Vogel, C. I., Wallace, C., Waterworth, D. M., Weedon, M. N., Willer, C. J., Wraight, V. L., Yuan, X., Zeggini, E., Hirschhorn,

J. N., Strachan, D. P., Ouwehand, W. H., Caulfield, M. J., Samani, N. J., Frayling, T. M., Vollenweider, P., Waeber, G., Mooser, V., Deloukas, P., McCarthy, M. I., Wareham, N. J., Barroso, I., Jacobs, K. B., Chanock, S. J., Hayes, R. B., Lamina, C., Gieger, C., Illig, T., Meitinger, T., Wichmann, H. E., Kraft, P., Hankinson, S. E., Hunter, D. J., Hu, F. B., Lyon, H. N., Voight, B. F., Ridderstrale, M., Groop, L., Scheet, P., Sanna, S., Abecasis, G. R., Albai, G., Nagaraja, R., Schlessinger, D., Jackson, A. U., Tuomilehto, J., Collins, F. S., Boehnke, M., and Mohlke, K. L. (2008). Common variants near MC4R are associated with fat mass, weight and risk of obesity. *Nature Genetics*, 40(6):768–775.

Lu, W., Zhang, H. H., and Zeng, D. (2013). Variable selection for optimal treatment decision. *Statistical Methods in Medical Research*, 22(5):493–504.

Luedtke, A. R. and van der Laan, M. J. (2016). Super-Learning of an Optimal Dynamic Treatment Rule. *International Journal of Biostatistics*, 12(1):305–332.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, pages 4768–4777.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis.* Multivariate applications series. Taylor & Francis Group/Lawrence Erlbaum Associates, New York, NY.

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., and Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, 7(1):83–104.

Mosteller, F. (1946). On Some Useful "Inefficient" Statistics. *The Annals of Mathematical Statistics*, 17(4):377 – 408.

Munafò, M. R., Tilling, K., Taylor, A. E., Evans, D. M., and Davey Smith, G. (2018). Collider scope: when selection bias can substantially influence observed associations. *International Journal of Epidemiology*, 47(1):226–235.

Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355.

Naimi, A. I., Cole, S. R., and Kennedy, E. H. (2017). An introduction to g methods. *International journal of epidemiology*, 46(2):756–762.

Neugebauer, R. and van der Laan, M. J. (2007). Nonparametric causal effects based on marginal structural models. *Journal of Statistical Planning and Inference*, 137(2):419–434.

Newey, W. K. and Robins, J. M. (2018). Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv*, pages 1–43.

Newey, W. K. and Smith, R. J. (2004). Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica*, 72(1):219–255.

Newey, W. K. and Stoker, T. M. (1993). Efficiency of Weighted Average Derivative Estimators and Index Models. *Econometrica*, 61(5):1199.

Newey, W. K. and West, K. D. (1987). Hypothesis Testing with Efficient Method of Moments Estimation. *International Economic Review*, 28(3):777.

Neyman, J. (1959). Optimal Asymptotic Tests of Composite Statistical Hypotheses. In Grenander, U., editor, *Probability and Statistics: The Harald Cramer Volume*, pages 213–234. Almqvist and Wiskell, Stockholm.

Nie, X. and Wager, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2):237–249.

Owen, A. B. and Prieur, C. (2017). On shapley value for measuring importance of dependent inputs. *SIAM-ASA Journal on Uncertainty Quantification*, 5(1):986–1002.

Pearl, J. (1986). Fusion, propagation, and structuring in belief networks. *Artificial Intelligence*, 29(3):241–288.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.

Pearl, J. (2000). *Causality: Models, Reasoning and Inference*.

Pearl, J. (2001). Direct and Indirect Effects. In *Proc. 17th Conference on Uncertainy in Articial Intelligence*, pages 411–420. San Francisco. Morgan Kaufmann.

Petersen, M. L. and van der Laan, M. J. (2014). Causal models and learning from data: Integrating causal modeling and statistical estimation. *Epidemiology*, 25(3):418–426.

Pfanzagl, J. (1990). Estimation in semiparametric models. In *Estimation in Semiparametric Models*, pages 17–22. Springer.

Pfanzagl, J. and Wefelmeyer, W. (1985). Contributions to a general asymptotic statistical theory. *Statistics & Risk Modeling*, 3(3-4):379–388.

Powell, J. L., Stock, J. H., and Stoker, T. M. (1989). Semiparametric Estimation of Index Coefficients. *Econometrica*, 57(6):1403.

Qin, J. and Lawless, J. (1994). Empirical Likelihood and General Estimating Equations. *The Annals of Statistics*, 22(1).

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. *NAACL-HLT 2016 - 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstrations Session*, pages 97–101.

Robins, J. M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics - Theory and Methods*, 23(8):2379–2412.

Robins, J. M. (2001). Data, design, and background knowledge in etiologic inference. *Epidemiology*, 12(3):313–320.

Robins, J. M. and Greenland, S. (1992). Identifiability and Exchangeability for Direct and Indirect Effects. *Epidemiology*, 3(2):143–155.

Robins, J. M., Li, L., Tchetgen Tchetgen, E. J., and van der Vaart, A. W. (2008). Higher order influence functions and minimax estimation of nonlinear functionals. *Probability and Statistics: Essays in Honor of David A. Freedman*, 2:335–421.

Robins, J. M. and Ritov, Y. (1997). Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in Medicine*, 16(1-3):285–319.

Robins, J. M. and Rotnitzky, A. (2001). Inference for semiparametric models: Some questions and an answer - Comments. *Statistica Sinica*, 11(4):920–936.

Robins, J. M. and van der Vaart, A. (2006). Adaptive nonparametric confidence sets. *Annals of Statistics*, 34(1):229–253.

Robinson, P. M. (1988). Root-N-Consistent Semiparametric Regression. *Econometrica*, 56(4):931.

Roman, S. M. and Gian-Carlo, R. (1978). The Umbral Calculus. *Advances in Mathematics*, 27:95 – 188.

Rosenkranz, G. (2020). *Exploratory Subgroup Analyses in Clinical Research*. Wiley.

Rothenhäusler, D. and Yu, B. (2019). Incremental causal effects. pages 1–34.

Rothwell, P. M. (2005). Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *The Lancet*, 365(9454):176–186.

Rotnitzky, A., Lei, Q., Sued, M., and Robins, J. M. (2012). Improved double-robust estimation in missing data and causal inference models. *Biometrika*, 99(2):439–456.

Rotnitzky, A., Li, L., and Li, X. (2010). A note on overadjustment in inverse probability weighted estimation. *Biometrika*, 97(4):997–1001.

Rotnitzky, A., Smucler, E., and Robins, J. M. (2021). Characterization of parameters with a mixed bias property. *Biometrika*, 108(1):231–238.

Rotnitzky, A. and Vansteelandt, S. (2014). Double-robust methods. In Molenberghs, G., Fitzmaurice, G., Kenward, M., Tsiatis, A., and Verbeke, G., editors, *Handbook of missing data methodology*, chapter 9, pages 185–212. CRC Press, New York.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331.

Schulz, J. and Moodie, E. E. (2021). Doubly Robust Estimation of Optimal Dosing Strategies. *Journal of the American Statistical Association*, 116(533):256–268.

Senn, S. (2001). Individual Therapy: New Dawn or False Dawn? *Drug Information Journal*, 35(4):1479–1494.

Shah, R. D. and Peters, J. (2018). The Hardness of Conditional Independence Testing and the Generalised Covariance Measure. (July 2017).

Sheehan, N. A. and Didelez, V. (2018). Human Genetics Epidemiology, Genetic Epidemiology and Mendelian Randomisation: more need than ever to attend to detail? Epidemiology, Genetic Epidemiology and Mendelian Randomisation: more need than ever to attend to detail? *Human Genetics*, (0123456789).

Shen, Y., Gao, C., Witten, D., and Han, F. (2020). Optimal estimation of variance in nonparametric regression with random design. *The Annals of Statistics*, 48(6):3589–3618.

Slamon, D. J., Leyland-Jones, B., Shak, S., Fuchs, H., Paton, V., Bajamonde, A., Fleming, T., Eiermann, W., Wolter, J., Pegram, M., Baselga, J., and Norton, L. (2001). Use of Chemotherapy plus a Monoclonal Antibody against HER2 for Metastatic Breast Cancer That Overexpresses HER2. *New England Journal of Medicine*, 344(11):783–792.

Sobel, M. E. (1982). Asymptotic Confidence Intervals for Indirect Effects in Structural Equation Models. *Sociological Methodology*, 13(1982):290.

Song, X., Ionita-Laza, I., Liu, M., Reibman, J., and Wei, Y. (2016). A general and robust framework for secondary traits analysis. *Genetics*, 202(4):1329–1343.

Speed, D. and Balding, D. J. (2014). Relatedness in the post-genomic era : is it still useful ? *Nature Publishing Group*, (November):1–12.

Speed, D., Hemani, G., Johnson, M. R., and Balding, D. J. (2012). Improved heritability estimation from genome-wide SNPs. *American Journal of Human Genetics*, 91(6):1011–1021.

Spiller, W., Slichter, D., Bowden, J., and Davey Smith, G. (2018). Detecting and correcting for bias in Mendelian randomization analyses using Gene-by-Environment interactions. *International Journal of Epidemiology*, pages 1–11.

Stigler, S. M. (1977). Fractional order statistics, with applications. *Journal of the American Statistical Association*, 72(359):544–550.

Stolzenberg, R. M. (1980). The Measurement and Decomposition of Causal Effects in Nonlinear and Nonadditive Models. *Sociological Methodology*, 11:459.

Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(1):25.

Swerdlow, D. I., Kuchenbaecker, K. B., Shah, S., Sofat, R., Holmes, M. V., White, J., Mindell, J. S., Kivimaki, M., Brunner, E. J., Whittaker, J. C., Casas, J. P., and Hingorani, A. D. (2016). Selecting instruments for Mendelian randomization in the wake of genome-wide association studies. *International Journal of Epidemiology*, 45(5):1600–1616.

Syrgkanis, V., Lei, V., Oprescu, M., Hei, M., Battocchi, K., and Lewis, G. (2019). Machine learning estimation of heterogeneous treatment effects with instruments. *Advances in Neural Information Processing Systems*, 32.

Ta, B. Q. (2015). Probabilistic approach to Appell polynomials. *Expositiones Mathematicae*, 33(3):269–294.

Taylor, S. J., Carnes, D., Homer, K., Pincus, T., Kahan, B. C., Hounsome, N., Eldridge, S., Spencer, A., Diaz-Ordaz, K., Rahman, A., Mars, T. S., Foell, J., Griffiths, C. J., and Underwood, M. R. (2016). Improving the self-management of chronic pain: COping with persistent Pain, Effectiveness Research in Self-management (COPERS). *Programme Grants for Applied Research*, 4(14):1–440.

Tchetgen Tchetgen, E. J. and Shpitser, I. (2012). Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness and sensitivity analysis. *Annals of Statistics*, 40(3):1816–1845.

Tchetgen Tchetgen, E. J. and Shpitser, I. (2014). Estimation of a semiparametric natural direct effect model incorporating baseline covariates. *Biometrika*, 101(4):849–864.

Tchetgen Tchetgen, E. J., Walter, S., Vansteelandt, S., Martinussen, T., and Glymour, M. (2015). Instrumental Variable Estimation in a Survival Context. *Epidemiology*, 26(3):402–410.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. Springer Series in Statistics. Springer New York, New York, NY.

van der Laan, M. J. (2013). Targeted Learning of an Optimal Dynamic Treatment , and Statistical Inference for its Mean Outcome Targeted Learning of an Optimal Dynamic Treatment , and Statistical Inference for its Mean Outcome. *UC Berkeley Division of Biostatistics Working Paper Series*, (317):1–90.

van der Laan, M. J. (2015). Statistics as a science, not an art: the way to survive in data science. *Amstat News*, 1.

van der Laan, M. J. and Gruber, S. (2012). Targeted minimum loss based estimation of causal effects of multiple time point interventions. *The International Journal of Biostatistics*, 8(1).

van der Laan, M. J. and Gruber, S. (2016). One-Step Targeted Minimum Loss-based Estimation Based on Universal Least Favorable One-Dimensional Submodels. *International Journal of Biostatistics*, 12(1):351–378.

van der Laan, M. J. and Luedtke, A. R. (2014). Targeted Learning of the Mean Outcome under an Optimal Dynamic Treatment Rule. *Journal of Causal Inference*, 3(1):61–95.

van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1).

van der Laan, M. J. and Rose, S. (2011). *Targeted Learning*, volume 27 of *Springer Series in Statistics*. Springer New York, New York, NY.

van der Laan, M. J. and Rubin, D. B. (2006). Targeted Maximum Likelihood Learning. *The International Journal of Biostatistics*.

van der Vaart, A. W. (1998a). Empirical Processes. In *Asymptotic Statistics*, pages 265–290. Cambridge University Press.

van der Vaart, A. W. (1998b). Functional delta method. In *Asymptotic Statistics*, page 291–303. Cambridge University Press.

van Garderen, K. J. and van Giersbergen, N. (2019). Almost Similar Tests for Mediation Effects Hypotheses with Singularities.

Vandenbroucke, J. P., Broadbent, A., and Pearce, N. (2016). Causality and causal inference in epidemiology: The need for a pluralistic approach. *International Journal of Epidemiology*, 45(6):1776–1786.

VanderWeele, T. J., Tchetgen Tchetgen, E. J., Cornelis, M., and Kraft, P. (2014). Methodological challenges in mendelian randomization. *Epidemiology (Cambridge, Mass.)*, 25(3):427.

VanderWeele, T. J. and Vansteelandt, S. (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface*, 2(4):457–468.

Vansteelandt, S. (2012). Understanding counterfactual-based mediation analysis approaches and their differences. *Epidemiology*, 23(6):889–891.

Vansteelandt, S. and Daniel, R. M. (2017). Interventional Effects for Mediation Analysis with Multiple Mediators. *Epidemiology*, 28(2):258–265.

Vansteelandt, S. and Dukes, O. (2022). Assumption-lean inference for generalised linear model parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(3):657–685.

Vansteelandt, S., Dukes, O., Lancker, K. V., and Martinussen, T. (2022). Assumption-lean cox regression. *Journal of the American Statistical Association*, 0(0):1–10.

Vansteelandt, S., Dukes, O., and Martinussen, T. (2018). Survivor bias in Mendelian randomization analysis. *Biostatistics*, 19(4):426–443.

Vansteelandt, S. and Joffe, M. (2014). Structural Nested Models and G-estimation: The Partially Realized Promise. *Statistical Science*, 29(4):707–731.

Verdinelli, I. and Wasserman, L. (2021). Decorrelated Variable Importance. *arXiv (2111.10853)*, pages 1–26.

Vermeulen, K. and Vansteelandt, S. (2015). Bias-Reduced Doubly Robust Estimation. *Journal of the American Statistical Association*, 110(511):1024–1036.

Verzelen, N. and Gassiat, E. (2018). Adaptive estimation of high-dimensional signal-to-noise ratios. *Bernoulli*, 24(4B):3683 – 3710.

Vilhjálmsson, B. J. and Nordborg, M. (2013). The nature of confounding in genome-wide association studies. *Nature Reviews Genetics*, 14(1):1–2.

Wager, S. and Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Wallace, M. P., Moodie, E. E., and Stephens, D. A. (2018). Reward ignorant modeling of dynamic treatment regimes. *Biometrical Journal*, 60(5):991–1002.

Wang, K. (2018). Understanding Power Anomalies in Mediation Analysis. *Psychometrika*, 83(2):387–406.

Wang, L., Brown, L. D., Cai, T. T., and Levine, M. (2008). Effect of mean on variance function estimation in nonparametric regression. *The Annals of Statistics*, 36(2):646–664.

Wasserman, L. (2006). *All of Nonparametric Statistics*. Springer Texts in Statistics. Springer New York.

Weedon, M. N., Lettre, G., Freathy, R. M., M, C., Voight, B. F., Perry, J. R. B., Elliott, K. S., Guiducci, C., Shields, B., Zeggini, E., Lango, H., Lyssenko, V., Timpson, N. J., Burtt, N. P., Rayner, N. W., Ardlie, K., Tobias, J. H., Ness, A. R., and Ring, S. M. (2011). A common variant of HMGA2 is associated with adult and childhood height in the general population. *Nature Genetics*, 39(10):1245–1250.

Willer, C. J., Sanna, S., Jackson, A. U., Scuteri, A., Bonnycastle, L. L., Clarke, R., Heath, S. C., Timpson, N. J., Najjar, S. S., Stringham, H. M., Strait, J., Duren, W. L., Maschio, A., Busonero, F., Mulas, A., Albai, G., Swift, A. J., Morken, M. A., Narisu, N., Bennett, D., Parish, S., Shen, H., Galan, P., Meneton, P., Hercberg, S., Zelenika, D., Chen, W. M., Li, Y., Scott, L. J., Scheet, P. A., Sundvall, J., Watanabe, R. M., Nagaraja, R., Ebrahim, S., Lawlor, D. A., Ben-Shlomo, Y., Davey-Smith, G., Shuldiner, A. R., Collins, R., Bergman, R. N., Uda, M., Tuomilehto, J., Cao, A., Collins, F. S., Lakatta, E., Lathrop, G. M., Boehnke, M., Schlessinger, D., Mohlke, K. L., and Abecasis, G. R. (2008). Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nature Genetics*, 40(2):161–169.

Williamson, B. D. and Feng, J. (2020). Efficient nonparametric statistical inference on population feature importance using Shapley values. *37th International Conference on Machine Learning, ICML 2020*, PartF168147-14:10213–10222.

Williamson, B. D., Gilbert, P. B., Carone, M., and Simon, N. (2021a). Nonparametric variable importance assessment using machine learning techniques. *Biometrics*, 77(1):9–22.

Williamson, B. D., Gilbert, P. B., Simon, N. R., and Carone, M. (2021b). A General Framework for Inference on Algorithm-Agnostic Variable Importance. *Journal of the American Statistical Association*, 0(0):1–38.

Windmeijer, F., Farbmacher, H., Davies, N., and Davey Smith, G. (2018). On the Use of the Lasso for Instrumental Variables Estimation with Some Invalid Instruments. *Journal of the American Statistical Association*, 1459.

Wood, S. N., Pya, N., and Säfken, B. (2016). Smoothing Parameter and Model Selection for General Smooth Models. *Journal of the American Statistical Association*, 111(516):1548–1563.

Wooldridge, J. M. and Zhu, Y. (2020). Inference in Approximately Sparse Correlated Random Effects Probit Models With Panel Data. *Journal of Business and Economic Statistics*, 38(1):1–18.

Wright, M. N. and Ziegler, A. (2017). Ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1).

Yu, K. and Zhang, J. (2005). A three-parameter asymmetric laplace distribution and its extension. *Communications in Statistics - Theory and Methods*, 34(9-10):1867–1879.

Zhang, L. and Janson, L. (2020). Floodgate: inference for model-free variable importance. *arXiv (2007.01283)*.

Zhang, M., Tsiatis, A. A., and Davidian, M. (2008). Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics*, 64(3):707–715.

Zhang, Y., Laber, E. B., Tsiatis, A., and Davidian, M. (2015). Using decision lists to construct interpretable and parsimonious treatment regimes. *Biometrics*, 71(4):895–904.

Zheng, W. and van der Laan, M. J. (2011). Cross-Validated Targeted Minimum-Loss-Based Estimation. In *Targeted Learning*, pages 459–474. Springer New York, New York, NY.

Zhou, X. and Stephens, M. (2014). Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature Methods*, 11(4):407–409.

Appendix A

Supplement to causality in genetics

A.1 Linear Mixed Models

Consider again the linear model in Eq.2.2. When the model parameters are estimated by OLS, one effectively makes no prior assumptions about the parameter values, other than that they are fixed to some true unknown value. Considering P as a random effect, however, we impose, in a Bayesian sense, a normally distributed prior for $\gamma \sim \mathcal{N}_p\left(0, \sigma_g^2 I_p\right)$, where I_p is a p by p identity matrix, σ_g^2 is a hyper parameter and $\mathcal{N}_p\left(\mu, \Sigma\right)$ is a p-multivariate normal distribution with mean μ and variance Σ.

By making this prior assumption we arrive at a LMM, which may be written as a model for the full n-dimensional observed phenotype vector, $\boldsymbol{Y}$. Here bold notation is used to refer to vector (or matrix) quantities with n entries (or rows), each representing a single individual in the cohort. Again $\boldsymbol{I_n}$ is the n by n identity matrix,

$$\boldsymbol{Y} \sim \mathcal{N}_n\left(\alpha \boldsymbol{G} + \boldsymbol{E}\beta, \sigma_g^2 \boldsymbol{K} + \sigma_e^2 \boldsymbol{I_n}\right) \tag{A.1}$$

where $\boldsymbol{K} = \boldsymbol{P}\boldsymbol{P}^\top$ and $\boldsymbol{P}$ is an n by q matrix where each row represents the vector of PCs for a particular individual. The n by n matrix, $\boldsymbol{K}$ is referred to as the genetic similarity matrix, since the entry K_{ij} is a measure of the genetic similarity between the i^{th} and j^{th} individuals in the cohort, obtained by comparing their PCs. In general one is not restricted to using PCs to define the genetic similarity matrix. In fact several different methods can be expressed by the LMM equation above, using different measures of genetic similarity Hoffman (2013).

Measures of Genetic Similarity

Methods for measuring genetic similarity may be broadly separated into two categories: Those related to the Principal Component Analysis (Principal Components like), and those where some biologically motivated measure of genetic similarity is made. We will refer to methods of the latter type as Identity By Descent like, since they often measure similarity by finding genetic regions which are thought to be identical by descent in two individuals. A brief overview of these approaches is provided below.

Principal Component like

In a conventional PC analysis, the variables from which PCs are constructed (in this case the SNP values) are standardised. Variations exist, however, in how the SNPs are selected, how they are weighted in the standardisation step, and how the resultant PCs are selected. These include:

1. Selection of which SNPs to use for PC analysis: It is possible to include all available SNPs, however, it has been suggested that only variants thought to be causally related to the phenotype of interest should be included Vilhjálmsson and Nordborg (2013); Lippert et al. (2013), since these are the ones which lie on the causal pathway between C and Y. The process of selecting SNPs is known as pruning or thinning.

2. The choice of SNP dependent scaling constant before constructing PCs: The intuition behind scaling the SNP value is that sharing a rare variant is greater evidence of common ancestry than sharing a common variant. Scaling values are often estimates of the SNP standard deviation. This may be estimated by the sample standard deviation or using the standard deviation under the Hardy-Weinberg equilibrium model.

   It has also been suggested that, rather than pruning SNPs, SNPs should be weighted according to their degree of LD, to account for replication of causal information by neighbouring, imputed, SNPs in LD Speed et al. (2012). Their proposal uses weights, chosen such that SNPs with high LD are down-weighted. This is implemented in their LDAK software package.

3. The number of PC dimensions chosen for inclusion in the linear model: This is often determined using heuristic measures. Each successive PC accounts for a smaller amount of genetic variation in the chosen SNPs. Most methods use estimates for the proportion of variance explained by each PC, for example selecting PCs to exceed some threshold of the total proportion of variance explained, or else choosing an arbitrary number of PCs.

   In the LMM, it is possible to include all PCs. This is the choice made in the GEMMA software package Zhou and Stephens (2014). This approach is equivalent to measuring the covariance between two individuals based on all chosen SNPs.

Identity By Descent like

Traditional measures for relatedness pre-date modern genomic study, and were originally used to study trait inheritance within pedigrees. Using known pedigree information one can construct the probabilities that genomic regions of two individuals are identical-by-descent (IBD) from a recent common ancestor ('recent' in so far as it is assumed that there is no intermediate mutation or recombination event).

Pedigree based relatedness measures are broadly obsolete in modern genomic analysis for several reasons Speed and Balding (2014): (i) When studying natural populations pedigree information is often unavailable or insufficient to account for population structure. (ii) Even when pedigree information is available, it is usually unrealistic to assume that pedigree founders have zero genetic similarity. (iii) The relatedness of any two individuals tends towards one, as the size of the pedigree is increased.

Rather than using pedigree information to estimate IBD probabilities, modern theories instead measure IBD by appealing to SNP data itself. These methods generally examine the length and frequencies of similar genomic regions in two individuals, and are based on biochemical theories regarding the process by which gametes divide and recombine from two parents. Examples include: FastIBD Browning and Browning (2011), which estimates the frequencies of shared haplotype distributions; and shared segment detection in PLINK Anderson et al. (2010). Reviewing these methods is beyond the scope of this review.

Appendix B

Supplement to partially linear mediation

B.1 Proofs

B.1.1 Proof of Lemma 1

In Step 1 the following expressions for each component of $E\{U(\beta^*, \gamma^*)\}$ are derived for $\beta^* = (\beta_1, \beta_2, \beta_3^*)$. In Step 2 we consider the behaviour of these expressions in each of the two misspecification cases.

$$E\{U_1(\beta^*, \gamma^*)\} = E\left[\{h(Z) - h(Z; \gamma_x^*)\}\{f(Z) - f(Z; \gamma_m^*)\}\right] \tag{B.1}$$
$$E\{U_2(\beta^*, \gamma^*)\} = E\left[\{f(Z) - f(Z; \gamma_m^*)\}\{g(X, Z) - \beta_3^* X - g(Z; \gamma_y^*)\}\right] \tag{B.2}$$
$$E\{U_3(\beta^*, \gamma^*)\} = E\left[\{X - h(Z; \gamma_x^*)\}\{g(X, Z) - \beta_3^* X - g(Z; \gamma_y^*)\}\right] \tag{B.3}$$

Step 1: For the first component we use the partial linearity to obtain

$$\begin{aligned} E\{U_1(\beta^*, \gamma^*)|X, Z\} &= \Big\{X - h(Z; \gamma_x^*)\Big\}\Big\{E(M - \beta_1 X|X, Z) - f(Z; \gamma_m^*)\Big\} \\ &= \Big\{X - h(Z; \gamma_x^*)\Big\}\Big\{f(Z) - f(Z; \gamma_m^*)\Big\} \\ E\{U_1(\beta^*, \gamma^*)|Z\} &= \Big\{h(Z) - h(Z; \gamma_x^*)\Big\}\Big\{f(Z) - f(Z; \gamma_m^*)\Big\} \end{aligned}$$

Similarly for the second component,

$$\begin{aligned} E\{U_2(\beta^*, \gamma^*)|M, X, Z\} &= \Big\{M - \beta_1 X - f(Z; \gamma_m^*)\Big\}\Big\{E(Y - \beta_2 M|M, X, Z) - \beta_3^* X - g(Z; \gamma_y^*)\Big\} \\ E\{U_2(\beta^*, \gamma^*)|X, Z\} &= \Big\{E(M - \beta_1 X|X, Z) - f(Z; \gamma_m^*)\Big\}\Big\{g(X, Z) - \beta_3^* X - g(Z; \gamma_y^*)\Big\} \\ &= \Big\{f(Z) - f(Z; \gamma_m^*)\Big\}\Big\{g(X, Z) - \beta_3^* X - g(Z; \gamma_y^*)\Big\} \end{aligned}$$

Finally for the third component,

$$\begin{aligned} E\{U_3(\beta^*, \gamma^*)|M, X, Z\} &= \Big\{X - h(Z; \gamma_x^*)\Big\}\Big\{E(Y - \beta_2 M|M, X, Z) - \beta_3^* X - g(Z; \gamma_y^*)\Big\} \\ E\{U_3(\beta^*, \gamma^*)|X, Z\} &= \Big\{X - h(Z; \gamma_x^*)\Big\}\Big\{g(X, Z) - \beta_3^* X - g(Z; \gamma_y^*)\Big\} \end{aligned}$$

Step 2: We shall consider the cases (i) and (ii) separately. In case (i) the conditions for assumption A2 are met, hence $f(Z) = f(Z; \gamma_m^*)$ and so (B.1) and (B.2) are exactly zero. The proof for case (i) is completed by letting β_3^* be the value which solves (B.3) equal to zero.

For case (ii) the conditions of A1 are met, hence $h(Z) = h(Z; \gamma_x^*)$ so (B.1) is exactly zero. Also there exists β_3 such that $g(X, Z) = \beta_3 X + g(Z)$ and for $\beta_3^* = \beta_3$ then the conditions in A3 are met so $g(Z) = g(Z; \gamma_y^*)$ and hence (B.2) and (B.3) are exactly zero, which completes the proof for case (ii).

B.1.2 Proof of Lemma 2

In Step 1 the following expressions for each component of $E\{U(\beta^*, \gamma^*)\}$ are derived for $\beta^* = (\beta_1^*, \beta_2, \beta_3)$. In Step 2 we consider the behaviour of these expressions in each of the two misspecification cases.

$$E\{U_1(\beta^*, \gamma^*)\} = E\left[\left\{X - h(Z; \gamma_x^*)\right\}\left\{f(X, Z) - \beta_1^* X - f(Z; \gamma_m^*)\right\}\right] \tag{B.4}$$

$$E\{U_2(\beta^*, \gamma^*)\} = E\left[\left\{f(X, Z) - \beta_1^* X - f(Z; \gamma_m^*)\right\}\left\{g(Z) - g(Z; \gamma_y^*)\right\}\right] \tag{B.5}$$

$$E\{U_3(\beta^*, \gamma^*)\} = E\left[\left\{h(Z) - h(Z; \gamma_x^*)\right\}\left\{g(Z) - g(Z; \gamma_y^*)\right\}\right] \tag{B.6}$$

Step 1. For the first component,

$$E\{U_1(\beta^*, \gamma^*)|X, Z\} = \left\{X - h(Z; \gamma_x^*)\right\}\left\{f(X, Z) - \beta_1^* X - f(Z; \gamma_m^*)\right\}$$

For the second component we use the partial linearity to obtain

$$\begin{aligned} E\{U_2(\beta^*, \gamma^*)|M, X, Z\} &= \left\{M - \beta_1^* X - f(Z; \gamma_m^*)\right\}\left\{E(Y - \beta_2 M - \beta_3 X|M, X, Z) - g(Z; \gamma_y^*)\right\} \\ &= \left\{M - \beta_1^* X - f(Z; \gamma_m^*)\right\}\left\{g(Z) - g(Z; \gamma_y^*)\right\} \\ E\{U_2(\beta^*, \gamma^*)|X, Z\} &= \left\{f(X, Z) - \beta_1^* X - f(Z; \gamma_m^*)\right\}\left\{g(Z) - g(Z; \gamma_y^*)\right\} \end{aligned}$$

Similarly for the third component,

$$\begin{aligned} E\{U_3(\beta^*, \gamma^*)|M, X, Z\} &= \left\{X - h(Z; \gamma_x^*)\right\}\left\{E(Y - \beta_2 M - \beta_3 X|M, X, Z) - g(Z; \gamma_y^*)\right\} \\ &= \left\{X - h(Z; \gamma_x^*)\right\}\left\{g(Z) - g(Z; \gamma_y^*)\right\} \\ E\{U_3(\beta^*, \gamma^*)|Z\} &= \left\{h(Z) - h(Z; \gamma_x^*)\right\}\left\{g(Z) - g(Z; \gamma_y^*)\right\} \end{aligned}$$

Step 2. We shall consider the cases (i) and (ii) separately. In case (i) the conditions for assumption A3 are met, hence $g(Z) = g(Z; \gamma_y^*)$ so (B.5) and (B.6) are exactly zero. Letting β_1^* be the value which solves (B.4) equal to zero completes the proof for case (i).

For case (ii) the conditions of A1 are met, so $h(Z) = h(Z; \gamma_x^*)$ and so (B.6) is zero. Also there exists β_1 such that $f(X, Z) = \beta_1 X + f(Z)$ and for $\beta_1^* = \beta_1$ then the conditions in A2 are met so $f(Z) = f(Z; \gamma_y^*)$ and hence (B.4) and (B.5) are exactly zero, which completes the proof for case (ii).

B.1.3 Proof of Theorem 1

Here we provide a sketch of the proof. Consider the Taylor Expansion

$$\begin{aligned} E_n\{U(\hat{\beta}, \hat{\gamma})\} = E_n\{U(\beta^*, \gamma^*)\} &+ E_n\left\{\frac{\partial U(\beta^*, \gamma^*)}{\partial \beta}\right\}(\hat{\beta} - \beta^*) \\ &+ E_n\left\{\frac{\partial U(\beta^*, \gamma^*)}{\partial \gamma}\right\}(\hat{\gamma} - \gamma^*) + o_p\left(n^{-1/2}\right) \end{aligned}$$

Since $E_n\{U(\hat{\beta}, \hat{\gamma})\} = 0$ then

$$\hat{\beta} - \beta^* = E_n\left\{-\frac{\partial U(\beta^*, \gamma^*)}{\partial \beta}\right\}^{-1}\left[E_n\{U(\beta^*, \gamma^*)\} + E_n\left\{\frac{\partial U(\beta^*, \gamma^*)}{\partial \gamma}\right\}(\hat{\gamma} - \gamma^*)\right] + o_p\left(n^{-1/2}\right)$$

Using the estimator in (3.10) and rearranging gives

$$\hat{\beta} - \beta^* = E_n \left(E_n \left\{ -\frac{\partial U(\beta^*, \gamma^*)}{\partial \beta} \right\}^{-1} \left[U(\beta^*, \gamma^*) + E_n \left\{ \frac{\partial U(\beta^*, \gamma^*)}{\partial \gamma} \right\} \phi(\beta^*, \gamma^*) \right] \right) + o_p \left(n^{-1/2} \right)$$

Applying the weak law of large numbers to the partial derivative terms gives the form of the influence function $\varphi(.)$ in (3.11). We must further show that $E\{\varphi(\beta^*, \gamma^*)\} = 0$

$$E\{\varphi(\beta^*, \gamma^*)\} = E \left\{ -\frac{\partial U(\beta^*, \gamma^*)}{\partial \beta} \right\}^{-1} \left[E\{U(\beta^*, \gamma^*)\} + E \left\{ \frac{\partial U(\beta^*, \gamma^*)}{\partial \gamma} \right\} E\{\phi(\beta^*, \gamma^*)\} \right]$$

Since $\phi(.)$ is an influence function, $E\{\phi(\beta^*, \gamma^*)\} = 0$. Therefore provided $E\{U(\beta^*, \gamma^*)\} = 0$ then $E\{\varphi(\beta^*, \gamma^*)\} = 0$ as required.

B.1.4 Derivation of Equation (3.12)

By Theorem 1,

$$\hat{\beta}_1 = \beta_1 + E_n\{\varphi_1(\beta^*, \gamma^*)\} + o_p \left(n^{-1/2} \right)$$
$$\hat{\beta}_2 = \beta_2 + E_n\{\varphi_2(\beta^*, \gamma^*)\} + o_p \left(n^{-1/2} \right)$$

Therefore, letting $A = E_n\{\varphi_1(\beta^*, \gamma^*)\}$ and $B = E_n\{\varphi_2(\beta^*, \gamma^*)\}$,

$$\hat{\beta}_1 \hat{\beta}_2 - \beta_1 \beta_2 = E_n\{\omega(\beta^*, \gamma^*)\} + AB + o_p \left(n^{-1/2} \right)$$

and the desired result follows provided that $AB = o_p \left(n^{-1/2} \right)$. Using Markov's inequality,

$$P(|n^{1/2} AB| \geq \epsilon) = P(n(AB)^2 \geq \epsilon^2) \leq \frac{nE\{(AB)^2\}}{\epsilon^2}$$

Examining the expectation term, we find a sum over four indices

$$E\{(AB)^2\} = n^{-4} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \sum_{l=1}^{n} E\{\varphi_1^{(i)}(\beta^*, \gamma^*) \varphi_1^{(j)}(\beta^*, \gamma^*) \varphi_2^{(k)}(\beta^*, \gamma^*) \varphi_2^{(l)}(\beta^*, \gamma^*)\}$$

where the superscript (i) denotes that the influence function is evaluated on the ith observation. Since the observations are iid and the influence function has mean zero, the terms of this quadruple sum can only be non-zero when their indices are paired, i.e when ($i = j$ and $k = l$) or ($i = k$ and $j = l$) or ($i = l$ and $j = k$). The number of non-zero terms in the sum is therefore of order n^2, and hence

$$P(|n^{1/2} AB| \geq \epsilon) \leq \mathcal{O} \left(n^{-1} \right)$$

where $\mathcal{O}$ denotes conventional big-O notation, i.e. for sufficiently large n there exists some constant k such that $|\mathcal{O} \left(n^{-1} \right)| \leq kn^{-1}$

B.1.5 Proof of Theorem 2

Here we adapt the proof from Section 5.1 of Dufour et al. (2017) to allow for orthogonal nuisance parameter estimation. We prove the results for the CUE estimator, however they are equally applicable to the two-step estimator. Our extension to the original results relies on three orthogonality-like derivative results for the test statistic of interest. These are derived assuming that the nuisance parameter estimator is orthogonal to the moment conditions in the sense that (3.13) holds. This may either be because all models are correctly specified or because a bias-reduced strategy is used to estimate nuisance parameters,

as described below. We begin by defining the CUE objective function, which we denote by M_n as in the original notation of Dufour et al. (2017),

$$M_n(\beta,\gamma) = D_n^\top(\beta,\gamma) I_n^{-1}(\beta,\gamma) D_n(\beta,\gamma)$$

where, for a target parameter moment function $U(\beta,\gamma)$,

$$\begin{aligned} D_n(\beta,\gamma) &= E_n[U(\beta,\gamma)] \\ C_n(\beta,\gamma) &= E_n\left[\frac{\partial U(\beta,\gamma)}{\partial \gamma}\right] = \frac{\partial D_n(\beta,\gamma)}{\partial \gamma} \\ I_n(\beta,\gamma) &= E_n[U(\beta,\gamma)U(\beta,\gamma)^\top] \end{aligned}$$

Theorem 2 in the text considers the exactly specified setting, i.e. $\text{Dim}(\beta) = \text{Dim}(U(\beta,\gamma))$. For the current proof, however, we consider the over-specified setting, i.e. $\text{Dim}(\beta) \leq \text{Dim}(U(\beta,\gamma))$. Also define the probability limits, β^*, γ^* as the (assumed to be) unique values such that

$$\begin{aligned} \frac{\partial M_n(\beta^*,\gamma^*)}{\partial \beta} &\xrightarrow{p} 0 \\ C_n(\beta^*,\gamma^*) &\xrightarrow{p} 0 \end{aligned}$$

The first of these is equivalent to $D_n(\beta^*,\gamma^*) \xrightarrow{p} 0$ in the exactly specified setting. By the central limit theorem,

$$\begin{aligned} \sqrt{n} D_n(\beta^*,\gamma^*) &\xrightarrow{d} \mathcal{N}(0, I_0) \\ I_n(\beta^*,\gamma^*) &\xrightarrow{p} I_0 = E[U(\beta^*,\gamma^*)U(\beta^*,\gamma^*)^\top] \end{aligned}$$

The unconstrained estimated values $\hat{\beta}, \hat{\gamma}$ are those which solve

$$\begin{aligned} \frac{\partial M_n(\hat{\beta},\hat{\gamma})}{\partial \beta} &= 0 \\ C_n(\hat{\beta},\hat{\gamma}) &= 0 \end{aligned} \tag{B.7}$$

Again, the first of these is equivalent to $D_n(\hat{\beta},\hat{\gamma}) = 0$ in the exactly specified setting. The constrained estimated values $\hat{\beta}_\psi, \hat{\gamma}_\psi$ are those which solve

$$\frac{\partial M_n(\hat{\beta}_\psi,\hat{\gamma}_\psi)}{\partial \beta} - \frac{\partial \psi(\hat{\beta}_\psi)}{\partial \beta}\lambda = 0 \tag{B.8}$$

$$C_n(\hat{\beta}_\psi,\hat{\gamma}_\psi) = 0$$

$$\psi(\hat{\beta}_\psi) = 0 \tag{B.9}$$

for a constraint function ψ and where λ is a Lagrange multiplier. The statement that we intend to prove is that

$$n[M_n(\hat{\beta}_\psi,\hat{\gamma}_\psi) - M_n(\hat{\beta},\hat{\gamma})] \xrightarrow{d} \chi^2_r \tag{B.10}$$

where r is the rank of $\partial\psi(\beta)/\partial\beta$ in a neighbourhood of β^*. In the exactly specified setting, $M_n(\hat{\beta},\hat{\gamma}) = 0$.

Three necessary derivative results

In this subsection we show that, since $C_n(\beta^*, \gamma^*) = o_p(1), D_n(\beta^*, \gamma^*) = o_p(1)$, and $\sqrt{n}D_n(\beta^*, \gamma^*) = O_p(1)$ then

$$\sqrt{n}\frac{\partial M_n(\beta^*, \gamma^*)}{\partial \gamma} = o_p(1) \tag{B.11}$$

$$\frac{\partial^2 M_n(\beta^*, \gamma^*)}{\partial \gamma \partial \beta} = o_p(1) \tag{B.12}$$

$$\frac{\partial^2 M_n(\beta^*, \gamma^*)}{\partial \gamma \partial \gamma} = o_p(1) \tag{B.13}$$

To do so it is easier to work in an index notation where D_i is the ith component of $D_n(\beta, \gamma)$, and I^{-1}_{ij} is the i, jth component of $I_n^{-1}(\beta, \gamma)$ and all quantities are evaluated at $(\beta, \gamma) = (\beta^*, \gamma^*)$. For example, letting $q = \text{Dim}(U(\beta, \gamma))$, then

$$M_n(\beta^*, \gamma^*) = \sum_{i=1}^{q}\sum_{j=1}^{q} D_i I^{-1}_{ij} D_j$$

For the first derivative term of interest,

$$\sqrt{n}\frac{\partial M_n(\beta^*, \gamma^*)}{\partial \gamma} = \sum_{i=1}^{q}\sum_{j=1}^{q}\left\{\left(2\frac{\partial D_i}{\partial \gamma} I^{-1}_{ij} + D_i \frac{\partial I^{-1}_{ij}}{\partial \gamma}\right)\sqrt{n}D_j\right\}$$

and for the second derivative term of interest,

$$\frac{\partial^2 M_n(\beta^*, \gamma^*)}{\partial \gamma \partial \beta} = \sum_{i=1}^{q}\sum_{j=1}^{q}\left\{2\frac{\partial D_i}{\partial \beta} I^{-1}_{ij}\frac{\partial D_j}{\partial \gamma} + \left(2\frac{\partial^2 D_i}{\partial \gamma \partial \beta} I^{-1}_{ij} + 2\frac{\partial D_i}{\partial \beta}\frac{\partial I^{-1}_{ij}}{\partial \gamma} + 2\frac{\partial D_i}{\partial \gamma}\frac{\partial I^{-1}_{ij}}{\partial \beta} + D_i \frac{\partial^2 I^{-1}_{ij}}{\partial \gamma \partial \beta}\right) D_j\right\}$$

For the third,

$$\frac{\partial^2 M_n(\beta^*, \gamma^*)}{\partial \gamma \partial \gamma} = \sum_{i=1}^{q}\sum_{j=1}^{q}\left\{2\frac{\partial D_i}{\partial \gamma} I^{-1}_{ij}\frac{\partial D_j}{\partial \gamma} + \left(2\frac{\partial^2 D_i}{\partial \gamma \partial \gamma} I^{-1}_{ij} + 2\frac{\partial D_i}{\partial \beta}\frac{\partial I^{-1}_{ij}}{\partial \gamma} + 2\frac{\partial D_i}{\partial \gamma}\frac{\partial I^{-1}_{ij}}{\partial \gamma} + D_i \frac{\partial^2 I^{-1}_{ij}}{\partial \gamma \partial \gamma}\right) D_j\right\}$$

By the orthogonality of the nuisance parameter estimator, $\partial D_j/\partial \gamma = o_p(1)$, and since $D_j = o_p(1)$, and $\sqrt{n}D_j = O_p(1)$ then the results follow.

Applying the derivative results

Consider the test statistic

$$\xi = M_n(\hat{\beta}_\psi, \hat{\gamma}_\psi) - M_n(\hat{\beta}, \hat{\gamma})$$

Under standard regularity conditions, and the rank condition in (3.20) (see Dufour et al. (2017) for details), $\hat{\gamma}, \hat{\gamma}_\psi, \hat{\beta}$ and $\hat{\beta}_\psi$ are CAN, hence expanding this test statistics to second order gives

$$\begin{aligned}
n\xi = & \sqrt{n}\frac{\partial M_n(\beta^*,\gamma^*)}{\partial \beta^\top}\left\{\sqrt{n}(\hat{\beta}_\psi-\beta^*)-\sqrt{n}(\hat{\beta}-\beta^*)\right\} \\
& + \sqrt{n}\frac{\partial M_n(\beta^*,\gamma^*)}{\partial \gamma^\top}\left\{\sqrt{n}(\hat{\gamma}_\psi-\gamma^*)-\sqrt{n}(\hat{\gamma}-\gamma^*)\right\} \\
& + \frac{1}{2}\left\{\sqrt{n}(\hat{\beta}_\psi-\beta^*)^\top\frac{\partial^2 M_n(\beta^*,\gamma^*)}{\partial\beta\partial\beta^\top}\sqrt{n}(\hat{\beta}_\psi-\beta^*)^\top-\sqrt{n}(\hat{\beta}-\beta^*)^\top\frac{\partial^2 M_n(\beta^*,\gamma^*)}{\partial\beta\partial\beta^\top}\sqrt{n}(\hat{\beta}-\beta^*)^\top\right\} \\
& + \frac{1}{2}\left\{\sqrt{n}(\hat{\gamma}_\psi-\gamma^*)^\top\frac{\partial^2 M_n(\beta^*,\gamma^*)}{\partial\gamma\partial\gamma^\top}\sqrt{n}(\hat{\gamma}_\psi-\gamma^*)^\top-\sqrt{n}(\hat{\gamma}-\gamma^*)^\top\frac{\partial^2 M_n(\beta^*,\gamma^*)}{\partial\gamma\partial\gamma^\top}\sqrt{n}(\hat{\gamma}-\gamma^*)^\top\right\} \\
& + \left\{\sqrt{n}(\hat{\gamma}_\psi-\gamma^*)^\top\frac{\partial^2 M_n(\beta^*,\gamma^*)}{\partial\gamma\partial\beta^\top}\sqrt{n}(\hat{\beta}_\psi-\beta^*)^\top-\sqrt{n}(\hat{\gamma}-\gamma^*)^\top\frac{\partial^2 M_n(\beta^*,\gamma^*)}{\partial\gamma\partial\beta^\top}\sqrt{n}(\hat{\beta}-\beta^*)^\top\right\} \\
& + o_p(1)
\end{aligned}$$

Using the derivative results (B.11) to (B.13) our expansion reduces to

$$\begin{aligned}
n\xi = & \sqrt{n}\frac{\partial M_n(\beta^*,\gamma^*)}{\partial \beta^\top}\left\{\sqrt{n}(\hat{\beta}_\psi-\beta^*)-\sqrt{n}(\hat{\beta}-\beta^*)\right\} \\
& + \frac{1}{2}\left\{\sqrt{n}(\hat{\beta}_\psi-\beta^*)^\top\frac{\partial^2 M_n(\beta^*,\gamma^*)}{\partial\beta\partial\beta^\top}\sqrt{n}(\hat{\beta}_\psi-\beta^*)^\top-\sqrt{n}(\hat{\beta}-\beta^*)^\top\frac{\partial^2 M_n(\beta^*,\gamma^*)}{\partial\beta\partial\beta^\top}\sqrt{n}(\hat{\beta}-\beta^*)^\top\right\} \\
& + o_p(1)
\end{aligned}$$

Next we consider the first order Taylor expansions of the estimating equations (B.7) to (B.9), taken about the probability limit values. Again, since $\hat{\gamma}, \hat{\gamma}_\psi, \hat{\beta}$ and $\hat{\beta}_\psi$ are CAN,

$$\begin{aligned}
0 &= \frac{\partial M_n(\beta^*,\gamma^*)}{\partial\beta}+\frac{\partial^2 M_n(\beta^*,\gamma^*)}{\partial\beta\partial\beta}(\hat{\beta}-\beta^*)+\frac{\partial^2 M_n(\beta^*,\gamma^*)}{\partial\gamma\partial\beta}(\hat{\gamma}-\gamma^*)+o_p(n^{-1/2}) \\
0 &= \frac{\partial M_n(\beta^*,\gamma^*)}{\partial\beta}+\frac{\partial^2 M_n(\beta^*,\gamma^*)}{\partial\beta\partial\beta}(\hat{\beta}_\psi-\beta^*)+\frac{\partial^2 M_n(\beta^*,\gamma^*)}{\partial\gamma\partial\beta}(\hat{\gamma}_\psi-\gamma^*)-\frac{\partial\psi(\hat{\beta}_\psi)}{\partial\beta}\lambda+o_p(n^{-1/2}) \\
0 &= \psi(\beta^*)+\frac{\partial\psi(\beta^*)}{\partial\beta}(\hat{\beta}_\psi-\beta^*)+o_p(n^{-1/2})
\end{aligned}$$

Under the null, $\psi(\beta^*)=0$, application of (B.12) gives

$$\begin{aligned}
0 &= X_n+V_0(\hat{\beta}-\beta^*)+o_p(n^{-1/2}) \\
0 &= X_n+V_0(\hat{\beta}_\psi-\beta^*)-P_0\lambda+o_p(n^{-1/2}) \\
0 &= P_0(\hat{\beta}_\psi-\beta^*)+o_p(n^{-1/2})
\end{aligned}$$

where,

$$\begin{aligned}
\frac{\partial\psi(\beta^*)}{\partial\beta} &= P_0 \\
\frac{\partial M_n(\beta^*,\gamma^*)}{\partial\beta} &= X_n \\
\frac{\partial^2 M_n(\beta^*,\gamma^*)}{\partial\beta\partial\beta} &\xrightarrow{p} V_0
\end{aligned}$$

It follows immediately from the original proof in Dufour et al. (2017) that

$$\xi=\frac{1}{2}X_n^\top V_0^{-1}P_0^\top(P_0V_0^{-1}P_0^\top)^{-1}P_0V_0^{-1}X_n+o_p(n^{-1})$$

The final result follows when $\sqrt{n}X_n \xrightarrow{d} \mathcal{N}(0,2V_0)$. This can be shown using the same derivative methods as used to show (B.11) to (B.13).

B.1.6 Proof of Equations (3.21) and (3.22)

We will prove (3.21) with the result for (3.22) proceeding in a similar fashion.

Consider Theorem 2 under the null hypothesis $\psi^{(0)}(\beta^*) = \beta_1\beta_2 = 0$. With the null parameter space given by $B_0 = \{\beta|\psi^{(0)}(\beta) = 0\}$, with

$$\text{Rank}\left(\frac{\partial\psi^{(0)}(\beta)}{\partial\beta}\right) = \begin{cases} 0 & \text{for } \beta_1 = \beta_2 = 0 \\ 1 & \text{otherwise} \end{cases}$$

One may decompose the supremum in (3.21) as

$$\sup_{\beta_* \in B_0} P_{\beta^*}\left(S_0 > x\right) = \max\left\{\sup_{\beta^* \in B_0 \setminus A} P_{\beta^*}\left(S_0 > x\right), \sup_{\beta^* \in A} P_{\beta^*}\left(S_0 > x\right)\right\} \tag{B.14}$$

where $A = \{\beta|\beta_1 = \beta_2 = 0\}$. For the first term in the max bracket above, the rank condition of Theorem 2 holds, so for all $\beta^* \in B_0 \setminus A$

$$P_{\beta^*}\left(S_0 > x\right) \to 1 - F_{\chi^2_1}(x)$$

Considering the second term, one may decompose the test statistic as

$$P_{\beta^*}\left(S_0 > x\right) = P_{\beta^*}\left(S_1 > x, S_2 > x\right) \leq P_{\beta^*}\left(S_1 > x\right)$$

where (for $j = 1, 2$) $S_j = \min_{\beta \in C_j} S(\beta)$ and $C_j = \{\beta|\beta_j = 0\}$. By Theorem 2, for all β^* in A,

$$P_{\beta^*}\left(S_1 > x\right) \to 1 - F_{\chi^2_1}(x)$$

Hence $P_{\beta^*}\left(S_0 > x\right)$ is asymptotically bounded from above by $1 - F_{\chi^2_1}(x)$ for all β^* in B_0, so (3.21) holds.

B.1.7 G-estimation when outcome model has exposure-mediator interaction

In the following we reason about the NIDE obtained by G-estimation using moment conditions (3.6) – (3.8), when one has erroneously excluded an interaction term from the outcome model, but the mediator model, $E(M|X, Z)$ is correctly specified and partially linear, i.e. in truth, (3.1) and (3.24) both hold. We define the probability limit as $\beta^* = (\beta_1^*, \beta_2^*, \beta_3^*)$ which solves $E\{U(\beta^*, \gamma^*)\} = 0$ and use assumption A2 as before. Let,

$$\begin{aligned}
\delta_j &= \beta_j - \beta_j^* \\
\epsilon_x &= X - E(X|Z) \\
\epsilon_m &= M - E(M|X, Z) \\
\beta_3 &= \frac{E(\epsilon_x g(X, Z))}{E(\epsilon_x X)} \\
\Delta_f &= f(Z) - f(Z; \gamma_m^*) \\
\Delta_g &= g(X, Z) - \beta_3 X - g(Z; \gamma_m^*) \\
\Delta_h &= h(Z) - h(Z; \gamma_m^*)
\end{aligned}$$

for $j = 1, 2, 3$. Here β_3 is the least squares coefficient of a regression of $g(X, Z)$ on X. It follows that the expected moment conditions can be written

$$\begin{aligned}
E[U_1(\beta^*, \gamma^*)] &= E\{(\epsilon_x + \Delta_h)(\delta_1 X + \Delta_f)\} \\
E[U_2(\beta^*, \gamma^*)] &= E\{(\epsilon_m + \delta_1 X + \Delta_f)(\delta_2 M + \theta MX + \delta_3 X + \Delta_g)\} \\
E[U_3(\beta^*, \gamma^*)] &= E\{(\epsilon_x + \Delta_h)(\delta_2 M + \theta MX + \delta_3 X + \Delta_g)\}
\end{aligned}$$

For the first equation, since $E(\epsilon_x \Delta_f) = 0$, then

$$E[U_1(\beta^*, \gamma^*)] = \delta_1 E[(\epsilon_x + \Delta_h)X] + E(\Delta_h \Delta_f)$$

We assume that $f(z)$ is modelled correctly and when $\delta_1 = 0$ then assumption A2 is satisfied and $\Delta_f = 0$. Using this fact, the second equation becomes

$$E[U_2(\beta^*, \gamma^*)] = \delta_2 E(\epsilon_m M) + \theta E(\epsilon_m MX)$$

where we have used the fact that $E(\epsilon_m X) = E(\epsilon_m \Delta_g) = 0$. Hence,

$$\delta_2 = -\theta \frac{E(\epsilon_m MX)}{E(\epsilon_m M)}$$

which, since $\delta_1 = 0$, gives the result

$$\beta_1^* \beta_2^* = \beta_1 \left(\beta_2 + \theta \bar{x}\right)$$

where

$$\bar{x} = \frac{E(\epsilon_m MX)}{E(\epsilon_m M)} = \frac{E[X \mathrm{var}(M|X, Z)]}{E[\mathrm{var}(M|X, Z)]} \tag{B.15}$$

can be thought of as a weighted average of X, or as a population least squares regression coefficient from regressing MX on M.

Appendix C

Supplement to influence curve based inference

C.1 Riesz Representation Theorem

Suppose P and $\widetilde{P}$ are both absolutely continuous w.r.t. some measure ν and denote the density functions $f(o) = dP(o)/d\nu(o)$ and $\tilde{f}(o) = d\widetilde{P}(o)/d\nu(o)$. The density of P_t w.r.t. ν is also well defined

$$f_t(o) \quad = \quad \frac{dP_t(o)}{d\nu(o)} = f(o) + t\left\{\tilde{f}(o) - f(o)\right\}.$$

The score function $S_t(o)$ is the derivative of the log density w.r.t. t

$$S_t(o) \quad = \quad \frac{d\log\{f_t(o)\}}{dt} = \frac{\tilde{f}(o) - f(o)}{f_t(o)}.$$

It follows that

$$\begin{aligned} S_t(o)dP_t(o) &= S_t(o)f_t(o)d\nu(o) \\ &= d\widetilde{P}(o) - dP(o). \end{aligned}$$

Hence $P_t\{S_t(O)\} = 0$. Now we consider the L_2 Hilbert space defined using the measure P_t. This is the set of functions $h(O)$ such that $P_t\{h(O)\} = 0, P_t\{h(O)^2\} < \infty$ and, letting $g(O)$ be another member of this space we define the inner product $P_t\{h(O)g(O)\}$. We refer the interested reader to Levy (2019) for an introduction to these Hilbert spaces. Now, assuming that $d\Psi(P_t)/dt$ is a continuous linear functional of $S_t(O)$, which is assumed to be a member of the Hilbert space, we use the Riesz Representation Theorem to obtain

$$\begin{aligned} \frac{d\Psi(P_t)}{dt} &= P_t\left\{\phi(O, P_t)S_t(O)\right\} \\ &= \int \phi(o, P_t)S_t(o)dP_t(o) \\ &= \int \phi(o, P_t)\{d\widetilde{P}(o) - dP(o)\} \\ &= (\widetilde{P} - P)\left\{\phi(O, P_t)\right\} \end{aligned}$$

It follows that this expansion holds for all t. Also note that $P_t\left\{\phi(O, P_t)\right\} = 0$. In the special cases $t = 0$ and $t = 1$ this allows us to write

$$\begin{aligned} \frac{d\Psi(P_t)}{dt}\Big|_{t=0} &= \widetilde{P}\left\{\phi(O, P)\right\} \\ \frac{d\Psi(P_t)}{dt}\Big|_{t=1} &= -P\left\{\phi(O, \widetilde{P})\right\}. \end{aligned}$$

C.2 Additional results

In this Appendix we derive the following results, which readers might find helpful for reference. Here, $F^{-1}(\tau)$ is the quantile function of Y for known $\tau \in [0,1]$, and $F(y|x)$ is the cumulative distribution function of Y given $X = x$. Also $\Theta(u)$ is a step function which takes the value 1 when $u \geq 0$ and 0 otherwise.

$$\partial_t F_t(y|x) = \frac{\mathbb{1}_{\tilde{x}}(x)}{f(x)} \{\Theta(y - \tilde{y}) - F(y|\tilde{x})\}$$
$$\partial_t E_{P_t}(Y|Y \leq y) = \frac{\Theta(y - \tilde{y})}{F(y)} \{\tilde{y} - E_P(Y|Y \leq y)\}$$
$$\partial_t F_t^{-1}(\tau) = \frac{\Theta\left\{\tilde{y} - F^{-1}(\tau)\right\} + \tau - 1}{f\left\{F^{-1}(\tau)\right\}}.$$

We also illustrate the steps described in the main paper through two further examples: the interventional direct effect, and the incremental propensity score intervention.

Conditional cumulative distribution function. Here we consider the conditional cumulative distribution function, $F(y|x)$ where, y and x are known,

$$F(y|x) = E_{\mathcal{P}} \{\Theta(Y - y)|X = x\}$$

It is fairly straightforward to recycle the result in (4.7), with Y replaced with $\Theta(Y - y)$ to recover the desired form.

Tail conditional expectation. Here we consider the tail conditional expectation, $E_P(Y|Y \leq y)$, where y is known:

$$E_P(Y|Y \leq y) = \frac{E_P\{\Theta(y - Y)Y\}}{F(y)}.$$

Now perturbing in the direction of the parametric submodel, and applying the quotient rule,

$$\begin{aligned}
\partial_t E_{P_t}(Y|Y \leq y) &= \frac{\partial_t E_{P_t}\{\Theta(y - Y)Y\}F(y) - E_P\{\Theta(y - Y)Y\}\partial_t F_t(y)}{F(y)^2} \\
&= \frac{[\Theta(y - \tilde{y})\tilde{y} - E_P\{\Theta(y - Y)Y\}]F(y) - E_P\{\Theta(y - Y)Y\}\{\Theta(y - \tilde{y}) - F(y)\}}{F(y)^2} \\
&= \frac{\Theta(y - \tilde{y})}{F(y)}\{\tilde{y} - E_P(Y|Y \leq y)\}.
\end{aligned}$$

We notice that the resultant efficient influence function is zero for observations where $\tilde{y} > y$. This coheres with our intuition that the distribution of Y outside the region $Y \leq y$ does not contribute to the asymptotic efficiency bound of $E(Y|Y \leq y)$.

Quantile function. Here we consider the quantile function, $F_t^{-1}(\tau)$, of a continuous random variable Y, where $\tau \in [0,1]$ is known. An alternative derivation of the influence curve can be found in van der Vaart (1998b). We define the estimand $\Psi = \Psi_\tau(P) = F^{-1}(\tau)$. The distribution quantile is implicitly defined by

$$\int_a^{\Psi_\tau(P)} f(y)dy = \tau,$$

where a denotes the lower boundary of the support of Y and $f(y)$ is the density function of Y. Under the parametric submodel,

$$\int_a^{\Psi_\tau(P_t)} f_t(y)dy = \tau.$$

Differentiating both sides with respect to t, the Leibniz integral rule gives us that

$$f_t\left\{\Psi_\tau(\mathcal{P}_t)\right\}\frac{d\Psi_\tau(\mathcal{P}_t)}{dt}+\int_a^{\Psi_\tau(\mathcal{P}_t)}\frac{df_t(y)}{dt}dy=0.$$

Hence,

$$\begin{aligned}\partial_t\Psi_\tau(P_t)&=\frac{-1}{f\left\{\Psi(P)\right\}}\int_a^{\Psi(P)}\left\{\mathbb{1}_{\tilde{y}}(y)-f(y)\right\}dy\\&=\frac{1}{f\left\{\Psi(P)\right\}}\left\{\int_a^{\Psi(P)}f(y)dy-\int_a^{\Psi(P)}\mathbb{1}_{\tilde{y}}dy\right\}\\&=\frac{\tau-\left[1-\Theta\left\{\tilde{y}-\Psi(P)\right\}\right]}{f\left\{\Psi(P)\right\}}.\end{aligned}$$

The resulting efficient influence function can be rewritten by defining the function $\rho'_\tau(u)=\Theta(u)+\tau-1$, which is the derivative (almost everywhere) of the standard quantile regression loss function, $\rho_\tau(u)=u[\Theta(u)+\tau-1]$. Doing so results in

$$\phi(y,P)=\rho'_\tau\left\{y-\Psi(P)\right\}/f\left\{\Psi(P)\right\}.$$

Interestingly, and as an aside, one might wonder how this estimand behaves for different distributions. Let's consider the median when Y follows a univariate normal distribution with mean μ and standard deviation σ. For the normal distribution the mean is equal to the median, so $\Psi(P)=\mu$. And hence

$$\begin{aligned}\phi(y,P)&=\sigma\sqrt{2\pi}\rho'_{1/2}(y-\mu)\\\phi(y,P)^2&=\frac{\pi}{2}\sigma^2.\end{aligned}$$

The standard error in the median estimator is therefore $E\left\{\phi(Y,P)^2/n\right\}^{1/2}\approx 1.253\frac{\sigma}{\sqrt{n}}$. This is 25% larger than the standard error in the sample mean, which (under the assumption of normality) estimates the same quantity, but achieves the Cramer-Rao lower bound.

Example 9 (interventional direct effect). In this example we will derive (one half of) the efficient influence function for the interventional direct effect for mediation, first defined by Vansteelandt and Daniel (2017), with an efficient influence function given in Benkeser (2020). This estimand is derived using a causal framework and is used to evaluate the effect of a binary outcome, X, on an outcome, Y, through a set of mediating variables, M, given a set of confounder variables, Z. Under standard causal assumptions the estimand may be written as a functional of the observed data. We shall not detail these assumptions here, since, once a functional of the data generating distribution is obtained, the causal assumptions are no longer required to derive estimators and efficiency results for it. For our purposes, it is sufficient to define the estimand over the set of variables, $O=(Y,M,X,Z)$, with conditional response surface, $b(m,x,z)=E(Y|M=m,X=x,Z=z)$,

$$\Psi(P)=\int b(m,x^1,z)f(m|x^0,z)f(z)dmdz, \tag{C.1}$$

where x^1 and x^0 are known values. Under the parametric submodel,

$$\Psi(P_t)=\int b_t(m,x^1,z)f_t(m|x^0,z)f_t(z)dmdz. \tag{C.2}$$

Applying the derivative operator gives

$$\begin{aligned}\partial_t \Psi(P_t) = \int \Bigg[& \partial_t b_t(m, x^1, z) f(m|x^0, z) f(z) \\ & + b(m, x^1, z) \partial_t f_t(m|x^0, z) f(z) \\ & + b(m, x^1, z) f(m|x^0, z) \partial_t f_t(z) \Bigg] dm dz.\end{aligned}$$

Evaluating these derivatives gives

$$\begin{aligned}\partial_t \Psi(P_t) = \int \Bigg[& \frac{\mathbb{1}_{\tilde{o}}(m, x^1, z)}{f(m, x^1, z)} \left\{ \tilde{y} - b(m, x^1, z) \right\} f(m|x^0, z) f(z) \\ & + b(m, x^1, z) \frac{\mathbb{1}_{\tilde{o}}(x^0, z)}{f(x^0, z)} \left\{ \mathbb{1}_{\tilde{m}}(m) - f(m|x^0, z) \right\} f(z) \\ & + b(m, x^1, z) f(m|x^0, z) \left\{ \mathbb{1}_{\tilde{z}}(z) - f(z) \right\} \Bigg] dm dz\end{aligned}$$

and evaluating the integral results in the efficient influence function

$$\frac{\mathbb{1}_{x^1}(X) f(M|x^0, Z)}{f(M, x^1|Z)} \left\{ Y - b(M, x^1, Z) \right\} + \frac{\mathbb{1}_{x^0}(X)}{f(x^0|Z)} \{ b(M, x^1, Z) - a(x^1, x^0, Z) \} + a(x^1, x^0, Z) - \Psi(\mathcal{P}),$$

where we define

$$a(x^1, x^0, z) = \int b(m, x^1, z) f(m|x^0, z) dm.$$

Example 10 (incremental propensity score intervention). The incremental propensity score intervention estimand is motivated by, and derived in the work of Kennedy (2019). It is an interesting example, since it uses a stochastic intervention which is a function of the true data generating distribution. We define the estimand over the set of variables $O = (Y, X, Z)$, where X is binary with propensity score $\pi(z) = E_P(X|Z = z)$, and conditional response surface, $m(x, z) = E(Y|X = x, Z = z)$,

$$\Psi(P) = \sum_{x=0}^{1} \int m(x, z) g_P(x|z) f(z) dz,$$

where $g_P(x|z)$ is a probability mass function, which is dependent on the true data generating distribution. Kennedy (2019) propose the 'propensity score intervention' indexed by a known value ϵ,

$$g_P(x|z) = \frac{x\epsilon\pi(z) + (1 - x) \{1 - \pi(z)\}}{\epsilon\pi(z) + 1 - \pi(z)}.$$

This propensity score intervention is motivated by a multiplication on the odds ratio scale,

$$\frac{g_P(1|z)}{g_P(0|z)} = \epsilon \frac{\pi(z)}{1 - \pi(z)}$$

although for the purposes of influence function derivation, we are not too concerned with interpretation of the estimand. Under the parametric submodel,

$$\Psi(P_t) = \sum_{x=0}^{1} \int m_t(x, z) g_{P_t}(x|z) f_t(z) dz.$$

Applying the ∂_t operator gives

$$\partial_t \Psi(P_t) = \sum_{x=0}^{1} \int \left[\frac{\mathbb{1}_{\tilde{o}}(x,z)}{f(x,z)} \{\tilde{y} - m(x,z)\} g_P(x|z) f(z) + m(x,z) \frac{dg_P(x|z)}{d\pi} \frac{\mathbb{1}_{\tilde{z}}(z)}{f(z)} \{\tilde{x} - \pi(z)\} f(z) \right.$$
$$\left. + m(x,z) g_P(x|z) \{\mathbb{1}_{\tilde{z}}(z) - f(z)\} \right] dz,$$

where

$$\frac{dg_P(x|z)}{d\pi} = \frac{(2x-1)\epsilon}{(\epsilon\pi(z) + 1 - \pi(z))^2}$$
$$= \frac{g_P(1|z) g_P(0|z)}{\pi(z)(1-\pi(z))} \{\mathbb{1}_1(x) - \mathbb{1}_0(x)\}.$$

Now, integrating over z, becomes

$$\partial_t \Psi(P_t) = \sum_{x=0}^{1} \left[\frac{\mathbb{1}_{\tilde{x}}(x)}{f(x|\tilde{z})} \{\tilde{y} - m(\tilde{x},\tilde{z})\} g_P(x|\tilde{z}) \right.$$
$$\left. + m(x,\tilde{z}) \frac{g_P(1|\tilde{z}) g_P(0|\tilde{z})}{\pi(\tilde{z}) \{1-\pi(\tilde{z})\}} \{\mathbb{1}_1(x) - \mathbb{1}_0(x)\} \{\tilde{x} - \pi(\tilde{z})\} + m(x,\tilde{z}) g_P(x|\tilde{z}) \right] - \Psi(P).$$

Performing the summation over x, the efficient influence function becomes

$$g_P(1|Z)\varphi_1(O,P) + g_P(0|Z)\varphi_0(O,P) + \frac{g_P(1|Z) g_P(0|Z)}{\pi(Z)\{1-\pi(Z)\}} \{X - \pi(Z)\} \{m(1,Z) - m(0,Z)\} - \Psi(P),$$

where $\varphi_x(O,P)$ is the 'uncentered' AIPW influence function as in (4.6).

Appendix D

Supplement to variable importance estimands

D.1 Derivation of Efficient Influence Curve

To derive the ICs in (5.3) and (5.5) we adopt the formalism given in Hines et al. (2022). Specifically we let P denote the true distribution of (Y, A, X) and let $\tilde{P}$ denote a point mass at $(\tilde{y}, \tilde{a}, \tilde{x})$. We further denote the parametric submodel $P_t = t\tilde{P} + (1-t)P$ where $t \in [0,1]$ is a scalar parameter, and we let ∂_t denote an operator such that for some function of $f(t)$, $\partial_t f(t) \equiv \frac{df(t)}{dt}|_{t=0}$.

We make use of the following lemma, which we demonstrate later in the proof. Letting $g_P(X)$ denote some functional of P, then

$$\partial_t E_{P_t}\{g_{P_t}(X)|X_{-s} = x_{-s}\} = \frac{\tilde{f}(x_{-s})}{f(x_{-s})}\left[g_P(\tilde{x}) - E_P\{g_P(X)|X_{-s} = x_{-s}\}\right] + E_P\{\partial_t g_{P_t}(X)|X_{-s} = x_{-s}\} \tag{D.1}$$

where $\tilde{f}(.)$ and $f(.)$ denote the marginal 'densities' of X_{-s} under $\tilde{P}$ and P respectively, which are both assumed to be absolutely continuous w.r.t. to a dominating measure. In practice this expression means that for discrete X_{-s} then $f(.)$ is a probability mass function and $\tilde{f}(.)$ is an indicator function. Similarly for continuous X_{-s} then $f(.)$ is a probability density function and $\tilde{f}(.)$ is a dirac delta function. In both cases $\frac{\tilde{f}(x_{-s})}{f(x_{-s})}$ is a probability point mass, which is zero when $\tilde{x}_{-s} \neq x_{-s}$.

It follows immediately from (D.1) that,

$$\partial_t E_{P_t}\{g_{P_t}(X)\} = g_P(\tilde{x}) - E_P\{g_P(X)\} + E_P\{\partial_t g_{P_t}(X)\} \tag{D.2}$$

By considering that

$$\text{var}_P\{g_P(X)|X_{-s} = x_{-s}\} = E_P\{g_P^2(X)|X_{-s} = x_{-s}\} - E_P\{g_P(X)|X_{-s} = x_{-s}\}^2$$

We can use (D.1) to show that that

$$\partial_t \text{var}_{P_t}\{g_{P_t}(X)|X_{-s} = x_{-s}\} = \frac{\tilde{f}(x_{-s})}{f(x_{-s})}\left[\{g_P(\tilde{x}) - E_P(g_P(X)|X_{-s} = x_{-s})\}^2 - \text{var}_P\{g_P(X)|X_{-s} = x_{-s}\}\right] \tag{D.3}$$

$$+ 2\text{cov}_P\{g_P(X), \partial_t g_{P_t}(X)|X_{-s} = x_{-s}\}$$

where $\text{cov}(A, B|C) \equiv E(\{A - E(A|C)\}B|C)$ denotes the conditional covariance. Using the results in (D.2) and (D.3), we obtain

$$\begin{aligned}\partial_t E_{P_t}\left[\text{var}_{P_t}\{g_{P_t}(X)|X_{-s}\}\right] = \{g_P(\tilde{x}) - E_P(g_P(X)|X_{-s} = \tilde{x}_{-s})\}^2 - E_P\left[\text{var}_P\{g_P(X)|X_{-s}\}\right] \\ + 2E_P\{\text{cov}_P\{g_P(X), \partial_t g_{P_t}(X)|X_{-s}\}\}\end{aligned} \tag{D.4}$$

Setting $g_P(X) = \tau(X)$, we use (D.1) and the fact that $\tau(x) = \mu(1,x) - \mu(0,x)$ to show that,

$$\partial_t g_{P_t}(x) = \frac{\tilde{f}(x)}{f(x)}\{\tilde{y} - \mu(\tilde{a},x)\}\frac{\tilde{a} - \pi(x)}{\pi(x)\{1-\pi(x)\}}$$

Hence, (D.4) implies the IC,

$$\begin{aligned}\phi_s(\tilde{z}) &= \{\tau(\tilde{x}) - \tau_s(\tilde{x})\}^2 - \Theta_s + 2\{\tau(\tilde{x}) - \tau_s(\tilde{x})\}\{\tilde{y} - \mu(\tilde{a},\tilde{x})\}\frac{\tilde{a} - \pi(\tilde{x})}{\pi(\tilde{x})\{1-\pi(\tilde{x})\}} \\ &= \{\tau(\tilde{x}) - \tau_s(\tilde{x})\}^2 - \Theta_s + 2\{\tau(\tilde{x}) - \tau_s(\tilde{x})\}\{\varphi(\tilde{z}) - \tau(\tilde{x})\}\end{aligned}$$

Completing the square of the expression above gives the result in (5.3). In replicating this proof, it is useful to note that for an arbitrary function $h(x)$

$$E_P\left\{\frac{\tilde{f}(X)}{f(X)}h(X)\right\} = h(\tilde{x})$$

Proof of Lemma in (D.1)

To demonstrate (D.1) we write the lefthand side as

$$\partial_t \int g_{P_t}(x^*)dP_{t,X_s|x_{-s}}(x_s^*) = \int g_P(x^*)\partial_t dP_{t,X_s|x_{-s}}(x_s^*) + \int\{\partial_t g_{P_t}(x^*)\}dP_{X_s|x_{-s}}(x_s^*)$$

where $dP_{t,X_s|x_{-s}}(.)$ is the conditional distribution of X_s given $X_{-s} = x_{-s}$ under the parametric submodel and $x^*_{-s} = x_{-s}$. The second integral on the righthand side recovers the final term in (D.1). Hence the lemma follows once we show that

$$\partial_t dP_{t,X_s|x_{-s}}(x_s^*) = \frac{\tilde{f}(x_{-s})}{f(x_{-s})}\{d\tilde{P}_{X_s}(x_s^*) - dP_{X_s|x_{-s}}(x_s^*)\}$$

To do so, let μ denote a dominating measure and write

$$\begin{aligned}dP_{t,X_s|x_{-s}}(x_s^*) &= f_{t,X_s|x_{-s}}(x_s^*)d\mu(x_s^*) \\ &= \frac{f_{t,X}(x^*)}{f_{t,X_{-s}}(x_{-s})}d\mu(x_s^*)\end{aligned}$$

where $f_{t,X}(.)$ and $f_{t,X_{-s}}(.)$ denote the marginal densities of X and X_{-s} under the parametric submodel, P_t, i.e. they are the Radon-Nikodym derivatives w.r.t. μ. Applying the quotient rule, we obtain

$$\partial_t dP_{t,X_s|x_{-s}}(x_s^*) = \frac{1}{f_{X_{-s}}(x_{-s})}\left[\partial_t f_{t,X}(x^*) - \frac{f_X(x^*)}{f_{X_{-s}}(x_{-s})}\partial_t f_{t,X_{-s}}(x_{-s})\right]d\mu(x_s^*)$$

We now evaluate the derivative parts. Since $\partial_t P_t = \tilde{P} - P$, the marginal density derivatives will have a similar structure, as shown in the first expression below, where $f_X(.)$ and $\tilde{f}_X(.)$ denote marginal densities of X under $\tilde{P}$ and P, with likewise for X_{-s}

$$\partial_t dP_{t,X_s|x_{-s}}(x^*) = \frac{1}{f_{X_{-s}}(x_{-s})}\left[\{\tilde{f}_X(x^*) - f_X(x^*)\} - \frac{f_X(x^*)}{f_{X_{-s}}(x_{-s})}\{\tilde{f}_{X_{-s}}(x_{-s}) - f_{X_{-s}}(x_{-s})\}\right]d\mu(x_s^*)$$

Since $\tilde{P}$ is a point mass, $\tilde{f}_X(x^*) = \tilde{f}_{X_s}(x_s^*)\tilde{f}_{X_{-s}}(x^*_{-s})$. Also $x^*_{-s} = x_{-s}$ hence,

$$\begin{aligned}\partial_t dP_{t,X_s|x_{-s}}(x^*) &= \frac{\tilde{f}_{X_{-s}}(x_{-s})}{f_{X_{-s}}(x_{-s})}\left[\tilde{f}_{X_s}(x_s^*) - \frac{f_X(x^*)}{f_{X_{-s}}(x_{-s})}\right]d\mu(x_s^*) \\ &= \frac{\tilde{f}_{X_{-s}}(x_{-s})}{f_{X_{-s}}(x_{-s})}\left[\tilde{f}_{X_s}(x_s^*) - f_{X_s|x_{-s}}(x_s^*)\right]d\mu(x_s^*)\end{aligned}$$

Thus, the result follows.

Corollary

An immediate consequence of these IC derivations is that the IC of $\text{var}\{\tau_s(X)\} = \Theta_p - \Theta_s$ is,

$$\phi_p(Z) - \phi_s(Z) = \{\varphi(Z) - \tau_p\}^2 - \{\varphi(Z) - \tau_s(X)\}^2 - \text{var}\{\tau_s(X)\}$$

This result is interesting since it holds even when Y is not independent of A given X_{-s}.

D.2 Estimator Asymptotic Distributions

D.2.1 Proof of Theorem 3

We demonstrate asymptotic regularity for the estimator $\hat{\Theta}_s$, with the result for $\hat{\Theta}_p$ following from the case $s = p$. Asymptotic regularity of Ψ_s follows using the ratio argument above.

Throughout we use superscript hat to denote functional estimators obtained from an independent sample, and we define,

$$\hat{\varphi}(z) = \{y - \hat{\mu}(a,x)\}\frac{a - \hat{\pi}(x)}{\hat{\pi}(x)\{1 - \hat{\pi}(x)\}} + \hat{\mu}(1,x) - \hat{\mu}(0,x)$$
$$\hat{\phi}_s(z) = \{\hat{\varphi}(z) - \hat{\tau}_s(x)\}^2 - \{\hat{\varphi}(z) - \hat{\tau}(x)\}^2 - \hat{\Theta}_s^0$$

where $\hat{\Theta}_s^0$ is an initial plug-in estimate of Θ_s. We make the following assumptions about these functional estimators,

(A1) The propensity score and outcome estimators are 'double robust' in the sense that $\{\pi(x) - \hat{\pi}(x)\}\{\mu(a,x) - \hat{\mu}(a,x)\}$ is $o_P(n^{-1/2})$ in $L_2(P)$ norm for $a = 0, 1$.

(A2) The differences $\tau(x) - \hat{\tau}(x)$ and $\tau_s(x) - \hat{\tau}_s(x)$ are both $o_P(n^{-1/4})$ in $L_2(P)$ norm.

(A3) The CATE difference estimates are bounded as $\{\hat{\tau}(x) - \hat{\tau}_s(x)\}^2 \leq \delta$ for some $\delta < \infty$ with probability 1.

(A4) The propensity score estimates are bounded as $\epsilon \leq \hat{\pi}(x) \leq 1 - \epsilon$ for some $\epsilon > 0$ with probability 1.

(A5) There exists a P-Donsker class $\mathcal{G}_0$ such that $P(\hat{\phi}_s(.) \in \mathcal{G}_0) \to 1$.

(A6) There exists a constant $K > 0$ such that each of $\tau(x)$, $\hat{\tau}(x)$, $\hat{\tau}_s(x)$ and $\text{var}(\varphi(Z)|X = x)$ has range uniformly contained in $(-K, K)$ with probability one as $n \to \infty$.

(A7) There exists a constant $K > 0$ such that $\text{var}(Y|X = x)$ and $\hat{\mu}(a,x)$ have range uniformly contained in $(-K, K)$ with probability one as $n \to \infty$.

Under these assumptions we show that the remainder term, R, in the expansion below is $o_P(n^{-1/2})$

$$\hat{\Theta}_s^0 - \Theta_s = -E\{\hat{\phi}_s(Z)\} + R$$

where we highlight that the expectation is conditional on the functional estimators, i.e. $\hat{\varphi}(z)$ is treated as a fixed function. We then show that

$$-E\{\hat{\phi}_s(Z)\} = n^{-1}\sum_{i=1}^{n}\phi_s(z_i) + H_n - n^{-1}\sum_{i=1}^{n}\hat{\phi}_s(z_i) \tag{D.5}$$

where H_n is an empirical process term, which is $o_P(n^{-1/2})$ under our assumptions. It follows therefore that for

$$\hat{\Theta}_s = \hat{\Theta}_s^0 + n^{-1}\sum_{i=1}^{n}\hat{\phi}_s(z_i)$$
$$\hat{\Theta}_s - \Theta_s = n^{-1}\sum_{i=1}^{n}\phi_s(z_i) + o_P(n^{-1/2}).$$

Formally $\hat{\Theta}_s$ is one-step plug-in bias correction estimator, but it is seen to be equivalent to the estimating equations estimator in the main text.

The remainder term

To simplify notation, in this subsection we largely omit function arguments, for example $\tau = \tau(X)$ with similar for $\hat{\tau}, \tau_s, \hat{\tau}_s, \pi, \hat{\pi}, \hat{\varphi}$. Evaluating the remainder $R \equiv E\{\hat{\phi}_s(Z) + \hat{\Theta}_s^0 - \Theta_s\}$ gives

$$R = E\left[\{\hat{\varphi} - \hat{\tau}_s\}^2 - \{\hat{\varphi} - \hat{\tau}\}^2 - \{\tau - \tau_s\}^2\right]$$

where we have used the fact that $\Theta_s = E[\{\tau - \tau_s\}^2]$. By algebraic manipulation, we write

$$R = E\left[\{\hat{\tau} - \hat{\tau}_s\}^2 - \{\tau - \tau_s\}^2 + 2\{\hat{\tau} - \hat{\tau}_s\}\{\hat{\varphi} - \hat{\tau}\}\right]$$

We then use the identity,

$$E\left[\{\hat{\tau} - \hat{\tau}_s\}^2 - \{\tau - \tau_s\}^2\right] = E\left[\{\tau_s - \hat{\tau}_s\}^2 - \{\tau - \hat{\tau}\}^2 + 2\{\hat{\tau} - \hat{\tau}_s\}\{\hat{\tau} - \tau\}\right]$$

to rewrite the remainder term as the sum of two error terms,

$$\begin{aligned}
R &= E\left[\{\hat{\tau} - \hat{\tau}_s\}^2 - \{\tau - \tau_s\}^2 + 2\{\hat{\tau} - \hat{\tau}_s\}\{\hat{\varphi} - \hat{\tau}\}\right] \\
&= E\left[\{\tau_s - \hat{\tau}_s\}^2 - \{\tau - \hat{\tau}\}^2 + 2\{\hat{\tau} - \hat{\tau}_s\}\{\hat{\tau} - \tau\} + 2\{\hat{\tau} - \hat{\tau}_s\}\{\hat{\varphi} - \hat{\tau}\}\right] \\
&= E\left[\{\tau_s - \hat{\tau}_s\}^2 - \{\tau - \hat{\tau}\}^2 + 2\{\hat{\tau} - \hat{\tau}_s\}\{\hat{\varphi} - \tau\}\right] \\
&= \underbrace{E\left[\{\tau_s - \hat{\tau}_s\}^2 - \{\tau - \hat{\tau}\}^2\right]}_{\text{CATE error}} + \underbrace{2E\left[\{\hat{\tau} - \hat{\tau}_s\}r\right]}_{\text{Pseudo-outcome error}}
\end{aligned}$$

where $r = r(X)$ is defined by

$$r(x) \equiv E\left[\hat{\varphi}|X = x\right] - \tau(x)$$

This represents a pseudo-outcome error in the sense that $r(x) = E\left[\hat{\varphi} - \varphi|X = x\right]$. Splitting the remainder in to two error terms allows us to consider that the CATE error is $o_P(n^{-1/2})$ when (A2) holds. For the pseudo-outcome error we use the Cauchy-Schwarz inequality to show that

$$E\left[\{\hat{\tau} - \hat{\tau}_s\}r\right]^2 \leq E\left[\{\hat{\tau} - \hat{\tau}_s\}^2\right] E\left[r^2\right] \leq \delta E\left[r^2\right]$$

with the second inequality following from (A3). Hence the pseudo-outcome error term is $o_P(n^{-1/2})$ if r is $o_P(n^{-1/2})$. By iterated expectation

$$r(x) = \left\{\frac{\pi(x)}{\hat{\pi}(x)} - 1\right\}\{\mu(1,x) - \hat{\mu}(1,x)\} - \left\{\frac{1-\pi(x)}{1-\hat{\pi}(x)} - 1\right\}\{\mu(0,x) - \hat{\mu}(0,x)\}$$

Using the inequality $(a+b)^2 \leq 2(a^2+b^2)$ then

$$\begin{aligned}
r^2(x) &\leq 2\left\{\frac{\pi(x)}{\hat{\pi}(x)} - 1\right\}^2\{\mu(1,x) - \hat{\mu}(1,x)\}^2 + 2\left\{\frac{1-\pi(x)}{1-\hat{\pi}(x)} - 1\right\}^2\{\mu(0,x) - \hat{\mu}(0,x)\}^2 \\
&\leq \left(\frac{2}{\epsilon^2}\right)\{\pi(x) - \hat{\pi}(x)\}^2\left[\{\mu(1,x) - \hat{\mu}(1,x)\}^2 + \{\mu(0,x) - \hat{\mu}(0,x)\}^2\right]
\end{aligned}$$

with the second inequality following from (A4). The final expression above is $o_P(n^{-1})$ under (A1), which completes the proof that R itself is $o_P(n^{-1/2})$.

The empirical process term

In this subsection we use a common empirical processes notation, where we define linear operators P and P_n such that for some function $h(Z)$, $P\{h(Z)\} \equiv E\{h(Z)\}$ and $P_n\{h(Z)\} \equiv n^{-1}\sum_{i=1}^{n} h(z_i)$. Hence we write $-E\{\hat{\phi}_s(Z)\}$ as

$$(P_n - P)\{\phi_s(Z)\} + (P_n - P)\{\hat{\phi}_s(Z) - \phi_s(Z)\} - P_n\{\hat{\phi}_s(Z)\}$$

which follows from adding and subtracting $(P_n - P)\{\phi_s(Z)\}$ and $P_n\{\hat{\phi}_s(Z)\}$ to $-P\{\hat{\phi}_s(Z)\}$. This expression recovers (D.5) since the IC is mean zero, in the sense that $P\{\phi_s(Z)\} = 0$, and we define the empirical process term

$$H_n \equiv (P_n - P)\{\hat{\phi}_s(Z) - \phi_s(Z)\}$$

By e.g. Lemma 19.24 of van der Vaart (1998a), H_n is $o_P(n^{-1/2})$ under (A5) provided that $P\left[\left\{\hat{\phi}_s(Z) - \phi_s(Z)\right\}^2\right]$ converges to zero in probability.

Start by writing,

$$\begin{aligned}\hat{\phi}_s - \phi_s =& 2(\hat{\varphi} - \varphi)(\hat{\tau} - \hat{\tau}_s) \\ &+ 2(\varphi - \tau_s)(\tau_s - \hat{\tau}_s) + (\tau_s - \hat{\tau}_s)^2 \\ &- 2(\varphi - \tau)(\tau - \hat{\tau}) - (\tau - \hat{\tau})^2 \\ &- (\hat{\Theta}_s^0 - \Theta_s)\end{aligned}$$

Using the inequality $(a+b)^2 \leq 2(a^2 + b^2)$,

$$\begin{aligned}P\left[\left\{\hat{\phi}_s(Z) - \phi_s(Z)\right\}^2\right] \leq \quad & 8P\left\{(\hat{\varphi} - \varphi)^2(\hat{\tau} - \hat{\tau}_s)^2\right] \\ &+ 2P\Big[\Big\{2(\varphi - \tau_s)(\tau_s - \hat{\tau}_s) + (\tau_s - \hat{\tau}_s)^2 \\ &\quad - 2(\varphi - \tau)(\tau - \hat{\tau}) - (\tau - \hat{\tau})^2 \\ &\quad - (\hat{\Theta}_s^0 - \Theta_s)\Big\}^2\Big]\end{aligned}$$

Letting $\hat{\Theta}_s^0 = P_n\{(\hat{\tau} - \hat{\tau}_s)^2\}$ then, in view of Theorem 1 of Williamson et al. (2021a), the second term converges to zero under (A2) and (A6). For the first of terms, we note that (A3) implies

$$P\left\{(\hat{\varphi} - \varphi)^2(\hat{\tau} - \hat{\tau}_s)^2\right] \leq \delta P\left\{(\hat{\varphi} - \varphi)^2\right].$$

Similar terms to $P\left\{(\hat{\varphi} - \varphi)^2\right]$ appear in the ATE empirical process literature. In view of Theorem 5.1 of Chernozhukov et al. (2018), this term is also converges to zero under (A1), (A4) and (A7).

Thus $H_n = o_P(n^{-1/2})$ which completes the proof.

D.2.2 Proof of Theorem 4

Under (A1)-(A7) and

(B1) The difference $\tau_p - \hat{\tau}_p$ is $o_P(n^{-1/4})$.

(B2) The CATE difference estimates are bounded as $\{\hat{\tau}(x) - \hat{\tau}_p\}^2 \leq \delta$ for some $\delta < \infty$ with probability 1.

(B3) There exists a P-Donsker class $\mathcal{G}_0$ such that $P(\hat{\phi}_p(.) \in \mathcal{G}_0) \to 1$.

(B4) There exists a constant $K > 0$ such that $\hat{\tau}_p \in (-K, K)$.

Then we have regular asymptotically linear estimators such that,

$$\hat{\Theta}_s - \Theta_s = n^{-1} \sum_{i=1}^{n} \phi_s(z_i) + o_p(n^{-1/2})$$
$$\hat{\Theta}_p - \Theta_p = n^{-1} \sum_{i=1}^{n} \phi_p(z_i) + o_p(n^{-1/2})$$

It follows by algebraic manipulations that,

$$\sqrt{n}(\hat{\Psi}_s - \Psi_s) = \frac{\Theta_p}{\hat{\Theta}_p} \left[n^{-1/2} \sum_{i=1}^{n} \Phi_s(z_i) + o_p(1) \right]$$

where $\Phi_s(z) = \{\phi_s(z) - \Psi_s \phi_p(z)\}/\Theta_p$ is the IC of Ψ_s. Next we use Slutsky's Theorem and the fact that $\hat{\Theta}_p/\Theta_p$ converges to 1 in probability, to write,

$$\lim_{n\to\infty} \sqrt{n}(\hat{\Psi}_s - \Psi_s) = \lim_{n\to\infty} n^{-1/2} \sum_{i=1}^{n} \Phi_s(z_i)$$

which gives the desired result due to the central limit theorem. We note that this set up is quite general when one considers estimands which are written as the ratio of two other estimands, such as Ψ_s in the present context.

D.3 Additional Plots

D.3.1 Simulation plots

Figure D.1 shows the performance of the estimator $\hat{\Psi}_2$, of the TE-VIM Ψ_2, as described in the simulation study in Section 5.3.

D.3.2 Applied example

Figure D.2 gives TE-VIM estimates from the ACTG175 where the discrete Super Learner (20 cross validation folds) is used for functional estimation. The plots for Algorithms 1A and 2A, i.e. the T-learner without and with sample splitting, appear highly uninformative. For Algorithm 1A this is due to the point estimate of the VTE being very close to zero -3.80×10^{-9} mm^{-6} (CI: -1.7×10^{-3},1.7×10^{-3}). For Algorithm 2A, the VTE estimate is also negative -107 mm^{-6} (CI:-159,-57), and does not overlap with zero. Since both point estimates are negative, the null hypothesis that the VTE$\leq$ 0 has p-value exactly equal to 1.

For Algorithm 2A, all of the TE-VIM point estimates are negative, meaning that all of the corresponding Θ_s estimates are positive. The negative VTE, therefore has the effect of reversing the order of importance, with CD4 count at baseline appearing at the bottom of the plot for Algorithm 2A in Figure D.2.

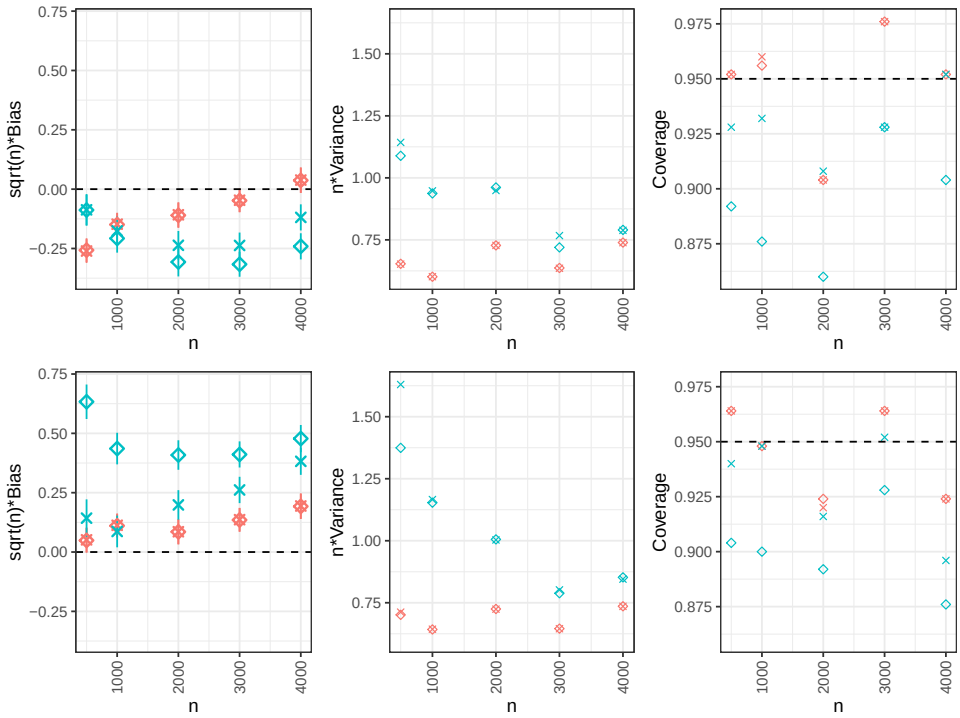


Figure D.1: Bias, variance and coverage for $\hat{\Psi}_2$ using 1000 sampled datasets. Red and blue points indicate that working models are fitted using generalised additive modelling and random forests respectively. Top row of plots corresponds to Algorithm 1 (no sample splitting) and the bottom row corresponds to Algorithm 2 (sample splitting). Square and crossed points indicate that the algorithm used the T-learner and DR-learner respectively for CATE estimation.

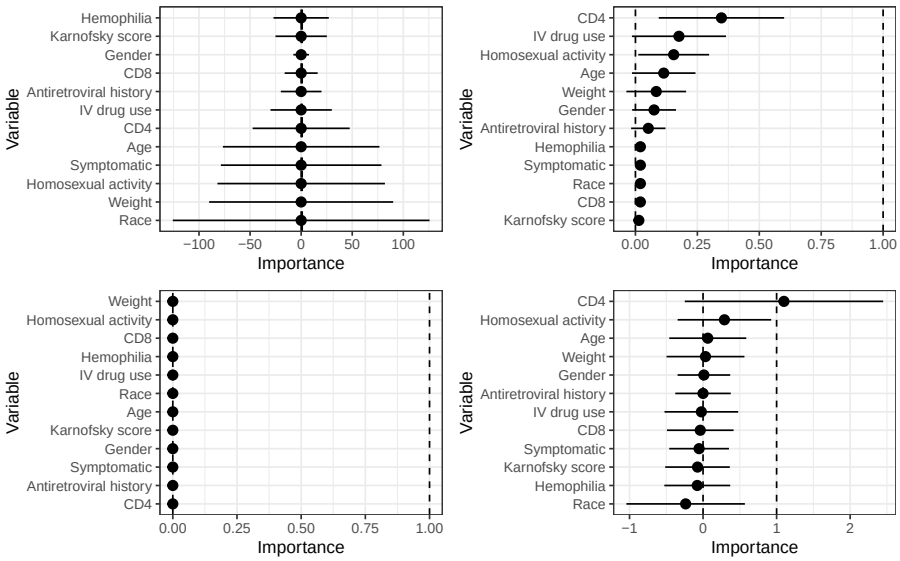


Figure D.2: TE-VIM estimates from the ACTG175 study using the discrete Super Learner for functional estimation. Top row: no sample splitting (Algorithm 1). Bottom row: with sample splitting (Algorithm 2). Left col: T-learner (A). Right col: DR-learner (B). Black lines indicate 95% Confidence intervals. Confidence intervals in the bottom left plot are so small that they are not visible. In each plot, covariates are sorted according to their TE-VIM point estimate. Dashed lines indicate the $[0, 1]$ support of the TE-VIM.

D.4 TE-CDF bounds

First note Chebyshev's inequality: For a variable V with mean μ and variance σ^2, for $k > 0$

$$\begin{aligned} & Pr(|V - \mu| \geq k\sigma) \leq k^{-2} \\ \implies & Pr(V \geq \mu + k\sigma) + Pr(V \leq \mu - k\sigma) \leq k^{-2} \end{aligned}$$

which implies the weaker inequality,

$$Pr(V \leq \mu - k\sigma) \leq k^{-2}$$

Let $\tau(X)$ be the CATE with ATE τ_p and VTE Θ_p then,

$$\beta(0) = Pr\left\{\tau(X) \leq 0\right\} = Pr\left\{\tau(X) \leq \tau_p - \left(\frac{\tau_p}{\sqrt{\Theta_p}}\right)\sqrt{\Theta_p}\right\} \leq \frac{\Theta_p}{\tau_p^2}$$

Where the inequality applies only when $\tau_p > 0$. It follows that, when the ATE is positive, the quantity on the RHS bounds $\beta(0)$ from above. The quotient rule gives that the IC (pathwise derivative) is,

$$\begin{aligned} \phi_\beta(Z) &= \frac{1}{\tau_p^2}\phi_p(Z) - 2\left(\frac{\Theta_p}{\tau_p^3}\right)\left\{\varphi(Z) - \tau_p\right\} \\ &= \frac{\left\{\varphi(Z) - \tau_p\right\}^2 - \left\{\varphi(Z) - \tau(X)\right\}^2 - \left(\frac{\Theta_p}{\tau_p^2}\right)\tau_p\{2\varphi(Z) - \tau_p\}}{\tau_p^2} \end{aligned}$$

where $\{\varphi(Z) - \tau_p\}$ is the IC of τ_p. An estimating equations estimator is that which solves

$$n^{-1}\sum_{i=1}^{n}\hat{\phi}_\beta(z_i) = 0$$

where $\hat{\phi}_\beta(z)$ is an estimate of $\phi_\beta(z)$. Therefore $\hat{\Theta}_p/\hat{\tau}_p^2$ is an estimating equations estimator where $\hat{\Theta}_p$ is the VTE estimator in the current paper and

$$\hat{\tau}_p = n^{-1}\sum_{i=1}^{n}\hat{\varphi}(z_i)$$

is the AIPW estimator of the ATE.

D.5 IC for continuous analogue estimands

Here we use the same fomalism as in Appendix D.1. To derive the ICs of interest we consider the results in (D.2) and (D.4) in the setting where we set $g_P(x) = \lambda(x)$. We will show that,

$$\partial_t g_{P_t}(x) = \frac{\tilde{f}(x)}{f(x)}\{\tilde{y} - \mu(x) - \lambda(x)\{\tilde{a} - \pi(x)\}\}\frac{\tilde{a} - \pi(x)}{\text{var}(A|X = x)} \tag{D.6}$$

and hence, letting

$$\varphi_\lambda(z) \equiv \{y - \mu(x) - \lambda(x)\{a - \pi(x)\}\}\frac{a - \pi(x)}{\text{var}(A|X = x)} + \lambda(x)$$

then by (D.2) the IC of $E\{\lambda(X)\}$ is,

$$\varphi_\lambda(z) - E\{\lambda(X)\}$$

and by (D.4), the IC of $E\left[\text{var}\{\lambda(X)|X_{-s}\}\right]$ is,

$$\{\varphi_\lambda(z) - \lambda_s(x)\} - \{\varphi_\lambda(z) - \lambda(x)\} - E\left[\text{var}\{\lambda(X)|X_{-s}\}\right]$$

where $\lambda_s(x) = E\{\lambda(X)|X_{-s} = x_{-s}\}$. The IC for $\text{var}\{\lambda(X)\}$ follows as a special case where s includes all the observed covariates. To demonstrate (D.6) we first note that, by (D.1),

$$\begin{aligned}\partial_t \text{cov}_{P_t}(A, Y|X = x) &= \partial_t E_{P_t}\{[A - E_{P_t}(A|X)][Y - E_{P_t}(Y|X)]|X = x\} \\ &= \frac{\tilde{f}(x)}{f(x)}\left[\{\tilde{a} - \pi(x)\}\{\tilde{y} - \mu(x)\} - \text{cov}_P(A, Y|X = x)\right]\end{aligned}$$

We also obtain $\partial_t \text{var}_{P_t}(A|X = x)$ as a special case of the above expression when $Y = A$. By the quotient rule,

$$\begin{aligned}\partial_t \frac{\text{cov}_{P_t}(A, Y|X = x)}{\text{var}_{P_t}(A|X = x)} &= \frac{\partial_t \text{cov}_{P_t}(A, Y|X = x)}{\text{var}_P(A|X = x)} - \frac{\text{cov}_P(A, Y|X = x)}{\text{var}_P(A|X = x)}\frac{\partial_t \text{var}_{P_t}(A, Y|X = x)}{\text{var}_P(A|X = x)} \\ &= \frac{\tilde{f}(x)}{f(x)}\{\tilde{y} - \mu(x) - \lambda(x)\{\tilde{a} - \pi(x)\}\}\frac{\tilde{a} - \pi(x)}{\text{var}(A|X = x)}\end{aligned}$$

Thus, the desired results follow.

Appendix E

Supplement to nonparametric score testing

E.1 Non-invariance to reparameterisation of Wald CIs

In Section 6.2.3 of the main text, we showed how the TMLE estimator achieves invariance to differentiable reparameterisations of the estimand by debiasing the initial plug-in estimator in the distribution space, rather than in the estimand space. Wald confidence sets, centred on TMLE point estimators, however, are not invariant to differentiable reparameterisations of the estimand, as we demonstrate here. Trivially, Wald confidence sets centered on one-step bias correction point estimators are also not invariant to differentiable reparameterisations of the estimand, since the point estimator itself is not invariant to such reparameterisations.

Consider the setting where $d = q = 1$ and $h(.)$ is monotonic, i.e. $\Psi(P_0)$ and $h\{\Psi(P_0)\}$ represent scalar estimands. Letting $\Psi(\hat{P}_n^*)$ denote a TMLE estimator of $\Psi(P_0)$, then since $h\{\Psi(\hat{P}_n^*)\}$ is an RAL estimator for $h\{\Psi(P_0)\}$

$$h\{\Psi(\hat{P}_n^*)\} - h\{\Psi(P_0)\} = h'\{\Psi(P_0)\}U_n(P_0) + o_p(n^{-1/2})$$
$$\sqrt{n}[h\{\Psi(\hat{P}_n^*)\} - h\{\Psi(P_0)\}] \xrightarrow{d} \mathcal{N}\left(0, h'\{\Psi(P_0)\}^2 I_0\right)$$

where $h'(.)$ denotes the derivative of $h(.)$. Estimating the variance, $h'\{\Psi(P_0)\}^2 I_0$ by $h'\{\Psi(\hat{P}_n^*)\}^2 I_n(\hat{P}_n^*)$ results in a Wald CI for $h\{\Psi(P_0)\}$, as the set of values h_0 which satisfy

$$\frac{n\left[h\{\Psi(\hat{P}_n^*)\} - h_0\right]^2}{h'\{\Psi(\hat{P}_n^*)\}^2 I_n(\hat{P}_n^*)} \leq c_\alpha^2$$

This inequality implies the Wald CI for $h\{\Psi(P_0)\}$

$$h\{\Psi(\hat{P}_n^*)\} \pm h'\{\Psi(\hat{P}_n^*)\}\sqrt{\frac{c_\alpha^2 I_n(\hat{P}_n^*)}{n}} \tag{E.1}$$

Alternatively, one might have first constructed a Wald CI for $\Psi(P_0)$ centred on the TMLE point estimator

$$\Psi(\hat{P}_n^*) \pm \sqrt{\frac{c_\alpha^2 I_n(\hat{P}_n^*)}{n}} \tag{E.2}$$

then, letting $h[.]$ denote the image of $h(.)$, one obtains a CI for $h\{\Psi(P_0)\}$ as

$$h\left[\Psi(\hat{P}_n^*) \pm \sqrt{\frac{c_\alpha^2 I_n(\hat{P}_n^*)}{n}}\right]. \tag{E.3}$$

Generally the interval in (E.3) is not equal to the interval in (E.1). It follows that the Wald type confidence sets are not invariant to differentiable reparameterisations of the estimand, even when they are centred on TMLE point estimators. Moreover, there is no guarantee that the interval in (E.2) lies in the domain of $h(.)$, in which case (E.3) is not well defined. When (E.2) does lie in the domain of $h(.)$, however, a Taylor expansion of (E.3) gives

$$h\{\Psi(\hat{P}_n^*)\} \pm h'\{\Psi(\hat{P}_n^*)\}\sqrt{\frac{c_\alpha^2 I_n(\hat{P}_n^*)}{n}} + o_P(n^{-1/2})$$

from which we conclude that the Wald intervals in (E.1) and (E.3) are asymptotically equivalent up to a term which is $o_p(n^{-1/2})$.

We reason that Wald confidence sets are generally not invariant to differentiable reparmeterizations of the estimand because they are constructed in the estimand space, rather than the distribution space. In the next section, we propose a score confidence set construction method that delivers invariant confidence sets by constructing the confidence set in the space of distributions. As such, the comparison between the proposed score confidence sets and nonparametric Wald confidence sets is analogous to the comparison between TMLE point estimators and one-step bias-correction point estimators, described in Section 6.2.2.

E.2 Significance Threshold discussion

Here we discuss three different methods for obtaining a threshold k which can be compared against the score statistic $M_n(P)$, i.e. where the score test is defined as the set of distributions for which $M_n(P) \leq k$. In the main text, we advocate the value $k = c_\alpha^2/n$, where c_α^2 is the $1-\alpha$ quantile of a χ_d^2 random variable. This is a natural choice, since $nM_n(P_0) \xrightarrow{d} \chi_d^2$.

Instead of basing inference on the score statistic $M_n(P)$, one could have based score type inference on an analogous statistic where the covariance matrix I_0 is estimated as the 'mean corrected' covariance matrix $\tilde{I}_n(P_0)$ in (6.6). The resulting 'mean corrected score statistic' $\tilde{M}_n(P) \equiv \boldsymbol{U}_n^\top(P)\tilde{I}_n^{-1}(P)\boldsymbol{U}_n(P)$ shares the same asymptotic distribution as the proposed score statistic $M_n(P)$ since $\tilde{I}_n(P_0) - I_n(P_0) \xrightarrow{p} 0$ as $n \to \infty$.

For the purposes of score interval construction, using the mean corrected score statistic in place of $M_n(P)$ is equivalent to using the score statistic $M_n(P)$, but with a different value for k. To see why, consider the 'mean corrected' score statistic

$$\tilde{M} = \tilde{M}_n(P) = u^\top(I - uu^\top)^{-1}u$$

where for convenience we let $u = U_n(P)$ and $I = I_n(P)$ so that the score statistic is written $M = M_n(P) = u^\top I^{-1}u$. Multiplying by $(1-M)$ gives

$$\begin{aligned}\tilde{M}(1-M) &= u^\top(I - uu^\top)^{-1}u(1 - u^\top I^{-1}u) \\ &= u^\top(I - uu^\top)^{-1}(I - uu^\top)I^{-1}u = M\end{aligned}$$

Hence

$$\tilde{M} = \frac{M}{1-M}$$
$$M = \frac{\tilde{M}}{1+\tilde{M}}$$

and we conclude that

$$\tilde{M} \leq k \iff M \leq \frac{k}{1+k}.$$

A score set based on the asymptotic χ_d^2 distribution of $\tilde{M}$ is therefore equivalent to score set based on M, which uses $k = c_\alpha^2/(n + c_\alpha^2)$. This choice of threshold is expected to affect inference only in small samples,

since the thresholds c_α^2/n and $c_\alpha^2/(n+c_\alpha^2)$ are asymptotically equivalent. We remark that asymptotically valid inference is obtained when k is set to any sequence k_n such that $nk_n \to c_\alpha^2$.

By construction, $\tilde{M} \geq 0$, which implies that $M \in [0,1)$. There is, however, no guarantee that c_α^2/n lies on the interval $[0,1)$, and it is possible for the threshold c_α^2/n to be outside of $[0,1)$, for example when the dimension d is large, n is small, and α is small, i.e. a wide confidence interval for a high-dimensional estimand with few observations. This represents an overly ambitious setting in practice e.g. for $\alpha = 0.05, d = 100, n = 124$ then $c_\alpha^2 \approx 1.002$, and hence no values will be excluded from the score set.

A philosophically appealing fix for this issue is to use consider the distribution of $M_n(P_0)$ when $\boldsymbol{\phi}(Z, P_0)$ is known to be normally distributed. In such a setting, the exact distribution of $M_n(P_0)$ is also known

$$M_n(P_0) \sim \text{Beta}\left(\frac{d}{2}, \frac{n-d}{2}\right) \tag{E.4}$$

To see why, note that when $\boldsymbol{\phi}(Z, P_0) \sim \mathcal{N}(0, I_0)$ is normally distributed with mean zero and covariance matrix I_0 then, $t^2 \equiv (n-1)\tilde{M}(P_0)$ is an 'Hotelling's t-squared statistic', a multivariate version of the more familiar 'Student's t-statistic' (Hotelling, 1931), and hence

$$\frac{n-d}{d}\tilde{M}(P_0) \sim F_{d,n-d}$$

where $F_{d,n-d}$ denotes an F distributed random variable with d and $n-d$ degrees of freedom. The result in (E.4) follows immediately from this observation, since, for an F-distributed random variable $A \sim F_{p,q}$

$$\frac{pA/q}{1+pA/q} \sim \text{Beta}\left(\frac{p}{2}, \frac{q}{2}\right).$$

This can be shown by manipulation of the relevant density functions, and it helps to use the fact that for $Y = aX/(1-aX)$

$$F_Y(y) = P\left(\frac{aX}{1-aX} < y\right) = P\left(X < \frac{y}{k-ay}\right) = F_X\left(\frac{y}{a-ay}\right)$$

The result in (E.4) for normally distributed ICs suggests that one could set $k = B_{\alpha,d,n}$ where $B_{\alpha,d,n}$ denotes the $1-\alpha$ quantile of a Beta$\{d/2, (n-d)/2\}$ distributed random variable. This threshold is asymptotically valid even when $\boldsymbol{\phi}(Z, P_0)$ is not normally distributed since $nB_{\alpha,d,n} \to c_\alpha^2$. Unlike c_α^2, however, the threshold k based on the beta distribution is guaranteed to lie on the interval $[0,1]$. The three thresholds values are plotted in Figure E.1 for some small sample sizes.

E.3 Review of generalised methods of moments

The score statistic $M_n(P)$ is analogous to parametric GMM hypothesis test statistics considered by Newey and West (1987); Hansen et al. (1996); Dufour et al. (2017). In this Section we sketch the relevant GMM hypothesis test statistics, with requisite regularity assumptions provided by Dufour et al. (2017). To make the analogy clear, we deliberately reuse the notation, which we used for the nonparametric score statistics in the main text. Consider inference of a parameter $\theta_0 \in \Theta \subseteq \mathbb{R}^p$ that uniquely satisfies a set of $d \leq p$ moment conditions $P_0\{\boldsymbol{\phi}(Z, \theta_0)\} = 0$, where $\boldsymbol{\phi}(z, \theta)$ represents a 'moment function'. Define

$$\boldsymbol{U}_n(\theta) = n^{-1}\sum_{i=1}^n \boldsymbol{\phi}(z_i, \theta)$$
$$I_n(\theta) = n^{-1}\sum_{i=1}^n \boldsymbol{\phi}(z_i, \theta)\boldsymbol{\phi}^\top(z_i, \theta)$$

such that, $\sqrt{n}\boldsymbol{U}_n(\theta_0) \xrightarrow{d} \mathcal{N}(0, I_0)$ where $I_0 = \text{plim}_{n\to\infty} I_n(\theta_0)$, and it is assumed that I_0 is nonsingular and $I_0 < \infty$ in the positive definite sense. Letting $M_n(\theta) = \boldsymbol{U}_n^\top(\theta) I_n^{-1}(\theta) \boldsymbol{U}_n(\theta)$, then, according to Hansen et al.

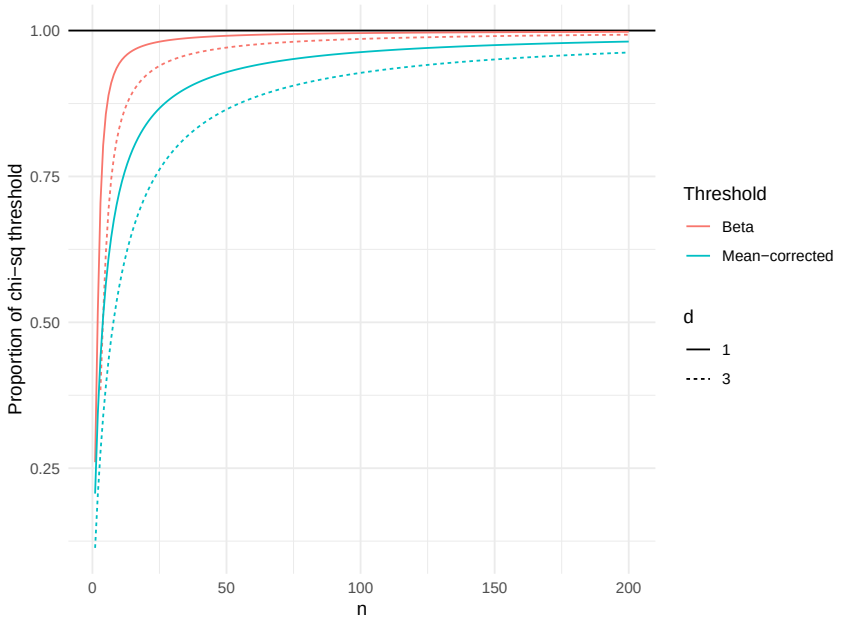


Figure E.1: Score threshold plotted against sample size for $\alpha = 0.05$ and $d = 1, 3$. Here 'Mean-corrected' refers to the threshold $k = c_\alpha^2/(n + c_\alpha^2)$, whilst 'Beta' refers to the threshold $k = B_{\alpha,d,n}$. In both cases these are normalised by the threshold $k = c_\alpha^2/n$, which is represented by the line at $y = 1$.

(1996) the so-called GMM-'continuously updated estimator'(GMM-CUE) of θ_0 is $\hat{\theta}^* = \arg\min_{\theta\in\Theta} M_n(\theta)$. If we assume that the Jacobian matrix V_0 below exists, then the GMM-CUE estimator is RAL in the sense that

$$\hat{\theta}^* - \theta_0 = -V_0 U_n(\theta_0) + o_p(n^{-1/2})$$
$$V_0 = P_0 \left\{ \frac{d\boldsymbol{\phi}(z_i, \theta_0)}{d\theta} \right\}$$

Next suppose that our goal is to infer some other quantity $\boldsymbol{\Psi}(\theta_0)$ where $\boldsymbol{\Psi} : \Theta \mapsto \mathbb{R}^d$. Technically the dimension of $\boldsymbol{\Psi}$ is required to be less than or equal to the dimension of the moment conditions, though for clarity we use d for both here. In this setting, Dufour et al. (2017) propose testing the hypothesis $H_0 : \boldsymbol{\Psi}(\theta_0) = \boldsymbol{\psi}_0$ using the statistic,

$$D_n \equiv n \left\{ M_n(\hat{\theta}) - M_n(\hat{\theta}^*) \right\}$$

where in a slight abuse of notation $H_0 \subseteq \Theta$ denotes the set of parameter values θ satisfying $\boldsymbol{\Psi}(\theta) = \boldsymbol{\psi}_0$ and

$$\hat{\theta} = \arg\min_{\theta \in H_0} M_n(\theta)$$

It follows from Dufour et al. (2017) that, provided requisite regularity conditions hold, then $D_n \xrightarrow{d} \chi_d^2$ under H_0. Moreover, when $\boldsymbol{U}_n(\hat{\theta}^*) = 0$ then $D_n = nM_n(\hat{\theta}) \xrightarrow{d} \chi_d^2$. Hence, a confidence set for $\boldsymbol{\Psi}$ at significance level α can be constructed as

$$\left\{ \boldsymbol{\Psi}(\theta) \text{ such that } M_n(\theta) \leq \frac{c_\alpha^2}{n} \right\}.$$

Comparing this description of GMM estimators to the discussion in the main text, we see that, our nonparametric developments are connected to the GMM in the sense that the nonparametric model

$\mathcal{M}$ plays the role of the parameter set Θ, and that the influence curve is treated as a GMM moment function. Since the GMM generalises likelihood based inference, the interpretation of the influence curve as a GMM moment function is compatible with the framework of TMLE, where the influence curve is treated like a score function associated with some likelihood. Just as TMLE follows likelihood based inference once a parametric submodel has been constructed, our proposed DIE interval estimators follow GMM based inference once a parametric submodel has been constructed around a TMLE distribution estimator. Connecting nonparametric inference and TMLE to the GMM is potentially significant since it is possible that other GMM techniques could be applied to nonparametric inference problems, e.g. empirical likelihoods and exponentially tilted GMM estimators (Owen, 1988; Qin and Lawless, 1994; Kitamura and Stutzer, 1997; Imbens, 1997; Corcoran, 1998; Imbens, 2002; Newey and Smith, 2004).

Appendix F

Supplement to optimally weighted average derivative effects

Derivation of (7.4)

Assume (C1), $\tilde{f}(a|\boldsymbol{z}) \equiv w(a|\boldsymbol{z})f(a|\boldsymbol{z})$ is differentiable, and (C2), $\tilde{f}(s|\boldsymbol{z}) = \tilde{f}(t|\boldsymbol{z}) = 0$, where s and t denote the boundary of the support of A. Integration by parts gives,

$$\begin{aligned} E\{w(A|\boldsymbol{Z})\mu'(A,\boldsymbol{Z})|\boldsymbol{Z}=\boldsymbol{z}\} &= \int_s^t \mu'(a,\boldsymbol{z})\tilde{f}(a|\boldsymbol{z})da \\ &= \mu(t,\boldsymbol{z})\tilde{f}(t|\boldsymbol{z}) - \mu(s,\boldsymbol{z})\tilde{f}(s|\boldsymbol{z}) - \int_s^t \mu(a,\boldsymbol{z})\tilde{f}'(a|\boldsymbol{z})da \\ &= E\{l(A|\boldsymbol{Z})\mu(A,\boldsymbol{Z})|\boldsymbol{Z}=\boldsymbol{z}\} = E\{l(A|\boldsymbol{Z})Y|\boldsymbol{Z}=\boldsymbol{z}\} \end{aligned}$$

To motivate Theorem 7, note that inverting (7.4) gives

$$\tilde{f}(a|\boldsymbol{z}) = -\int_s^a l(a^*|\boldsymbol{z})f(a^*|\boldsymbol{z})da^* = -E\{l(A|\boldsymbol{Z})|A \leq a, \boldsymbol{Z}=\boldsymbol{z}\}F(a|\boldsymbol{z})$$

Proof of Theorem 7

Theorem 7 essentially follows by the integration by parts argument above. Rather than work with the exposure weight in (7.5) directly, we consider the function $\tilde{f}(a|\boldsymbol{z}) \equiv w(a|\boldsymbol{z})f(a|\boldsymbol{z})$. Our goal is to show that this integrates to 1, i.e. $E\{w(A|\boldsymbol{Z})|\boldsymbol{Z}\} = 1$, and that it satisfies (C2). Note (C1) is satisfied since $\tilde{f}'(a|\boldsymbol{z}) = -l(a|\boldsymbol{z})f(a|\boldsymbol{z})$ by the fundamental theorem of calculus. To do so, let $\Theta(u)$ denote a step function which is 1 for $u \geq 0$ and 0 for $u < 0$ and hence,

$$\tilde{f}(a|\boldsymbol{z}) = -\int_s^t \Theta(a-a^*)l(a^*|\boldsymbol{z})dP(a^*|\boldsymbol{z})$$

where $dP(a|\boldsymbol{z})$, is the probability measure of A given $\boldsymbol{Z}$. Integrating over a, gives

$$\int_s^t \tilde{f}(a|\boldsymbol{z})da = \int_s^t \left[\int_s^t -\Theta(a-a^*)da\right] l(a^*|\boldsymbol{z})dP(a^*|\boldsymbol{z})$$

For the part in the square brackets,

$$\int_s^t -\Theta(a-a^*)da = \int_{a^*}^t -1dx = a^* - t$$

Hence,

$$\int_s^t \tilde{f}(a|\boldsymbol{z})da = E\{l(A|\boldsymbol{Z})A|\boldsymbol{Z}=\boldsymbol{z}\} - tE\{l(A|\boldsymbol{Z})|\boldsymbol{Z}=\boldsymbol{z}\} = 1$$

Thus $\tilde{f}(a|\boldsymbol{z})$ integrates to 1. As a side note, when the exposure weight is non-negative then $\tilde{f}(a|\boldsymbol{z})$ is a density function. Next we show $\tilde{f}(a|\boldsymbol{z})$ satisfies (C2) when A is a continuous random variable. Since $a^* \leq t$, then $\Theta(t-a^*)=1$, hence

$$\tilde{f}(t|\boldsymbol{z}) = -\int_s^t l(a^*|\boldsymbol{z})dP(a^*|\boldsymbol{z}) = -E\{l(A|\boldsymbol{Z})|\boldsymbol{Z}=\boldsymbol{z}\} = 0$$

Similarly, since $a^* \geq s$, then $\Theta(s-a^*)=1$ only for $a^*=s$,

$$\tilde{f}(s|\boldsymbol{z}) = -\int_s^s l(a^*|\boldsymbol{z})dP(a^*|\boldsymbol{z}) = -l(s|\boldsymbol{z})P(A=s|\boldsymbol{Z}=\boldsymbol{z})$$

Since A is continuous, $P(A=s|\boldsymbol{Z}=\boldsymbol{z})=0$. This completes the proof.

Proof of Lemma 4

First we prove that Theorem 7 is satisfied when $l(a|\boldsymbol{z})$ is monotonically increasing and (D1) $E\{l(A|\boldsymbol{Z})|\boldsymbol{Z}\} = 0$ and (D2) $E\{l(A|\boldsymbol{Z})A|\boldsymbol{Z}\} = 1$ almost surely. We split $l(a|\boldsymbol{z})$ into a positive and negative part by defining two non-negative functions, $l^+(a|\boldsymbol{z}) = \max\{l(a|\boldsymbol{z}),0\}$ and $l^-(a|\boldsymbol{z}) = \max\{-l(a|\boldsymbol{z}),0\}$ such that, $l(a|\boldsymbol{z}) = l^+(a|\boldsymbol{z}) - l^-(a|\boldsymbol{z})$. It follows from (D1) that,

$$E\{l^+(A|\boldsymbol{Z})|\boldsymbol{Z}\} = E\{l^-(A|\boldsymbol{Z})|\boldsymbol{Z}\}$$

This equality is satisfied by $l^+(a|\boldsymbol{z}) = l^-(a|\boldsymbol{z}) = 0$, however this solution violates (D2), hence the positive and negative parts are both non-zero. Since, $l(a|\boldsymbol{z})$ is monotonically increasing there must be some value, $c = c(\boldsymbol{z})$, on the support of A, such that the positive part is zero for $a < c$ and the negative part is zero for $a \geq c$, i.e.

$$l(a|\boldsymbol{z}) = l^+(a|\boldsymbol{z})\Theta(a-c) - l^-(a|\boldsymbol{z})\{1-\Theta(a-c)\}$$

First consider the inequality in (7.5) when $a < c$,

$$\int_s^a l(a^*|\boldsymbol{z})dP(a^*|\boldsymbol{z}) = -\int_s^a l^-(a^*|\boldsymbol{z})dP(a^*|\boldsymbol{z}) \leq 0$$

When $a \geq c$,

$$\int_s^a l(a^*|\boldsymbol{z})dP(a^*|\boldsymbol{z}) = \int_c^a l^+(a^*|\boldsymbol{z})dP(a^*|\boldsymbol{z}) - \int_s^c l^-(a^*|\boldsymbol{z})dP(a^*|\boldsymbol{z})$$

The first part on the right hand side is $\leq E\{l^+(A|\boldsymbol{Z})|\boldsymbol{Z}=\boldsymbol{z}\}$ and the second part is $= E\{l^-(A|\boldsymbol{Z})|\boldsymbol{Z}=\boldsymbol{z}\}$ therefore, in both cases,

$$\int_s^a l(a^*|\boldsymbol{z})dP(a^*|\boldsymbol{z}) \leq 0$$

Hence $\tilde{f}(a|\boldsymbol{z}) \geq 0$, so the inequality in (7.5) is satisfied. The proof is completed by verifying that the contrast function in (7.7) is monotonically increasing when $v(a,\boldsymbol{z})$ is monotonically increasing or decreasing but not constant. This is fairly straight forward and we note that the decreasing case follows from the increasing case since (7.7) is invariant to replacing $v(a,\boldsymbol{z})$ with $-v(a,\boldsymbol{z})$.

Proof of Theorem 8

An efficient estimator of θ_w is regular asymptotically linear, such that

$$\hat{\theta}_w = \theta_w + n^{-1}\sum_{i=1}^{n}\phi_{\theta,w}(\boldsymbol{o}_i) + o_p(n^{-1/2})$$

where $\phi_{\theta,w}(\boldsymbol{o})$ is the influence curve of θ_w in (7.10). Hence,

$$n^{1/2}(\hat{\theta}_w - \theta_{w,S}) = n^{-1/2}\sum_{i=1}^{n} w(\boldsymbol{z}_i)l(a_i|\boldsymbol{z}_i)\{y_i - \mu(a_i,\boldsymbol{z}_i)\} + o_p(1)$$

By the central limit theorem, $n^{1/2}(\hat{\theta}_w - \theta_{w,S}) \xrightarrow{d} \mathcal{N}(0,V)$, where the efficiency bound is

$$\begin{aligned} V &= E\left\{w(\boldsymbol{Z})^2 l(A|\boldsymbol{Z})^2\{Y-\mu(A,\boldsymbol{Z})\}^2\right\} \\ &= E\left\{w(\boldsymbol{Z})^2 l(A|\boldsymbol{Z})^2 E\left(\{Y-\mu(A,\boldsymbol{Z})\}^2|X,Z\right)\right\} \\ &= E\left\{w(\boldsymbol{Z})^2 l(A|\boldsymbol{Z})^2\sigma^2(A,\boldsymbol{Z})\right\} \\ &= E\left\{w(\boldsymbol{Z})^2 E[l(A|\boldsymbol{Z})^2\sigma^2(A,\boldsymbol{Z})|\boldsymbol{Z}]\right\} \end{aligned}$$

Minimizing $E\{l^2(A|\boldsymbol{Z})\sigma^2(A,\boldsymbol{Z})|\boldsymbol{Z}=\boldsymbol{z}\}$ subject to $E\{l(A|\boldsymbol{Z})|\boldsymbol{Z}=\boldsymbol{z}\}=0$ and $E\{l(A|\boldsymbol{Z})A|\boldsymbol{Z}=\boldsymbol{z}\}=1$. Using Lagrange multipliers, λ_1, λ_2, which are both constant given $\boldsymbol{Z}=\boldsymbol{z}$,

$$\int l^2(a|\boldsymbol{z})\sigma^2(a,\boldsymbol{z}) - 2\lambda_1 l(a|\boldsymbol{z}) - 2\lambda_2\{l(a|\boldsymbol{z})a - 1\}dP(a|\boldsymbol{z})$$

Differentiating the Lagrangian with respect to $l(a|\boldsymbol{z})$ and setting equal to zero gives.

$$l(a|\boldsymbol{z}) = \frac{\lambda_1 + \lambda_2 a}{\sigma^2(a,\boldsymbol{z})}$$

Applying the two constraints fixes λ_1 and λ_2, giving the contrast function stated in the main theorem. Next we consider optimizing for $w(\boldsymbol{z})$ under the constraint $E\{w(\boldsymbol{Z})\}=1$. Again, the use of Lagrange multipliers gives

$$\int w(\boldsymbol{z})^2 E\{l^2(A|\boldsymbol{Z})\sigma^2(A,\boldsymbol{Z})|\boldsymbol{Z}=\boldsymbol{z}\} - 2\lambda_3\{w(\boldsymbol{z})-1\}dP(\boldsymbol{z})$$

and differentiating the Lagrangian with respect to $w(\boldsymbol{z})$ and setting equal to zero gives

$$w(\boldsymbol{z}) = \frac{\lambda_3}{E\{l^2(A|\boldsymbol{Z})\sigma^2(A,\boldsymbol{Z})|\boldsymbol{Z}=\boldsymbol{z}\}}$$

The constant, λ_3 is fixed by the constraint, completing the proof.

Proof of Lemma 5

Let $\alpha(w)$ be a function such that $\alpha(w)=0$ for $w=0$ and $\alpha(w)=1/w$ otherwise. Now consider the expectation

$$\begin{aligned} E\left[W\left\{V\alpha(W) - g(\boldsymbol{U})\right\}^2\right] &= E\left\{V^2\alpha^2(W)W\right\} + E\left\{g^2(\boldsymbol{U})W - 2g(\boldsymbol{U})VW\alpha(W)\right\} \\ &= E\left\{V^2\alpha^2(W)W\right\} + E\left\{g^2(\boldsymbol{U})E(W|\boldsymbol{U}) - 2g(\boldsymbol{U})E\{VW\alpha(W)|\boldsymbol{U}\}\right\} \end{aligned}$$

For the purposes of minimization over $g(.)$ the first term on the right hand side can be discarded since it does not depend on $g(.)$. Hence

$$\arg\min_{g(.)} E\left[W\left\{V\alpha(W) - g(\boldsymbol{U})\right\}^2\right] = \arg\min_{g(.)} E\left\{g^2(\boldsymbol{U})E(W|\boldsymbol{U}) - 2g(\boldsymbol{U})E\{VW\alpha(W)|\boldsymbol{U}\}\right\}$$

By the calculus of variations, the minimiser $g^*(u)$ satisfies,

$$g^*(u)E(W|\boldsymbol{U}=\boldsymbol{u}) = E\{VW\alpha(W)|\boldsymbol{U}=\boldsymbol{u}\}$$

Since $W \neq 0$ almost surely,

$$\begin{aligned} E\{VW\alpha(W)|\boldsymbol{U}\} &= E\{VW\alpha(W)|W\neq 0,\boldsymbol{U}\} \\ &= E(V|W\neq 0,\boldsymbol{U}) \\ &= E(V|\boldsymbol{U}). \end{aligned}$$

Hence the result follows provided that $E(W|\boldsymbol{U}) \neq 0$ which is true since $W > 0$ almost surely.

F.1 Estimator Asymptotic Distribution

F.1.1 Estimator of ψ

Throughout we use superscript hat to denote functional estimators obtained from an independent sample, and we define,

$$\hat{\phi}_\psi(\boldsymbol{o}) = \frac{\{y-\hat{\mu}(\boldsymbol{z})\}\{a-\hat{\pi}(\boldsymbol{z})\} - \hat{\lambda}(\boldsymbol{z})\{a-\hat{\pi}(\boldsymbol{z})\}^2}{\hat{\beta}(\boldsymbol{z})} + \hat{\lambda}(\boldsymbol{z}) - \hat{\psi}_0$$

where $\hat{\psi}_0$ denotes an initial plug-in estimate of ψ. We make the following assumptions, where $||f|| \equiv E\{f^2(\boldsymbol{O})\}^{1/2}$ denotes the $L_2(P)$ norm.

(A1) The propensity score error, $||\pi - \hat{\pi}||$ is $o_P(n^{-1/4-\delta})$ for some $\delta \geq 0$.

(A2) The outcome error, $||\mu - \hat{\mu}||$ is $o_P(n^{-1/4+\delta})$.

(A3) The product of $||\lambda - \hat{\lambda}||$ and $||\beta - \hat{\beta}||$ is $o_P(n^{-1/2})$.

(A4) The variance estimates are bounded as $\hat{\beta}(\boldsymbol{z}) \geq \epsilon$ for some $\epsilon > 0$ with probability 1.

(A5) There exists a constant $K > 0$ such that each of $\hat{\lambda}(\boldsymbol{z}), ...$ has a range uniformly contained in $(-K, K)$, with probability one as $n \to \infty$.

(A6) There exists a P-Donsker class $\mathcal{G}_0$ such that $P(\hat{\phi}_\psi(.) \in \mathcal{G}_0) \to 1$.

Under these assumptions we show that the remainder term, R, in the expansion below is $o_P(n^{-1/2})$

$$\hat{\psi}_0 - \psi = -E\{\hat{\phi}_\psi(\boldsymbol{O})\} + R$$

where we highlight that the expectation is conditional on the independent functional estimators, e.g. $\hat{\lambda}(\boldsymbol{z})$ is treated as a fixed function. We then show that

$$-E\{\hat{\phi}_\psi(\boldsymbol{O})\} = n^{-1}\sum_{i=1}^n \phi_\psi(\boldsymbol{o}_i) + H_n - n^{-1}\sum_{i=1}^n \hat{\phi}_\psi(\boldsymbol{o}_i) \tag{F.1}$$

where H_n as an empirical process term, which is $o_P(n^{-1/2})$ under our assumptions. It follows therefore that for

$$\hat{\psi} = \hat{\psi}_0 + n^{-1}\sum_{i=1}^n \hat{\phi}_\psi(\boldsymbol{o}_i)$$

$$\hat{\psi} - \psi = n^{-1}\sum_{i=1}^n \phi_\psi(\boldsymbol{o}_i) + o_P(n^{-1/2})$$

Formally $\hat{\psi}$ is a one-step bias correction estimator, but it is seen to be equivalent to the estimating equations estimator in the main text.

The remainder term

To simplify notation, in this subsection we largely omit function arguments, for example $\mu = \mu(\boldsymbol{Z})$ with similar for $\hat{\mu}, \lambda, \hat{\lambda}, \pi, \hat{\pi}, \hat{\beta}, \beta$. Evaluating the remainder $R \equiv E\{\hat{\phi}_\psi(\boldsymbol{O}) + \hat{\psi}_0 - \psi\}$ gives

$$R = E\left[\frac{(Y-\hat{\mu})(A-\hat{\pi}) - \hat{\lambda}(A-\hat{\pi})^2}{\hat{\beta}} + \hat{\lambda} - \lambda\right]$$

where we have used the fact that $\psi = E[\lambda]$. Next we use the results,

$$E[(Y-\hat{\mu})(A-\hat{\pi})|\boldsymbol{Z}] = \lambda\beta + (\mu-\hat{\mu})(\pi-\hat{\pi})$$
$$E[(A-\hat{\pi})^2|\boldsymbol{Z}] = \beta + (\pi-\hat{\pi})^2$$

to obtain

$$\begin{aligned} R &= E\left[\frac{(\pi-\hat{\pi})\{\mu-\hat{\mu}-\hat{\lambda}(\pi-\hat{\pi})\} + (\lambda-\hat{\lambda})(\beta-\hat{\beta})}{\hat{\beta}}\right] \\ &\leq \left(\frac{1}{\epsilon}\right) E\left[(\pi-\hat{\pi})\{\mu-\hat{\mu}-\hat{\lambda}(\pi-\hat{\pi})\} + (\lambda-\hat{\lambda})(\beta-\hat{\beta})\right] \end{aligned}$$

where the inequality follows from (A4). Using the inequality, $(a+b)^2 \leq 2(a^2+b^2)$,

$$R^2 \leq \left(\frac{2}{\epsilon^2}\right)\left(\underbrace{E\left[(\pi-\hat{\pi})\{\mu-\hat{\mu}-\hat{\lambda}(\pi-\hat{\pi})\}\right]^2}_{\text{first remainder}} + \underbrace{E\left[(\lambda-\hat{\lambda})(\beta-\hat{\beta})\right]^2}_{\text{second remainder}}\right)$$

and we show that the two marked remainder terms above are $o_P(n^{-1})$, and hence R itself is $o_P(n^{-1/2})$. For the second remainder term, the Cauchy-Schwarz inequality gives

$$E\left[(\lambda-\hat{\lambda})(\beta-\hat{\beta})\right]^2 \leq E\left[(\lambda-\hat{\lambda})^2\right] E\left[(\beta-\hat{\beta})^2\right]$$

which is $o_P(n^{-1})$ under (A2). Similarly, for the first remainder term the Cauchy-Schwarz inequality gives,

$$\begin{aligned} E\left[(\pi-\hat{\pi})\{\mu-\hat{\mu}-\hat{\lambda}(\pi-\hat{\pi})\}\right]^2 &\leq E\left[(\pi-\hat{\pi})^2\right] E\left[\{\mu-\hat{\mu}-\hat{\lambda}(\pi-\hat{\pi})\}^2\right] \\ &\leq 2E\left[(\pi-\hat{\pi})^2\right]\left\{E\left[(\mu-\hat{\mu})^2\right] + E\left[\hat{\lambda}^2(\pi-\hat{\pi})^2\right]\right\} \\ &\leq 2E\left[(\pi-\hat{\pi})^2\right]\left\{E\left[(\mu-\hat{\mu})^2\right] + K^2 E\left[(\pi-\hat{\pi})^2\right]\right\} \end{aligned}$$

where the second inequality follows from the inequality, $(a+b)^2 \leq 2(a^2+b^2)$, and the third inequality follows by (A5). The first remainder term is therefore $o_P(n^{-1})$ under (A1) and (A2), hence R is $o_P(n^{-1/2})$

The empirical process term

In this subsection we use a common empirical processes notation, where we define linear operators P and $\mathbb{P}_n$ such that for some function $h(\boldsymbol{O})$, $P\{h(\boldsymbol{O})\} \equiv E\{h(\boldsymbol{O})\}$ and $\mathbb{P}_n\{h(\boldsymbol{O})\} \equiv n^{-1}\sum_{i=1}^n h(\boldsymbol{o}_i)$. Hence we write $E\{\hat{\phi}_\psi(\boldsymbol{O})\}$ as

$$(\mathbb{P}_n - P)\{\phi_\psi(\boldsymbol{O})\} + (\mathbb{P}_n - P)\{\hat{\phi}_\psi(\boldsymbol{O}) - \phi_\psi(\boldsymbol{O})\} - \mathbb{P}_n\{\hat{\phi}_\psi(\boldsymbol{O})\}$$

which follows from adding and subtracting $(\mathbb{P}_n - P)\{\phi_\psi(\boldsymbol{O})\}$ and $\mathbb{P}_n\{\hat{\phi}_\psi(\boldsymbol{O})\}$ to $P\{\hat{\phi}_\psi(\boldsymbol{O})\}$. This expression recovers (F.1) since the IC is mean zero, in the sense that $P\{\phi_\psi(\boldsymbol{O})\} = 0$, and we define the empirical process term

$$\begin{aligned} H_n &\equiv (\mathbb{P}_n - P)\{\hat{\phi}_\psi(\boldsymbol{O}) - \phi_\psi(\boldsymbol{O})\} \\ &= (\mathbb{P}_n - P)\{\hat{f}(\boldsymbol{O}) - f(\boldsymbol{O})\} \end{aligned}$$

where

$$\begin{aligned} \hat{f}(\boldsymbol{O}) - f(\boldsymbol{O}) = &\frac{\{Y - \hat{\mu}(\boldsymbol{Z})\}\{A - \hat{\pi}(\boldsymbol{Z})\} - \hat{\lambda}(\boldsymbol{Z})\{A - \hat{\pi}(\boldsymbol{Z})\}^2}{\hat{\beta}(\boldsymbol{Z})} + \hat{\lambda}(\boldsymbol{Z}) \\ &- \frac{\{Y - \mu(\boldsymbol{Z})\}\{A - \pi(\boldsymbol{Z})\} - \lambda(\boldsymbol{Z})\{A - \pi(\boldsymbol{Z})\}^2}{\beta(\boldsymbol{Z})} + \lambda(\boldsymbol{Z}) \end{aligned}$$

and we use the fact that $(\mathbb{P}_n - P)\{\hat{\psi}_0 - \psi\} = 0$. By e.g. Lemma 19.24 of van der Vaart (1998b), H_n is $o_P(n^{-1/2})$ provided that $||\hat{f}(\boldsymbol{O}) - f(\boldsymbol{O})|| = o_P(1)$.

F.1.2 Estimator of Ψ

We consider an estimator for $\Psi = C/D$ where

$$\begin{aligned} C &\equiv E\left[\{Y - \mu(\boldsymbol{Z})\}\{A - \pi(\boldsymbol{Z})\}\right] \\ D &\equiv E\left[\{A - \pi(\boldsymbol{Z})\}^2\right] \end{aligned}$$

Our goal is to consider the estimator $\hat{\Psi} = \hat{C}/\hat{D}$ where

$$\begin{aligned} \hat{C} &\equiv n^{-1}\sum_{i=1}^{n}\{y_i - \hat{\mu}(\boldsymbol{z}_i)\}\{a_i - \hat{\pi}(\boldsymbol{z}_i)\} \\ \hat{D} &\equiv n^{-1}\sum_{i=1}^{n}\{a_i - \hat{\pi}(\boldsymbol{z}_i)\}^2 \end{aligned}$$

We will show that, under (A1), (A2), and [Assumptions not complete]

(B1) There exists a constant $K > 0$ such that each of $\hat{\mu}(\boldsymbol{z}), ...$ has a range uniformly contained in $(-K, K)$, with probability one as $n \to \infty$.

(B2) There exists a P-Donsker class $\mathcal{G}_0$ such that $P(\hat{\phi}_c(.) \in \mathcal{G}_0) \to 1$ and $P(\hat{\phi}_d(.) \in \mathcal{G}_0) \to 1$.

then $\hat{C}$ and $\hat{D}$ are regular asymptotically linear in the sense that,

$$\begin{aligned} \hat{C} - C &= n^{-1}\sum_{i=1}^{n}\phi_c(\boldsymbol{o}_i) + o_p(n^{-1/2}) \\ \hat{D} - D &= n^{-1}\sum_{i=1}^{n}\phi_d(\boldsymbol{o}_i) + o_p(n^{-1/2}) \end{aligned}$$

where

$$\begin{aligned} \phi_c(\boldsymbol{o}) &\equiv \{y - \mu(\boldsymbol{z})\}\{a - \pi(\boldsymbol{z})\} - C \\ \phi_d(\boldsymbol{o}) &\equiv \{a - \pi(\boldsymbol{z})\}^2 - D \end{aligned}$$

denote the ICs of C and D respectively. It follows by algebraic manipulations that,

$$\sqrt{n}(\hat{\Psi} - \Psi) = \frac{D}{\hat{D}}\left[n^{-1/2}\sum_{i=1}^{n}\phi_\Psi(\boldsymbol{o}_i) + o_p(1)\right]$$

where $\phi_\Psi(\boldsymbol{o}) = \{\phi_c(\boldsymbol{o}) - \Psi\phi_d(\boldsymbol{o})\}/D$ is the IC of Ψ. Next we use Slutsky's Theorem and the fact that $\hat{D}/D$ converges to 1 in probability, to write,

$$\lim_{n\to\infty} \sqrt{n}(\hat{\Psi} - \Psi) = \lim_{n\to\infty} n^{-1/2} \sum_{i=1}^{n} \phi_\Psi(\boldsymbol{z}_i)$$

which gives the desired result due to the central limit theorem. We note that this set up is quite general when one considers estimands which are written as the ratio of two other estimands, such as Ψ in the present context. Since C and D are of the same form, we show regular asymptotically linearity of the numerator C with the result for D following as the special case when $Y = A$ almost surely.

Numerator estimator

Throughout we use superscript hat to denote functional estimators obtained from an independent sample, and we define,

$$\hat{\phi}_c(\boldsymbol{o}) \equiv \{y - \hat{\mu}(\boldsymbol{z})\}\{a - \hat{\pi}(\boldsymbol{z})\} - \hat{C}_0$$

where $\hat{C}_0$ denotes an initial plug-in estimate of C. Under (A1), (A2) and (B1) to (Bn) we show that the remainder term, R, in the expansion below is $o_P(n^{-1/2})$

$$\hat{C}_0 - C = -E\{\hat{\phi}_c(\boldsymbol{O})\} + R$$

where we highlight that the expectation is conditional on the independent functional estimators, e.g. $\hat{\pi}(\boldsymbol{z})$ is treated as a fixed function. We then show that

$$-E\{\hat{\phi}_c(\boldsymbol{O})\} = n^{-1} \sum_{i=1}^{n} \phi_c(\boldsymbol{o}_i) + H_n - n^{-1} \sum_{i=1}^{n} \hat{\phi}_c(\boldsymbol{o}_i) \tag{F.2}$$

where H_n as an empirical process term, which is $o_P(n^{-1/2})$ under our assumptions. It follows therefore that for

$$\hat{C} = \hat{C}_0 + n^{-1} \sum_{i=1}^{n} \hat{\phi}_c(\boldsymbol{o}_i)$$

$$\hat{C} - C = n^{-1} \sum_{i=1}^{n} \phi_c(\boldsymbol{o}_i) + o_P(n^{-1/2})$$

Formally $\hat{C}$ is a one-step bias correction estimator, but it is seen to be equivalent to the definition of $\hat{C}$ above.

Remainder term

To simplify notation, in this subsection we largely omit function arguments, for example $\mu = \mu(\boldsymbol{Z})$ with similar for $\hat{\mu}, \pi, \hat{\pi}$. Evaluating the remainder $R \equiv E\{\hat{\phi}_c(\boldsymbol{O}) + \hat{C}_0 - C\}$ gives

$$\begin{aligned} R &= E\left[(Y - \hat{\mu})(A - \hat{\pi}) - (Y - \mu)(A - \pi)\right] \\ &= E\left[(\mu - \hat{\mu})(\pi - \hat{\pi})\right] \end{aligned}$$

where we have used the fact that $\Psi = E\left[(Y - \mu)(A - \pi)\right]$. By the Cauchy-Schwarz inequality,

$$R^2 = E\left[(\mu - \hat{\mu})(\pi - \hat{\pi})\right]^2 \leq E\left[(\mu - \hat{\mu})^2\right] E\left[(\pi - \hat{\pi})^2\right]$$

which is $o_P(n^{-1})$ provided that the product of $||\mu - \hat{\mu}||$ and $||\pi - \hat{\pi}||$ is $o_P(n^{-1/2})$. In the special case where $Y = A$ almost surely, i.e. for the denominator estimator $\hat{D}$, we require that $||\pi - \hat{\pi}||^2 = o_P(n^{-1/2})$. Both of these conditions are satisfied under (A1) and (A2).

The empirical process term

Similar to the empirical process argument in Appendix F.1.1, we define the empirical process term

$$H_n \equiv (\mathbb{P}_n - P)\{\hat{\phi}_c(\boldsymbol{O}) - \phi_c(\boldsymbol{O})\}$$

which, by e.g. Lemma 19.24 of van der Vaart (1998b), is $o_P(n^{-1/2})$ provided that $P\left[\{\hat{\phi}_c(\boldsymbol{O}) - \phi_c(\boldsymbol{O})\}^2\right]$ converges to zero in probability.

F.2 Additional illustrated results

Table F.1: Least squares estimands applied to IWPC data, using the discrete superlearner algorithm for model fitting. Values indicate point estimates, given in INR/(mg/week), with 95% Wald confidence intervals given in parentheses. P-values represent those obtained from a Wald based test of the null hypothesis that the estimand is 0.

Estimand	Algorithm	Result
Ψ	1	1.87×10^{-3} (0.648×10^{-3},3.09×10^{-3}) p=0.003
Ψ	2	1.87×10^{-3} (0.661×10^{-3},3.08×10^{-3}) p=0.002
ψ	1A	-9.23×10^{-2} (-1.84,1.65) p=0.92
ψ	2A	-0.704 (-1.74,0.335) p=0.18
ψ	1B	1.47×10^{-3} (-0.125×10^{-3},3.06×10^{-3}) p=0.07
ψ	2B	1.51×10^{-3} (-0.136×10^{-3},3.17×10^{-3}) p=0.07

Appendix G

Supplement to causal derivative effects for continuous exposures

Moment and Cumulant Functions of the least squares intervention distribution

Consider a random variable, X with measure $dP_0(x)$, mean μ, variance σ^2, and support $[a, b]$. The least squares intervention density is,

$$\tilde{f}(x) = \int_a^b \Theta(x - x^*) \frac{\mu - x^*}{\sigma^2} dP_0(x^*)$$

where $\Theta(.)$ is the unit step function. The characteristic function of this density is

$$\begin{aligned}
\tilde{\varphi}(t) &= \int_a^b e^{itx} \tilde{f}(x) dx \\
&= \int_a^b \int_a^b e^{itx} \Theta(x - x^*) \frac{\mu - x^*}{\sigma^2} dP_0(x^*) dx \\
&= \int_a^b \frac{\mu - x^*}{\sigma^2} \left[\int_a^b e^{itx} \Theta(x - x^*) dx \right] dP_0(x^*).
\end{aligned}$$

For the part inside the square brackets,

$$\int_a^b e^{itx} \Theta(x - x^*) dx = \int_{x^*}^b e^{itx} dx = \frac{e^{itb} - e^{itx^*}}{it}$$

Hence,

$$\begin{aligned}
\tilde{\varphi}(t) &= \int_a^b \frac{\mu - x^*}{it\sigma^2} \left\{ e^{itb} - e^{itx^*} \right\} dP_0(x^*) = \int_a^b \frac{x^* - \mu}{it\sigma^2} e^{itx^*} dP_0(x^*) \\
&= \frac{-1}{t\sigma^2} \left\{ \varphi'(t) - i\mu\varphi(t) \right\} \qquad \text{(G.1)}
\end{aligned}$$

where $\varphi(t) = E(e^{itX})$ is the characteristic function of the density f, with derivative $\varphi'(t) = E(iXe^{itX})$. Similarly, if the moment function $M(t) = E(e^{tX})$ exists, then the moment function of the least squares intervention density is

$$\begin{aligned}
\tilde{M}(t) &= \frac{1}{t\sigma^2} \left\{ M'(t) - \mu M(t) \right\} = \frac{M(t)}{t\sigma^2} \left\{ \frac{M'(t)}{M(t)} - \mu \right\} \\
&= \frac{\exp\{K(t)\}}{t\sigma^2} \left\{ K'(t) - \mu \right\}
\end{aligned}$$

where $K(t) = \log\{M(t)\}$. Hence, letting $\tilde{K}(t) = \log\{\tilde{M}(t)\}$, the cumulant function of the least squares intervention density is given by (8.14), where we note that the mean and variance are the first and second cumulants of $K(t)$ respectively.

Proof of Theorem 12

To demonstrate symmetry of the ALSE transformation, we use the standard result that the characteristic function of a random variable is real if and only if the distribution of the corresponding random variable is symmetric about 0, see Feller (1966), Chapter XV. We let X be a symmetrically distributed random variable with finite mean, μ and variance, σ^2 and write the transformed variable $\tilde{X}$, which is distributed according to the least squares intervention distribution associated with X. It follows from (G.1) that,

$$E_{\tilde{P}}\{e^{it(X-\mu)}\} = \frac{-1}{t\sigma^2}\frac{d}{dt}E\{e^{it(X-\mu)}\}$$

Here $E\{e^{it(X-\mu)}\}$ is the characteristic function of the random variable $X-\mu$ which is symmetric about 0, and hence the RHS is real. The LHS is the characteristic function of the random variable $\tilde{X}-\mu$ which must also be real, and hence symmetric about 0.

Least squares intervention distribution for certain exposure distributions

The Gamma distribution with shape parameter, α and rate parameter, β, and cumulants, $\kappa_1 = \alpha/\beta$ and $\kappa_2 = \kappa_1/\beta$ and cumulant generating function,

$$K(t;\alpha,\beta) = -\alpha\log\left(1-\frac{t}{\beta}\right)$$

Therefore, by (8.14),

$$\tilde{K}(t;\alpha,\beta) = -(\alpha+1)\log\left(1-\frac{t}{\beta}\right)$$

which happens to be $K(t;\alpha+1,\beta)$. Note the Chi-squared Distribution is a special case of the Gamma distribution with, $\alpha = k/2$ and $\beta = 1/2$.

The beta distribution, with shape parameters, α, and β, has the density

$$f(x|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$$

for $x \in [0,1]$ and 0 otherwise. The mean is $\mu = \alpha/(\alpha+\beta)$ and we note that $xf(x|\alpha,\beta) = \mu f(x|\alpha+1,\beta)$, hence,

$$\tilde{f}(x|\alpha,\beta) = \int_0^x \frac{\mu-x^*}{\sigma^2}f(x^*|\alpha,\beta)dx^* = \frac{\mu}{\sigma^2}\{F(x|\alpha,\beta) - F(x|\alpha+1,\beta)\}$$

The distribution function, $F(x|\alpha,\beta)$ has the property that

$$F(x|\alpha+1,\beta) = F(x|\alpha,\beta) - \frac{\Gamma(\alpha+\beta)}{\alpha\Gamma(\alpha)\Gamma(\beta)}x^{\alpha}(1-x)^{\beta}$$

Therefore, using the fact that $\mu/\sigma^2 = (\alpha+\beta)(\alpha+\beta+1)/\beta$, we recover the result, $\tilde{f}(x|\alpha,\beta) = f(x|\alpha+1,\beta+1)$.

The beta-prime distribution, with shape parameters, α, and β, has the density

$$f(x|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1+x)^{-\alpha-\beta}$$

for $x \geq 0$. Whilst the result for the beta-prime distribution can be derived in a similar way to the beta distribution, it is sufficient to verify that

$$\frac{d}{dx} f(x|\alpha+1, \beta-2) = \frac{\mu - x}{\sigma^2} f(x|\alpha, \beta)$$

The result follows since, by (8.8),

$$\frac{d}{dx} \tilde{f}(x|\alpha, \beta) = \frac{\mu - x}{\sigma^2} f(x|\alpha, \beta)$$

Appendix H

Supplement to conclusion and outlook

H.1 Proof of Theorem 13

This proof makes use of the following Lemma (see Section H.2 for proof),

$$k_n(x+\delta) = \sum_{j=0}^{n} \binom{n}{j} k_j(x)\delta^{n-j} \tag{H.1}$$

where $\binom{n}{j}$ is a binomial coefficient. We start by considering the remainder term,

$$R_n(x) = f(x) - \sum_{i=0}^{n} \frac{E\{f^{(i)}(X)\}}{i!} k_i(x)$$

Applying the Lemma gives,

$$\begin{aligned} R_n(x+\delta) &= f(x+\delta) - \sum_{i=0}^{n} \frac{E\{f^{(i)}(X)\}}{i!} \left\{ \sum_{j=0}^{i} \binom{i}{j} k_j(x)\delta^{i-j} \right\} \\ &= f(x+\delta) - \sum_{j=0}^{n} \frac{k_j(x)}{j!} \sum_{i=j}^{n} \frac{E\{f^{(i)}(X)\}}{(i-j)!}\delta^{i-j} \\ &= f(x+\delta) - \sum_{j=0}^{n} \frac{k_j(x)}{j!} \sum_{i=0}^{n-j} \frac{E\{f^{(j+i)}(X)\}}{i!}\delta^{i} \end{aligned}$$

Since $E_P\{k_j(X)\} = 0$ for $j > 0$ and $E_P\{k_0(X)\} = 1$, it follows that

$$E\{R_n(X+\delta)\} = E\left\{ f(X+\delta) - \sum_{i=0}^{n} \frac{f^{(i)}(X)}{i!}\delta^{i} \right\}$$

By Taylor's theorem (see below),

$$E\{R_n(X+\delta)\} = E\left\{r_n(X+\delta)\right\}$$

where $r_n(.)$ is a function such that,

$$\lim_{\delta \to 0} \frac{r_n(X+\delta)}{\delta^n} = 0$$

Hence,

$$\lim_{\delta \to 0} \frac{E\{R_n(X+\delta)\}}{\delta^n} = E\left\{ \lim_{\delta \to 0} \frac{r_n(X+\delta)}{\delta^n} \right\} = 0$$

which completes the proof.

Theorem 14 (Taylor's Theorem) *Let, $f(x)$, be a function which is $n \geq 1$ times differentiable at a point x^*. Then there exists a function $r_n(x)$ such that*

$$f(x) = \sum_{j=0}^{n} \frac{f^{(j)}(x^*)}{j!}(x - x^*)^j + r_n(x)$$

and

$$\lim_{\delta \to 0} \frac{r_n(x^* + \delta)}{\delta^n} = 0$$

H.2 Proof of the Lemma in (H.1)

This Lemma is exactly that in Proposition 2.5 of Ta (2015). For completeness we provide a proof below. From the generating function of $k_n(x)$ it follows that

$$\begin{aligned}\sum_{n=0}^{\infty} k_n(x+\delta)\frac{t^n}{n!} &= \left(\frac{e^{tx}}{M_X(t)}\right)\left(e^{t\delta}\right) \\ &= \left\{\sum_{n=0}^{\infty} k_n(x)\frac{t^n}{n!}\right\}\left\{\sum_{n=0}^{\infty}\delta^n\frac{t^n}{n!}\right\}\end{aligned}$$

Applying the Cauchy product gives,

$$\begin{aligned}\sum_{n=0}^{\infty} k_n(x+\delta)\frac{t^n}{n!} &= \sum_{n=0}^{\infty}\left\{\sum_{j=0}^{n} k_j(x)\frac{t^j}{j!}\delta^{n-j}\frac{t^{n-j}}{(n-j)!}\right\} \\ &= \sum_{n=0}^{\infty}\left\{\sum_{j=0}^{n}\binom{n}{j}k_j(x)\delta^{n-j}\right\}\frac{t^n}{n!}\end{aligned}$$

Note that the order of summation can be changed due to the absolute convergence of the series. This completes the proof.

H.3 Note on expressing Appell polynomials

Letting $K_X(t) = \log M_X(t)$ denote the cumulant generating function, we write the Appell polynomial generating functions as,

$$\begin{aligned}\sum_{n=0}^{\infty}\frac{t^n}{n!}k_n(x) &= e^{tx - K_X(t)} \\ &= \exp\left(tx - \sum_{i=1}^{\infty}\kappa_i\frac{t^i}{i!}\right)\end{aligned}$$

This is of the form of the complete Bell polynomials, which are defined as the polynomials $B_n(y_1, \ldots, y_n)$ with generating function,

$$\sum_{n=0}^{\infty} B_n(y_1, \ldots, y_n)\frac{t^n}{n!} = \exp\left(\sum_{j=1}^{\infty} y_j\frac{t^n}{j!}\right)$$

These have their origin in counting set partitions and have functional forms which are well known, see e.g. 12.3.7 of Andrews (1984). The first few complete Bell polynomials are

$$\begin{aligned} B_0 &= 1 \\ B_1(y_1) &= y_1 \\ B_2(y_1, y_2) &= y_1^2 + y_2 \\ B_3(y_1, y_2, y_3) &= y_1^3 + 3y_1y_2 + y_3 \\ B_4(y_1, y_2, y_3, y_4) &= y_1^4 + 6y_1^2y_2 + 4y_1y_3 + 3y_2^2 + y_4 \end{aligned}$$