Assumption-lean inference for generalised linear model parameters

Stijn Vansteelandt

Ghent University, Ghent, Belgium, and the London School of Hygiene and Tropical Medicine, London, U.K.

E-mail: stijn.vansteelandt@ugent.be

Oliver Dukes

Ghent University, Ghent, Belgium.

[To be read before The Royal Statistical Society at the Discussion Meeting held online on Tuesday, 6 July 2021, Professor Guy Nason in the Chair]

Summary. Inference for the parameters indexing generalised linear models is routinely based on the assumption that the model is correct and a priori specified. This is unsatisfactory because the chosen model is usually the result of a data-adaptive model selection process, which may induce excess uncertainty that is not usually acknowledged. Moreover, the assumptions encoded in the chosen model rarely represent some a priori known, ground truth, making standard inferences prone to bias, but also failing to give a pure reflection of the information that is contained in the data. Inspired by developments on assumption-free inference for so-called projection parameters, we here propose novel nonparametric definitions of main effect estimands and effect modification estimands. These reduce to standard main effect and effect modification parameters in generalised linear models when these models are correctly specified, but have the advantage that they continue to capture respectively the (conditional) association between two variables, or the degree to which two variables interact in their association with outcome, even when these models are misspecified. We achieve an assumption-lean inference for these estimands on the basis of their efficient influence function under the nonparametric model while invoking flexible data-adaptive (e.g., machine learning) procedures.

Key words: bias; conditional treatment effect; estimand; influence function; interaction; model misspecification; nonparametric inference.

1. Introduction

Statistical analyses routinely invoke modelling assumptions. These include smoothness assumptions, implied by parametric or semiparametric model specifications, for instance, but also sparsity assumptions that underlie variable selection procedures. Such assumptions are generally a necessity. The curse of dimensionality indeed forces one to borrow information across strata of subjects with different covariate values, as well as to reduce the dimensions of the possibly many measured variables. Modelling assumptions are often also a deliberate choice. With a continuous exposure, for instance, one would often not be interested in knowing exactly how the outcome changes with each increase in exposure, but might content oneself with a 'simple' and parsimonious summary of the exposure effect. Models enable one to create such summaries. This distinction in the nature of the assumptions is rarely made in how we approach a data analysis, but is nonetheless an essential one that will turn out to be key to the strategy that we advocate.

Regardless of this distinction, modelling assumptions are almost always a pure mathematical convenience, and not reflecting a priori knowledge that we had prior to seeing the data. Ideally, in such cases, data analyses should therefore only extract information from the data, and not from the assumptions. This realisation is not new. It became very dominant in the 90's in work on non-ignorable incomplete data. Rotnitzky and Robins (e.g., Rotnitzky and Robins (1997); Rotnitzky et al. (1998); Scharfstein et al. (1999)), amongst others, then increased awareness that modelling assumptions, such as normality and linearity assumptions, may sometimes permit identification of parameters in the absence of missing data assumptions. There is now a fairly general agreement that such identification is dishonest when these modelling assumptions are made for convenience. In spite of this, once we have stated structural assumptions (e.g., missing data assumptions) needed for identification, we often fall back into our routine. We continue to rely on modelling assumptions more than we may realise, and treat them as representing some ground truth in how we approach inference.

For instance, likelihood-based or semiparametric estimation approaches extract information not only from the data, but also from the model as if it were known to contain the truth. In fact, maximum likelihood estimators, maximum a posteriori estimators and semiparametric efficient estimators precisely succeed to increase efficiency by taking modelling assumptions as given, and extracting information from them. This makes the resulting data analysis no longer purely evidence-based. We usually try to make up for this by adopting model or variable selection procedures. However, the inferences that are commonly provided, continue to pretend that the model delivered by these procedures was a priori given and known, which can sometimes make things worse. All of this is raising questions to what extent the data analyses that we produce are effectively (purely) evidence-based.

Motivated by these concerns, enormous progress has been made over the past several decades in terms of how to develop an inference that is 'assumption-free', across several different literatures. White (1980) developed the so-called 'sandwich estimator' of the standard error for ordinary least squares (OLS); this delivers a valid measure of uncertainty around the regression coefficient estimates, even if key model-based assumptions of OLS (linearity, homoscedasticity) are not met. Freedman (2006) noted that although the sandwich estimator is unbiased under nonlinearity, the resulting confidence intervals and tests are not useful given that it may be unclear what the model coefficients represent. Several proposals for restoring meaning to regression estimates have been made, seeing a model coefficient as a projection parameter (van der Laan and Rose, 2011; Neugebauer and van der Laan, 2007; Kennedy et al., 2019; Buja et al., 2019a,b,c), or variable importance measure (Chambaz et al., 2012), both ideas which have gained traction in high-dimensional statistics (Berk et al., 2013; Wasserman, 2014). In terms of doing causal inference, Lin (2013) gave a 'model-agnostic' approach to the adjustment for baseline covariates in randomised experiments. He noted that "one does not need to believe in the classical linear model to tolerate or even advocate OLS adjustment." Related work has explored how OLS estimates can in certain settings be interpreted as weighted averages of treatment effects, even when the linear model is wrong (Angrist and Krueger, 1999; Angrist and Pischke, 2009; Aronow and Samii, 2016; Graham and Pinto, 2018; Słoczyński, 2020). Many of the above approaches start with a common estimator of a parameter indexing a parametric regression model. They then characterise to what estimand (i.e., functional of the data distribution) the estimator converges, without assuming that the model is true. In contrast, Mark van der Laan and collaborators take an alternative approach in their scientific 'roadmap' (van der Laan and Rubin, 2006; van der Laan and Rose, 2011). They first define an estimand which characterises what we aim to infer from the data, and next develop estimation and inference based on its efficient influence function (provided the estimand is pathwise-differentiable under the nonparametric model; see Section 5), with all nuisance functionals estimated nonparametrically (e.g., via machine learning). Reliance on the efficient influence function is essential to this development, as it enables valid inference even when the analysis is based on data-adaptive procedures, such as machine learning, variable selection, model selection, etc. Attention in their work is mainly given to causal inference applications where the focus is on the average (total or (in)direct) effect of a binary, possibly time-varying treatment on a binary or continuous outcome.

Key to the latter developments is changing the starting point of the analysis from the postulation of a statistical model to the postulation of an estimand. This change of focus brings many advantages. It forces one to work with well-understood estimands that target the scientific question from the start. It enables one to separate modelling assumptions made for parsimony, which will be used to define the estimand, from assumptions imposed to handle the curse of dimensionality. It prevents reliance on these assumptions, as inference for the estimand can be developed under the nonparametric model. Finally, the resulting analysis can be pre-specified, which is essential if one aims for an honest data analysis that reflects all uncertainties, including the uncertainty surrounding the model that is used.

Changing this focus of the analysis is non-trivial, however. It turns the difficulty of postulating a model, to which we have grown to become familiar, into the difficulty of choosing an estimand, for which infinitely many choices can typically be conceived. While there is some experience in choosing meaningful estimands in causal inference applications, complications easily arise when e.g. considering continuous exposures, or when general association measures (e.g. measures of a time trend) rather than causal effect measures are of interest. It calls for the development of specific estimands that can be used quite generically (in a sense that we will make specific later) and connect to regression parameters that practitioners have grown to become familiar with. In this way, they can provide an assumption-lean inference for those standard regression parameters, which uses the underlying model only with the aim to summarise and to deliver a familiar interpretation, but relates to flexible statistical or machine learning procedures running in the background to assure valid inference. In this paper, we will show how we believe this is best done when the aim is to infer regression parameters indexing generalised linear models. In particular, we propose novel estimands for conditional association measures between two variables, and for the degree to which two variables interact in their association with outcome, which are well defined in a nonparametric sense (i.e., regardless of what is the underlying data-generating distribution). We achieve an assumption-lean inference for these estimands by deriving their efficient influence function under the nonparametric model and invoking flexible data-adaptive (e.g., parametric model selection or machine learning) procedures. Since the proposed estimands reduce to standard main effect and interaction parameters in generalised linear models when these models are correctly specified, we thus generalise standard inference for such parameters to give a pure reflection of the information that is contained in the data. Our developments thus provide a novel framework for fitting generalised linear models, and at a broader level, also shed light on what defines an adequate estimand, and how it can be constructed.

In Section 2, we illustrate the above concerns about parametric and semiparametric methods with a simple example. This is followed by proposals for novel main effect and interaction estimands in Sections 3 and 4, respectively. Nonparametric inference is developed for these estimands in Section 5, and the empirical performance of the resulting estimators is assessed in Section 6 via simulation studies. In Section 7 we apply our framework in an analysis of the effect of the First Steps program on infant birth-weight, before closing the paper with a discussion in Section 8.

2. Illustration

To clarify the points made in the introduction, we provide a simple illustration with artificial, independent data for $n = 50$ subjects on a scalar standard normal variate L, a dichotomous exposure A, coded 0 or 1, with $P(A = 1|L) = \text{expit}(L - L^2)$ and a normally distributed outcome with mean $A - L + 4.5AL + 0.5L^2 - 2.25AL^2$ and unit (residual) variance. The ordinary least squares estimator for β under model

$$E(Y|A, L) = \alpha_0 + \alpha_1 L + \beta A,$$

can be shown to converge to

$$\frac{E\left[\pi(L)\left\{1 - \tilde{\pi}(L)\right\}\left\{E(Y|A = 1, L) - E(Y|A = 0, L)\right\}\right]}{E\left[\pi(L)\left\{1 - \tilde{\pi}(L)\right\}\right]} + \frac{E\left[\left\{\pi(L) - \tilde{\pi}(L)\right\}E(Y|A = 0, L)\right]}{E\left[\pi(L)\left\{1 - \tilde{\pi}(L)\right\}\right]},$$

where $\pi(L) = P(A = 1|L)$ is the so-called propensity score and $\tilde{\pi}(L)$ denotes the population least squares projection of A onto 1 and L. This displayed 'estimand' consists of two contributions. The first is a weighted average of the contrasts $E(Y|A = 1, L) - E(Y|A = 0, L)$. It is informative about the conditional association between A and Y. The second contribution is a weighted average of the contrasts $\pi(L) - \tilde{\pi}(L)$. It is not informative about the conditional association between A and Y and is generally non-zero, except when the linear outcome model is correctly specified or $\pi(L)$ happens to be a linear function of L (see e.g. Robins et al. (1992); Vansteelandt et al. (2014)). This is disturbing. It makes the estimand targeted by the ordinary least squares estimator a questionable summary of the conditional association between A and Y, given L, when the linear model is misspecified.

A more attractive approach is based on the partially linear model

$$E(Y|A, L) = \omega(L) + \beta A, \qquad (1)$$

where β and $\omega(L)$ are unknown. Here, $\hat{\beta}$ can be obtained as the E-estimator

$$\frac{\sum_{i=1}^{n} \{A_i - \hat{\pi}(L_i)\} \{Y_i - \hat{\omega}(L_i)\}}{\sum_{i=1}^{n} \{A_i - \hat{\pi}(L_i)\} A_i}, \tag{2}$$

(Robins et al., 1992), where $\hat{\pi}(.)$ and $\hat{\omega}(.)$ are possibly data-adaptive estimators of $\pi(.)$ and $\omega(.)$, respectively. In the illustration in the next paragraph, for instance, we have based $\pi(.)$ and $\omega(.)$ on a logistic and linear additive model, respectively, using smoothing splines. The ability to use data-adaptive procedures, makes it more plausible to reason under the assumption that $\hat{\pi}(.)$ converges to $\pi(.)$, which we will make. In that case, the above estimator has been shown (Vansteelandt and Daniel, 2014) to converge to the weighted contrast

$$\frac{E\left[\pi(L)\{1-\pi(L)\}\{E(Y|A=1,L)-E(Y|A=0,L)\}\right]}{E\left[\pi(L)\{1-\pi(L)\}\right]}, \tag{3}$$

of the conditional outcome mean at $A = 1$ versus $A = 0$, even when model (1) is misspecified, e.g. because A and L interact in their association with outcome.

It follows from the above reasoning that the E-estimator, as opposed to the ordinary least squares estimator, is not crucially relying on the restrictions imposed by the outcome model: it returns a meaningful estimand that is directly informative about the conditional association between A and L, even when model (1) is misspecified. Even so, caution is warranted as the calculation of standard errors and confidence intervals may still invoke the restrictions of model (1), thereby resulting in overly optimistic inferences about the conditional association between A and Y, given L. This is indeed the case. Standard inference is based on standard errors estimated as 1 over root-n times the sample standard deviation of the so-called (estimated) influence function of $\hat{\beta}$ under model (1):

$$\frac{\{A_i - \hat{\pi}(L_i)\}\left\{Y_i - \hat{\beta}A_i - \hat{\omega}(L_i)\right\}}{n^{-1/2}\sum_{i=1}^{n}\{A_i - \hat{\pi}(L_i)\} A_i}$$

(Robins et al., 1992). This is valid when model (1) is correctly specified, but ignores that when it is misspecified, then different choices of $\pi(L)$ in (3) return estimands of a possibly different magnitude. This explains why excess variability, not expressed by the standard deviation of these influence functions, may be observed when repeated samples deliver different estimates of $\pi(L)$; Buja et al. (2019a) make a related remark that such excess variability may lead to differences between fixed- versus random-covariate designs. More formally, as we will see in Section 5, under model misspecification $\hat{\pi}(L_i)$ contributes to the first-order bias of the E-estimator. This is especially worrying when $\hat{\pi}(L_i)$ converges (in terms of root mean squared error) at a rate slower than $n^{-1/2}$. This may well be the case when smoothing splines are used, and is such that $\hat{\pi}(L_i)$ will then dominate the behaviour of the E-estimator. Accommodating this can be a daunting task when the behaviour of $\hat{\pi}(.)$ over repeated samples is ill understood (e.g., because of being based on smoothing splines).

In a simulation study under the above data-generating mechanisms, we found the empirical standard deviation of the E-estimator to be 16.7% larger than estimated, resulting in 87.3% coverage of 95% confidence intervals for (3), despite the lack of bias in $\hat{\beta}$. In contrast, the nonparametric approach that we will develop later in this article, resulted in estimators with similar bias, and empirical standard deviation of the E-estimator being only 3.0% larger than estimated (and being only 2.6% larger than that of the E-estimator), resulting in 94.9% coverage of 95% confidence intervals for (3), despite the small sample size ($n = 50$).

3. Main effect estimands

Suppose that interest lies in the association between a possibly continuous variable or exposure A and an arbitrary outcome Y, conditional on measured variables L. One logical starting point would be the generalised partially linear model

$$g\{E(Y|A, L)\} = \beta A + \omega(L), \tag{4}$$

where $g(\cdot)$ is a known link function and β and $\omega(L)$ are unknown. This model choice reflects the fact that in many regression analyses only a small subset of the parameters are of key scientific interest, and an analyst may prefer to be agnostic about the nuisance parameters. Model (4) assumes a linear association as well as the absence of A-L interactions (on the scale of the link function). It does so for reasons of parsimony, e.g. because we may want to summarise the association between A and Y into a single number, but not necessarily because it reflects the ground truth. The general question, which we will work out in this paper, is then how to develop inference for β in a way that does not rely on these assumptions.

The starting point of such analysis is to come up with an estimand that is meaningful when the above model does not hold, but reduces to β when the model holds; this then subsequently allows for nonparametric inference to be developed for that estimand. One relatively simple and generic strategy would be to define the estimand as a 'projection' of the actual data distribution onto the (semiparametric) model, such as the maximiser of the population expectation of the loglikelihood or some weighted least squares projection (e.g., van der Laan and Rose (2011); Buja et al. (2019b); Neugebauer and van der Laan (2007); Kennedy et al. (2019)). This suggestion is useful, and we will effectively build on it, but it may deliver estimands that are complicated to interpret. It is moreover vague as there will often be infinitely many such projection estimands. Indeed, each consistent estimator under the (semi)parametric model maps into a projection estimand, being defined as its probability limit under the nonparametric model.

This calls for guidance concerning the choice of estimand in practice. In our development below, we will use three criteria for choosing an estimand. Firstly, when the parametric assumptions hold, it should reduce to the target parameter of interest, in this case the parameter β indexing (4), to assure that the proposal does not hinder a familiar interpretation of the final result. Second, it should be generic, in the sense of being well defined regardless of whether A is continuous or discrete. Indeed, the fact that parametric methods can flexibly incorporate any type of regressor no doubt contributes to their continuing appeal. It should also be generic in the sense that its efficient influence function should not demand the modelling of a (conditional) density, as flexible statistical or machine learning techniques are currently not well-adapted to density estimation, and moreover, density estimators may be slowly converging. This criterion distinguishes our development from related work in the causal inference literature, where focus is usually (though not exclusively) given to binary exposures (and effect modifiers). Third, the estimand must capture what one is aiming for (e.g., a conditional association), which was not the case for ordinary least squares in Section 2. This is for instance satisfied when it equals some L-dependent weighted average of the estimand one would choose to report for a subset of individuals with given L (e.g. of the average outcome difference between subjects with $A = 1$ versus $A = 0$ and the same level of L), but is not guaranteed by all projection estimands (see the discussion section).

To distinguish assumptions aimed at parsimony from other, more substantive assumptions, let us start by assuming that the main difficulty of the problem had already been solved. Suppose in particular we already knew $E(Y|A = a, L)$ for all levels a in the support of A and all covariate levels L over the support of L. Then we would generally not be interested in reporting exactly how $E(Y|A = a, L)$ changes over a and L. We would content ourselves with a parsimonious summary of the exposure effect. At each level of L, a useful summary would be the population least squares projection of $g\left\{E(Y|A, L)\right\}$ onto A, given L. This reduces to

$$g\left\{E(Y|A = 1, L)\right\} - g\left\{E(Y|A = 0, L)\right\},$$

when A is dichotomous (coded 0 or 1). This is clearly capturing a summary of the conditional association between A and Y, given L, regardless of whether some model holds. This L-specific estimand can next be summarised across levels of L by taking a weighted average with weights given by

$$\frac{\text{Var}\left(A|L\right)}{E\left\{\text{Var}\left(A|L\right)\right\}};$$

this choice of weights will be motivated later in this section. For dichotomous A, this delivers the estimand

$$\frac{E\left(\pi(L)\left\{1 - \pi(L)\right\}\left[g\left\{E(Y|A = 1, L)\right\} - g\left\{E(Y|A = 0, L)\right\}\right]\right)}{E\left[\pi(L)\left\{1 - \pi(L)\right\}\right]}.$$

More generally, it gives rise to the estimand

$$\frac{E\left(\text{Cov}\left[A, g\left\{E(Y|A, L)\right\}|L\right]\right)}{E\left\{\text{Var}\left(A|L\right)\right\}}, \tag{5}$$

which reduces to β under model (4), but remains unambiguously defined when this model is misspecified. It will therefore enable us to do inference for β in model (4) without relying on this model restriction. Interpretation of β can still be done in the familiar way, relating to model (4), but with the additional assurance that it continues to represent a summary of the conditional association between A and Y, given L, when that model is misspecified. Such assurance is not attained for standard maximum likelihood estimators, for instance, as we saw in Section 2. We note that the interpretation of our proposed estimand may be more complicated when the L-specific estimand varies dramatically over levels of L; however, other summary measures would also then need to be interpreted with care. Summary measures remain of interest in statistics with the aim to provide insight, as they may represent all that one can realistically infer with reasonable precision in the face of the curse of dimensionality.

The estimand (5) with $g(.)$ the identity link has been studied by a number of authors, e.g. Robins et al. (2008); Newey and Robins (2018); Whitney et al. (2019). We will here extend inference for it to arbitrary link functions. Such extension is non-trivial, if one considers the major difficulties that have been experienced in drawing inference for β under the partially linear logistic model (Tchetgen Tchetgen et al., 2010; Tan, 2019), which have resulted in elegant, but complex proposals that require the modelling of the conditional density or mean of the exposure, given outcome and covariates; relying on such models is arguably less desirable when information about the conditional density of the exposure, given covariates but not outcome, is a priori available (as in randomised experiments, for instance). These complications will be avoided with our choice of estimand (5), which also reduces to β under model (4) with $g(.)$ the logit link, for which we develop nonparametric inference in Section 5. This extension is moreover important since the probability limits of popular estimators of parameters indexing non-linear models have no simple closed-form representation (unlike the case for the OLS estimator in Section 2), thus rendering their behaviour poorly understood when the model restrictions fail to hold. In particular, estimators for β based on the semiparametric efficient score under the logistic partially linear model will generally fail to converge to (5). We emphasise moreover that our estimand does not require knowing the 'true' link function under which the data was generated, since it is defined nonparametrically. Standard advice for fitting generalised linear models is that a link should be chosen that provides a scale where linearity/additivity of the effects of A and L is at least plausible. To maintain the connection between our estimand and the parameter in a semiparametric generalised linear model (4), following such advice appears reasonable, although the identity link may yield the simplest interpretation.

When the exposure is dichotomous (taking values 0 and 1), $g(.)$ is the identity link and moreover L is sufficient to adjust for confounding (in the sense that A is independent of the counterfactual outcome Y^a to exposure level a, given L), then (5) reduces to

$$\frac{E\left[\pi(L)\left\{1-\pi(L)\right\}\left(Y^1-Y^0\right)\right]}{E\left[\pi(L)\left\{1-\pi(L)\right\}\right]}. \tag{6}$$

This effect, which was also considered in Crump et al. (2006) and Vansteelandt and Daniel (2014), gives highest weight to covariate regions where both treated and untreated subjects are found. It expresses the exposure effect that would be observed in a randomised experiment where the chance of recruitment is proportional to both the probability of being treated as well as the probability of being untreated. In that case, subjects with a 10% chance of receiving treatment (or no treatment) are roughly 10 times more likely to be recruited than subjects with a 1% chance of receiving treatment (or no treatment), while subjects whose chance of receiving treatment lies between 25% and 75% are nearly equally likely to be recruited (their chance of recruitment deviates at most 33% in relative terms). Although such recruitment probabilities are not readily applied in a real-life setting, the resulting effect may well approximate that which would be found in a real-life randomised experiment, where the eligibility criteria would exclude patients who are unlikely to receive treatment or no treatment in practice. Regarding the optimality properties of this estimand, Crump et al. (2006) consider the class of weighted sample average treatment effects $\sum_{i=1}^n w(L_i)(Y_i^1 - Y_i^0)/\sum_{i=1}^n w(L_i)$ where $w(L)$ is a (known) weight. They show that, under homoscedasticity, the choice $w(L) = \pi(L)\{1-\pi(L)\}$ delivers the parameter that can be estimated with the greatest precision across the entire class.

The estimand (5) thus generalises the propensity-overlap-weighted effects to more general exposures and arbitrary link functions. Such generalisation becomes essential when the exposure is continuous, in view of the need to summarise the (now high-dimensional) exposure effect. For binary exposures, an alternative approach which prevents excessive extrapolations would be to consider overlap-weighted effects on other, non-additive scales, e.g.

$$\frac{E\left[\pi(L)\left\{1-\pi(L)\right\}Y^1\right]}{E\left[\pi(L)\left\{1-\pi(L)\right\}Y^0\right]}$$

(Vansteelandt and Daniel, 2014). Such estimands directly target marginal causal effects, as opposed to taking a weighted average of conditional causal effects. They may thus be easier to interpret than (5). However, they do not easily generalise to arbitrary exposures. Moreover, they do not generally reduce to parameters indexing a well-understood generalised linear model, making them arguably more difficult to communicate.

4. Effect modification estimands

Suppose next that interest lies in the interaction between two possibly continuous variables A_1 and A_2 in their association with a continuous outcome Y, conditional on measured variables L. One logical starting point is the generalised partially linear interaction model (Vansteelandt et al., 2008)

$$g\{E(Y|A_1, A_2, L)\} = \omega_1(A_1, L) + \omega_2(A_2, L) + \beta A_1 A_2, \tag{7}$$

where $\beta, \omega_1(A_1, L)$ and $\omega_2(A_2, L)$ are unknown. The construction of a generic estimand that reduces to β when model (7) is correctly specified, turns out to be a non-trivial task. We are not aware of existing estimands for interaction parameters that satisfy the criteria in Section 3; even if we accept estimands whose efficient influence function requires modelling a density, current proposals are limited to binary A_1 and A_2 (van der Laan and Rose, 2011).

In this paper, we propose to work with the following estimand:

$$\frac{E\left[\Pi(A_1 A_2) g\left\{E(Y|A_1, A_2, L)\right\}\right]}{E\left[\Pi(A_1 A_2)^2\right]}, \tag{8}$$

where $\Pi(.)$ is an orthogonal projection operator (w.r.t. the covariance inner product), which projects an arbitrary function of (A_1, A_2, L) onto the space of functions of (A_1, A_2, L) with mean zero, conditional on A_1, L as well as conditional on A_2, L. Such projection eliminates from $g\left\{E(Y|A_1, A_2, L)\right\}$ all main effects of A_1 and L (as well as their (additive) interactions) and all main effects of A_2 and L (as well as their (additive) interactions), thus leaving only its dependence on functions of both A_1 and A_2 (and L) that cannot be additively separated into functions of (A_1, L) or (A_2, L); such functions define additive interactions between A_1 and A_2 on the scale of the link function $g(.)$. It follows that (8) reduces to β when model (7) is correctly specified. However, a key advantage in pre-specifying such an estimand (relative to standard inference for interactions) is that it continues to capture the interaction between both exposures in their association with outcome, even when this model is misspecified. This is best understood for dichotomous exposures. From the results in Vansteelandt et al. (2008), it then follows that

$$\Pi(A_1A_2) = \frac{w(L)}{P(A_1, A_2|L)} \left\{I(A_1 = A_2) - I(A_1 \neq A_2)\right\}$$

with

$$w(L) = \left\{\frac{1}{\pi_{11}(L)} + \frac{1}{\pi_{10}(L)} + \frac{1}{\pi_{01}(L)} + \frac{1}{\pi_{00}(L)}\right\}^{-1},$$

where $\pi_{a_1a_2}(L) \equiv P(A_1 = a_1, A_2 = a_2|L)$ for $a_1, a_2 = 0, 1$. With this definition, it can now be shown that (8) reduces to a weighted average of L-conditional interactions. Indeed, for dichotomous exposures we can always write

$$\begin{aligned} g\left\{E(Y|A_1, A_2, L)\right\} &= c_0(L) + c_1(L)A_1 + c_2(L)A_2 \\ &\quad + \left\{\mu_{11}(L) + \mu_{00}(L) - \mu_{10}(L) - \mu_{01}(L)\right\} A_1A_2, \end{aligned}$$

for certain functions $c_j(L), j = 1, 2, 3$ and $\mu_{a_1a_2}(L) \equiv g\{E(Y|A_1 = a_1, A_2 = a_2, L)\}$. Here,

$$\mu_{11}(L) + \mu_{00}(L) - \mu_{10}(L) - \mu_{01}(L),$$

captures the interaction between both exposures in their association (on the scale of the link function) with outcome, at the considered level of L. This and the fact that $c_0(L)+c_1(L)A_1+c_2(L)A_2$ is orthogonal (w.r.t. the covariance inner product) to $\Pi(A_1A_2)$ implies that the estimand (8) reduces to

$$\frac{E\left[w(L)\left\{\mu_{11}(L) + \mu_{00}(L) - \mu_{10}(L) - \mu_{01}(L)\right\}\right]}{E\left\{w(L)\right\}}.$$

Here, the weights $w(L)$ naturally generalise the propensity-overlap-weights

$$\pi(L)\left\{1 - \pi(L)\right\} = \left\{\frac{1}{\pi(L)} + \frac{1}{1 - \pi(L)}\right\}^{-1},$$

to the setting of interactions between two dichotomous exposures. They assign highest weight to subjects for whom each exposure combination is sufficiently likely, so as to avoid extrapolation towards covariate strata that carry little or no information about interaction. In particular, they down-weigh those strata L in which at least one of the four possible realisations of (A_1, A_2) is unlikely to be observed. When L is sufficient to adjust for confounding for the effect of both exposures (in the sense that (A_1, A_2) is independent of the counterfactual outcome $Y^{a_1a_2}$ to exposure (a_1, a_2), given L) and $g(\cdot)$ is the identity link, then estimand (8) can also be written as

$$\frac{E\left\{w(L)\left(Y^{11} - Y^{10} - Y^{01} + Y^{00}\right)\right\}}{E\left\{w(L)\right\}}. \tag{9}$$

Consider next the special case where A_1 and A_2 are conditionally independent, given L. This is relevant in settings where A_1 or A_2 is under the control of the investigator (such that A_1 is independent of A_2 is known to hold by design); for example in summarising how the effect of a randomised treatment A_1 is modified by a continuous covariate A_2. It is moreover relevant in gene-environment interaction studies (where it is usually assumed that genetic and environmental factors are independent in the population). Then it further follows from Vansteelandt et al. (2008) that

$$\Pi(A_1A_2) = \{A_1 - E(A_1|L)\}\{A_2 - E(A_2|L)\},$$

regardless of whether the exposures are dichotomous or not. In that case, the estimand (8) can also be written as

$$\frac{E\left[\{A_1 - E(A_1|L)\}\{A_2 - E(A_2|L)\}g\{E(Y|A_1, A_2, L)\}\right]}{E\left[\{A_1 - E(A_1|L)\}^2\{A_2 - E(A_2|L)\}^2\right]}. \tag{10}$$

When A_1 and A_2 are dichotomous, this simplifies further to

$$\frac{E\left[\pi_1(L)\{1-\pi_1(L)\}\pi_2(L)\{1-\pi_2(L)\}\{\mu_{11}(L)+\mu_{00}(L)-\mu_{10}(L)-\mu_{01}(L)\}\right]}{E\left[\pi_1(L)\{1-\pi_1(L)\}\pi_2(L)\{1-\pi_2(L)\}\right]}, \tag{11}$$

where $\pi_1(L) = P(A_1 = 1|L)$ and $\pi_2(L) = P(A_2 = 1|L)$.

In the more general case, the projection operator is not obtainable in closed-form but can be obtained via the alternating conditional expectations (ACE) algorithm (Bickel et al., 1993). This involves first taking the difference U_1 between A_1A_2 and its conditional expectation given A_1 and L, next taking the difference U_2 between U_1 and its conditional expectation given A_2 and L, next taking the difference U_3 between U_2 and its conditional expectation given A_1 and L, and so on..., eventually delivering the projection U_∞. Importantly, this algorithm does not demand knowledge of the entire joint density of both exposures, conditional on L, and moreover avoids inverse weighting by such density. This is essential for enabling an inference that is generic (e.g, can be used for continuous exposures), and has been a key challenge in proposing a generic estimand such as (8).

5. Nonparametric inference

In the previous sections, we have shown how modelling assumptions can be invoked to summarise the (conditional) association between two variables, which may itself be high-dimensional, or the extent to which two variables interact in their association with an outcome. To prevent that these convenience assumptions are used as a ground truth, we next develop inference for the resulting estimands under a nonparametric model.

5.1. Main effect estimands

5.1.1. The plug-in estimator

A natural estimator of the main effect estimand β, given by (5), is

$$\frac{\sum_{i=1}^{n}\left\{A_i - \hat{E}(A_i|L_i)\right\} g^{-1}\{\hat{E}(Y_i|A_i, L_i)\}}{\sum_{i=1}^{n}\left\{A_i - \hat{E}(A_i|L_i)\right\} A_i}. \tag{12}$$

We call it a 'plug-in' estimator, as it equals the sample analogue of (5) with estimators $\hat{E}(A|L)$ and $\hat{E}(Y|A, L)$ of the unknown conditional expectations 'plugged in'. In the spirit of being 'assumption-free' (or at least, assumption-lean) it is natural to learn these conditional expectations without pre-specification of parametric models. One could therefore adopt variable/model selection procedures, or use traditional nonparametric estimators (e.g. kernel methods, sieve estimators, regression trees) or even machine learning approaches (random forests, neural networks, support vector machines) which are particularly effective when the dimension of the covariates is large. Machine learning techniques learn a (potentially very complex) 'model' from the data, whilst using regularisation (in combination with cross-validation) to minimise issues of overfitting and optimise out-of-sample predictive performance. The analyst does not need choose between different estimators now available in statistical software; ensemble learners, such as the Super Learner (Van der Laan et al., 2007), aim to take the optimal weighted combination of candidate (parametric and nonparametric) estimators.

Traditionally, statisticians have been hesitant to routinely incorporate machine learning when analysing data. This is in part because the tuning parameters used to control the degree of regularisation in the data-adaptive estimators $\hat{E}(A|L)$ and $\hat{E}(Y|A, L)$ are typically chosen to balance bias and variance in a way that is optimal for prediction purposes. Unfortunately, this choice is usually *suboptimal* for estimation of the target parameter; the 'plug-in' estimator of β given in (12) can inherit the potentially large biases from $\hat{E}(A|L)$ and $\hat{E}(Y|A, L)$. The consequence is that the bias of the naive estimator may be of the order $n^{-1/2}$ or larger, and hence the use of standard confidence intervals is not justified. A further issue is that even if parametric-rate confidence intervals could be constructed, it is unclear how one would acount for the uncertainty in the estimation of the nuisance parameters, given that these may follow a complex distribution.

5.1.2. The efficient influence function

To overcome the problems associated with plug-in estimators, we will develop inference for the estimand under a nonparametric model based on its so-called efficient influence function (Pfanzagl, 1990; Bickel et al., 1993). Technically, this is mean zero functional of the observed data and the data-generating distribution, which characterises the estimand's sensitivity to arbitrary (smooth) changes in the data-generating law. The efficient influence function for the proposed estimand is given below.

THEOREM 1. *Under the nonparametric model, the main effect estimand β, defined* *by (5), has efficient influence function*

$$\frac{\{A - E(A|L)\}\left[\mu(Y, A, L) - \beta\{A - E(A|L)\}\right]}{E\left[\{A - E(A|L)\}^2\right]} \tag{13}$$

where $g'(x) = \partial g(x)/\partial x$ *and*

$$\mu(Y, A, L) = g'\{E(Y|A, L)\}\{Y - E(Y|A, L)\} + g\{E(Y|A, L)\} - E[g\{E(Y|A, L)\}|L].$$

The proof of this and all other results is given in Section 1 of the Supplemental Materials.

If the conditional expectations indexing the efficient influence function were known, then it would follow from its mean zero property that a consistent estimator $\widetilde{\beta}$ of β could be obtained as the value of β that makes the sample average of the influence functions zero. The resulting estimator's asymptotic distribution would be governed by this influence function in the sense that

$$\sqrt{n}\left(\widetilde{\beta} - \beta\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\{A_i - E(A|L_i)\}\left[\mu(Y_i, A_i, L_i) - \beta\{A_i - E(A|L_i)\}\right]}{E\left[\{A - E(A|L)\}^2\right]} + o_p(1). \tag{14}$$

The fact that the difference between the estimator and the truth can be approximated by the sample average of a mean-zero random variable implies that $\widetilde{\beta}$ is asymptotically normally distributed with bias that shrinks to zero faster than the standard error, and with a variance that can be estimated as the sample variance of the efficient influence function (where population expectations and the value of β can be substituted by consistent estimates).

The fact that the efficient influence function involves unknown conditional expectations, makes the estimator $\widetilde{\beta}$ suggested in the previous paragraph infeasible. As in the previous section, we will therefore substitute these by consistent estimators and denote the resulting estimator $\hat{\beta}$. The power of basing the estimator on the efficient influence function is that it behaves the same asymptotically whether it be based on known conditional expectations or consistent estimators thereof, provided that these converge sufficiently fast in a relatively weak sense (made specific in the following theorem). Throughout this section, $\mathbb{P}_n$ denotes the empirical measure (i.e., sample average) and for a function $f(O)$ of the data O we use the notation $\mathbb{P}\{f(O)\} = \int f(O)\mathbb{P}(O)dO$ where $\mathbb{P}(O)$ denotes the density of the data; for an estimator $\hat{f}$, $\mathbb{P}\{\hat{f}(O)\}$ averages over O but not $\hat{f}$.

Theorem 2. *Let $\hat{\beta}$ refer to the proposed estimator of β based on estimators $\hat{E}(A|L)$ and $\hat{\mu}(Y, A, L)$ which are consistent for $E(A|L)$ and $\mu(Y, A, L)$, respectively (see details in the Appendix). Suppose that the weak positivity assumptions at both the population and sample level hold that $\mathbb{P}\left[\{A - E(A|L)\}^2\right] > \sigma$, $\mathbb{P}_n\left[\{A - E(A|L)\}^2\right] > \sigma$ and $\mathbb{P}_n\left[\left\{A - \hat{E}(A|L)\right\}^2\right] > \sigma$ for some $\sigma > 0$. Suppose furthermore that at least one of the following two conditions hold:*

(a) (Sample-splitting) $\hat{E}(A|L)$ and $\hat{\mu}(Y, A, L)$ are obtained from a sample independent from the one used to construct $\hat{\beta}$.

(b) (Donsker condition) The quantity

$$\frac{\left\{A-\hat{E}(A|L)\right\}\left[\hat{\mu}(Y,A,L)-\hat{\beta}\left\{A-\hat{E}(A|L)\right\}\right]}{\mathbb{P}_n\left[\left\{A-\hat{E}(A|L)\right\}^2\right]}$$

falls within a $\mathbb{P}$-Donsker class with probability approaching 1.

Finally, assume that $A-\hat{E}(A|L)=O_p(1)$ and that sufficient rates of convergence are attained so that the following terms are $o_p(n^{-1/2})$:

$$\mathbb{P}\left[\left\{E(Y|A,L)-\hat{E}(Y|A,L)\right\}^2\right],$$

$$\mathbb{P}\left[\left\{E(A|L)-\hat{E}(A|L)\right\}^2\right],$$

$$\mathbb{P}\left[\left\{E(A|L)-\hat{E}(A|L)\right\}^2\right]^{1/2}\mathbb{P}\left\{\left(E\left[g\left\{E(Y|A,L)\right\}|L\right]-\hat{E}\left[g\left\{\hat{E}(Y|A,L)\right\}|L\right]\right)^2\right\}^{1/2},$$

Then it follows that (14) holds with $\hat{\beta}$ in lieu of $\widetilde{\beta}$.

A detailed discussion of the above assumptions is saved for later on in this section. A consequence of this result is that the variance of $\hat{\beta}$ can be estimated as previously suggested, namely as 1 over n times the sample variance of the efficient influence functions, as if these conditional expectations were given. It implies in particular that the uncertainty that the estimators of the conditional expectations add to the analysis can be ignored when drawing inference about β, even when these are based on variable selection or machine learning procedures, whose uncertainty is difficult to quantify.

We can thus obtain an estimator and confidence interval via the simple recipe below:

(a) Obtain estimates $\hat{E}(A|L)$ and $\hat{E}(Y|A,L)$, e.g. using machine learning.

(b) If A is binary, estimate $E[g\{E(Y|A,L)\}|L]$ as

$$\hat{E}[g\{\hat{E}(Y|A,L)\}|L]=g\{\hat{E}(Y|A=1,L)\}\hat{E}(A|L)+g\{\hat{E}(Y|A=0,L)\}\{1-\hat{E}(A|L)\}$$

otherwise, use an additional data-adaptive fit (with $g\{\hat{E}(Y|A,L)\}$ as outcome).

(c) Fit a linear regression of

$$\begin{aligned}\hat{\mu}(Y,A,L) &= g^{-1}\{\hat{E}(Y|A,L)\}\{Y-\hat{E}(Y|A,L)\}\\ &\quad +g\{\hat{E}(Y|A,L)\}-\hat{E}[g\{\hat{E}(Y|A,L)\}|L]\end{aligned}$$

on the sole predictor $A-\hat{E}(A|L)$ (without an intercept) using OLS in order to obtain an estimate $\hat{\beta}$ of β.

The variance of $\hat{\beta}$ can be consistently estimated as

$$\hat{V}(\hat{\beta})=\frac{n^{-2}\sum_{i=1}^n\left\{A_i-\hat{E}(A_i|L_i)\right\}^2\left[\hat{\mu}(Y_i,A_i,L_i)-\hat{\beta}\left\{A_i-\hat{E}(A_i|L_i)\right\}\right]^2}{\left[n^{-1}\sum_{i=1}^n\left\{A_i-\hat{E}(A_i|L_i)\right\}^2\right]^2}.$$

It is readily obtained by requesting that the software provide a sandwich estimator in step (c). A confidence interval can be constructed as $\hat{\beta} \pm 1.96\sqrt{\hat{V}(\hat{\beta})}$.

The rate conditions required in Theorem 2 will hold if all nuisance parameters are consistently estimated at a rate faster than $n^{1/4}$; under certain smoothness/sparsity assumptions, these are attainable for many data adaptive methods (see Chernozhukov et al. (2018) for a summary). The Donsker condition, which restricts the complexity of the estimators involved, is unlikely to be satisfied for very flexible machine learning methods. A simple solution is to use sample-splitting; split the data in half, estimate the nuisance parameters in the 'training' split, and perform inference on β in the 'validation' sample. This has a disadvantage of halving the sample size. However, efficiency can be asymptotically recovered via cross-fitting (Zheng and van der Laan, 2011; Chernozhukov et al., 2018); e.g. one can reverse the training and validation samples, construct a second estimate of β and average the pair. Confidence intervals can be constructed by combining the estimated influence functions across the different splits, replacing β with the averaged rather than split-specific estimate. As before, one can then estimate the variance of cross-fit estimator of β as 1 over n times the sample variance of the (estimated) influence functions.

The combination of efficient influence function-based estimators with cross-fitting facilitates the use of machine learning to estimate parts of the data distribution of no scientific interest. These important results have only been highlighted relatively recently (Zheng and van der Laan, 2011; Chernozhukov et al., 2018), and many open questions remain. Firstly, there is yet to be firm guidance on the number of splits to use in the cross-fitting. Moreover, since the machine learning methods typically perform better with more data, it may be that no splitting can sometimes yield estimators of the target parameter with smaller bias (though potentially more biased standard errors) compared to cross-fitting. At the other extreme, due to the similarity of our estimator to that of Robinson (1988), it may be possible to obtain much sharper results on the nuisance estimators by using a more specific variant of cross-fitting in combination with so-called 'undersmoothing' (Newey and Robins, 2018). This is left to future work. For now, if cross-fitting is adopted, we recommend 10-fold cross-fitting, each time using nine tenths as training sample and the remainder as validation sample.

5.1.3. Illustration - inference under the partially linear model

We return to the case study in Section 2. So long as the partially linear model (1) holds, it turns out that there are several different ways of constructing estimators of β that are desensitised to 'plug-in' bias of machine learners. Chernozhukov et al. (2018) propose using either the E-estimator (2) described in Section 2 or the 'partialling out' estimator

$$\frac{\sum_{i=1}^{n}\left\{A_i - \hat{E}(A_i|L_i)\right\}\{Y_i - \hat{E}(Y_i|L_i)\}}{\sum_{i=1}^{n}\left\{A_i - \hat{E}(A_i|L_i)\right\}^2}, \tag{15}$$

(Robinson, 1988), where all nuisance parameters are estimated via machine learning. So long as the semiparametric model restriction holds, under standard conditions both estimation approaches discussed in the previous paragraph are first-order equivalent. The E-estimator has an influence function

$$\frac{\{A - E(A|L)\}\{Y - \beta A - E(Y|A=0,L)\}}{E\left[\{A - E(A|L)\}^2\right]}$$

that coincides with the influence function

$$\frac{\{A - E(A|L)\}\left[Y - E(Y|L) - \beta\{A - E(A|L)\}\right]}{E\left[\{A - E(A|L)\}^2\right]}$$

for the 'partialling out' estimator when model (1) holds. The latter reduces to (13) for the identity link, given in Theorem 1; indeed, the estimators proposed in the previous subsection generalise the 'partialling out' estimator to arbitrary link functions. Further, assuming the residual variance of Y conditional on A and L is a constant σ^2, both estimators have an asymptotic variance equal to the semiparametric efficiency bound $\sigma^2/E\{\text{var}(A|L)\}$. The asymptotic bias of both approaches depends in part on the product of two errors - either

$$\{E(A|L) - \hat{E}(A|L)\}\{E(Y|A=0,L) - \hat{E}(Y|A=0,L)\}$$

for the E-estimator or

$$\{E(A|L) - \hat{E}(A|L)\}[E(Y|L) - \hat{E}(Y|L) - \beta\{E(A|L) - \hat{E}(A|L)\}] \qquad (16)$$

for the 'partialling out' estimator. As long as each estimator converges to the truth, then the product of two errors will tend to shrink at least as fast (and usually much faster) than an individual error.

However, the situation is quite different when restriction (1) fails (Whitney et al., 2019). The asymptotic bias of the E-estimator, relative to estimand (5), is now proportional to

$$E\left[\left\{E(A|L) - \tilde{E}(A|L)\right\}\left\{E(Y - \beta A|L) - \tilde{E}(Y|A=0,L)\right\}\right],$$

where $\tilde{E}(A|L)$ is the probability limit of $\hat{E}(A|L)$ and $\tilde{E}(Y|A=0,L)$ is the probability limit of $\hat{E}(Y|A=0,L)$; note that $E(Y|A=0,L) = E(Y - \beta A|L)$ under the partially linear model but not otherwise. Because the error $E(Y - \beta A|L) - \hat{E}(Y|A=0,L)$ will no longer shrink to zero, the bias of the E-estimator will be determined by $E(A|L) - \hat{E}(A|L)$. As discussed above, the situation may be much worse for semiparametric estimators in nonlinear models, since the bias w.r.t (5) may now even diverge. By considering (16), it follows that the same issues are not true for the 'partialling out' estimator, which makes the sample average of the influence functions for the estimand (5) evaluated at the machine learning predictions equal to zero. This highlights the benefits of estimation using the influence function obtained under a nonparametric model; it incorporates an implicit bias-correction, as the bias of the estimator of the target parameter is usually smaller in magnitude than that of the first stage estimators. Moreover, this property is not dependent on any semiparametric modelling assumptions.

Note also that when model (1) is misspecified, each change of $\pi(L)$ also changes the estimand targeted by the E-estimator. In particular, different estimates of the propensity score may then be viewed as targeting different effect estimands. The resulting excess variability is not acknowledged when basing inference on the influence function of the E-estimator, as this is assuming model (1) to be correctly specified. This was indeed what was observed in the simulation study described in Section 2. As Buja et al. (2019c) note, for certain choices of nuisance parameter estimators (specifically, series methods or twicing kernels) the E-estimator and the proposed influence function-based estimator can exactly coincide. However, since we wish to work in greater generality, and in the following section consider arbitrary machine learners for the nuisances, we do not consider this subtlety any further.

5.2. Effect modification estimands

The following theorem gives the efficient influence function for the effect modification estimand β, given by (8), under the nonparametric model.

Theorem 3. *Under the nonparametric model, the effect modification estimand β, defined by (11), has efficient influence function*

$$\frac{\Pi(A_1A_2)}{E\left\{\Pi(A_1A_2)^2\right\}}\left\{\mu(Y,A_1,A_2,L)-\beta\Pi\left(A_1A_2\right)\right\}$$

where

$$\mu(Y,A_1,A_2,L)\equiv g'\left\{E(Y|A_1,A_2,L)\right\}\left\{Y-E(Y|A_1,A_2,L)\right\}+\Pi\left[g\left\{E(Y|A_1,A_2,L)\right\}\right].$$

A root-n consistent estimator of β can thus be obtained as

$$\hat{\beta}=\left\{\frac{1}{n}\sum_{i=1}^{n}\hat{\Pi}^2(A_{i1}A_{i2})\right\}^{-1}\frac{1}{n}\sum_{i=1}^{n}\hat{\Pi}(A_{i1}A_{i2})\hat{\mu}(Y_i,A_{i1},A_{i2},L_i),$$

where

$$\hat{\mu}(Y,A_1,A_2,L)=g'\left\{\hat{E}(Y|A_1,A_2,L)\right\}\left\{Y-\hat{E}(Y|A_1,A_2,L)\right\}+\hat{\Pi}\left[g\left\{\hat{E}(Y|A_1,A_2,L)\right\}\right].$$

Here, $\hat{E}(Y|A_1,A_2,L_i)$ denotes a data-adaptive prediction (e.g., obtained using machine learning or a flexible parametric model building procedure). Further, the projection $\hat{\Pi}(A_{i1}A_{i2})$ can be obtained via the alternating conditional expectations (ACE) algorithm (Bickel et al., 1993). This involves first data-adaptively predicting $A_{i1}A_{i2}$ on the basis of A_{i1} and L_i and taking the residuals; next, data-adaptively predict these residuals on the basis of A_{i2} and L_i and take the residuals; next, data-adaptively predict these residuals on the basis of A_{i1} and L_i and take the residuals; and so forth. This process can be aborted when the variance of the predicted residuals reaches a value very close to zero. To ensure a decreasing variance, we recommend in each step tuning the obtained predictions of the residuals by substituting these by the ordinary least squares prediction of those residuals onto the obtained data-adaptive predictions. The projection $\hat{\Pi}\left\{\hat{E}\left(Y|A_{i1},A_{i2},L_i\right)\right\}$ is likewise obtained, starting from $\hat{E}\left(Y|A_{i1},A_{i2},L_i\right)$. The following theorem outlines the necessary conditions on the nuisance parameters, in order to obtain valid inference.

Theorem 4. *Suppose that estimators $\hat{\Pi}(.)$ and $\hat{E}(Y|A_1, A_2, L)$ are consistent for $\Pi(.)$ and $E(Y|A_1, A_2, L)$, respectively (see details in the Appendix). Suppose that the weak positivity assumptions at both the population and sample level hold that $\mathbb{P}\left\{\Pi\left(A_1A_2\right)^2\right\} > \sigma$, $\mathbb{P}_n\left\{\Pi\left(A_1A_2\right)^2\right\} > \sigma$ and $\mathbb{P}_n\left\{\hat{\Pi}\left(A_1A_2\right)^2\right\} > \sigma$ for some $\sigma > 0$. Suppose furthermore that at least one of the following two conditions hold:*

(a) (Sample-splitting) $\hat{\Pi}(A_1A_2)$, $\hat{E}(Y|A_1, A_2, L)$ and $\hat{\Pi}[g\{\hat{E}(Y|A_1, A_2, L)\}]$ are obtained from a sample independent to the one used to construct $\hat{\beta}$.

(b) (Donsker condition) The quantity

$$\frac{\hat{\Pi}(A_1A_2)}{\mathbb{P}_n\left\{\hat{\Pi}(A_1A_2)^2\right\}}\left\{\hat{\mu}(Y, A_1, A_2, L) - \hat{\beta}\hat{\Pi}\left(A_1A_2\right)\right\}$$

falls within a $\mathbb{P}$-Donsker class with probability approaching 1.

Further, assume that $\hat{\Pi}(A_1A_2) = O_p(1)$ and that sufficient rates of convergence are attained so that the following terms are $o_p(n^{-1/2})$:

$$\mathbb{P}\left[\left\{E(Y|A, L) - \hat{E}(Y|A, L)\right\}^2\right],$$

$$\mathbb{P}\left[\left\{\Pi(A_1A_2) - \hat{\Pi}(A_1A_2)\right\}^2\right],$$

$$\mathbb{P}\left(\left\{\Pi(A_1A_2) - \hat{\Pi}(A_1A_2)\right\}^2\right)^{1/2}\mathbb{P}\left(\left[\Pi\left[g\left\{E(Y|A, L)\right\}\right] - \hat{\Pi}\left[g\left\{\hat{E}(Y|A, L)\right\}\right]\right]^2\right)^{1/2},$$

where, for a random variable V, $\hat{\Pi}^*(V) \equiv V - \hat{\Pi}(V)$. Then it follows that

$$\sqrt{n}(\hat{\beta} - \beta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\Pi(A_{i1}A_{i2})}{E\left\{\Pi(A_{i1}A_{i2})^2\right\}}\left\{\mu(Y_i, A_{i1}, A_{i2}, L_i) - \beta\Pi\left(A_{i1}A_{i2}\right)\right\} + o_p(1).$$

The variance of both considered estimators is obtained as 1 over n times the variance of the corresponding influence function, with conditional expectations substituted by data-adaptive predictions, marginal expectations by sample averages and β by $\hat{\beta}$.

6. Simulation studies

To provide insight into different aspects of the proposal, we provide results on 4 sets of simulation experiments. In all experiments, we report Monte Carlo bias and standard deviation (SD), as well as standard errors (SE) estimated as 1 over root-n times the sample standard deviation of the estimated influence functions and coverage of corresponding 95% Wald confidence intervals. Throughout, we will refer to the proposal as 'AL' for 'Assumption-Lean'.

6.1. Main effects, binary exposure

In the first experiment, we study inference for the main effect estimand (5) with $g(.)$ the logit link. The aim of this experiment is to contrast our proposal based on random forest regression with 3 competing estimators. In particular, we considered the maximum likelihood estimator (MLE) of β obtained by fitting the logistic regression model $\text{logit}\{E(Y|A,L)\} = \beta A + \alpha_0 + \alpha_1^T L$. We also included two estimators designed for the partially linear logistic model $\text{logit}\{E(Y|A,L)\} = \beta A + \omega(L)$; the first estimator 'ES' solves the semiparametric efficient score equations e.g. in Kosorok (2007):

$$0 = \sum_{i=1}^{n} \left(A_i - \frac{\hat{E}\left[A_i \hat{E}(Y_i|A_i,L_i)\{1-\hat{E}(Y_i|A_i,L_i)\}\middle|L_i\right]}{\hat{E}\left[\hat{E}(Y_i|A_i,L_i)\{1-\hat{E}(Y_i|A_i,L_i)\}\middle|L_i\right]} \right) \\ \times \left(Y_i - \text{expit}\left[\beta A_i + \text{logit}\{\hat{E}(Y_i|A_i=0,L_i)\}\right]\right)$$

whereas the second is the simple doubly robust (DR) estimator proposed in Tchetgen Tchetgen (2013), which solves the equations

$$0 = \sum_{i=1}^{n} \left\{ A_i - \hat{E}(A_i|Y_i=0,L_i) \right\} \left\{ Y_i - \hat{E}(Y|A_i=0,L_i) \right\} \exp(-\beta A_i Y_i).$$

For this purpose, we generated a 10-dimensional covariate $L \sim N(0,\Sigma)$, where Σ was (once) randomly generated with variances between 2 and 10 and correlations up to 0.72 in absolute value and then fixed across simulations; and $A \sim \text{Bern}(\gamma^T L - 0.15 L_1^2)$, where γ is the 10-dimensional unit vector scaled by $1/\sqrt{40}$ and L_k is the kth entry of L. For generating the outcome Y, we considered 4 separate settings: 1) $Y \sim \text{Bern}(\text{expit}(0.3A + \delta^T L_{[1:5]}))$ where δ is a 5-dimensional unit vector scaled by 1/10; 2) $Y \sim \text{Bern}(\text{expit}(0.3A + \delta^T L_{[1:5]} + 0.1L_1^2))$; 3) $Y \sim \text{Bern}(\text{expit}(L_1(1.5A-1) + \delta^T L_{[1:5]}))$; and 4) $Y \sim \text{Bern}(\text{expit}(0.1/(1+\exp(0.1L_3 - 0.1L_2)) + 0.3A/(1+\exp(-0.1L_2)) + 0.5AL_6 + 0.025L_1^2))$. Only in the first two settings does the partially linear logistic model restriction hold; the fourth setting is especially challenging, in light of the complex functional form of the interaction between A and L.

For the ES estimator as well as the proposal, random forests (via the 'grf' package described in Athey et al. (2019)) were used to learn $E(Y|A,L)$ and $E(A|L)$ and yield predictions. These could then be plugged into the relevant estimating equations via application of the law of total probability. For the DR estimator, random forests were used to learn $E(A|Y,L)$ so that predictions of $E(A|Y=0,L)$ could be obtained, as this reflects how this conditional expectation would likely be estimated in practice using machine learning. In experiments 3 and 4, in order not to privilege the proposed estimand, we chose to report bias and coverage relative to the population limit of the estimator. The latter was approximated by generating 500 datasets with sample size 100,000, using the true conditional expectations for the nuisance functionals where possible, and taking the average of the resulting estimates.

In Table 1, we see that the MLE does not always target an estimand that summarises the conditional association of scientific interest. This is confirmed in experiment 2, where the limit of the MLE is in a different direction to the parameter in the partially linear model, which is especially worrisome. We see that the two semiparametric approaches perform well when the model restriction holds. In experiment 3 and 4 however we see that outside of the model, coverage can sharply decrease as sample size increases. This is the result of excess variability not reflected by standard errors when the model is misspecified. This is particularly the case for the DR estimator, where the bias inherited from the random forests appears to be substantial. Our proposal had better coverage than competing approaches in experiment 3 across the different sample sizes; this is due to both lower bias, and estimated standard errors that at least in large samples more accurately reflect the variability of the estimator. Reassuringly, despite our inferences being assumption-lean, the empirical standard deviations show that this does not come with a loss of precision.

Table 1 about here.

6.2. Effect modification, binary exposure

In a second set of simulation experiments, we considered inference for effect modification estimand (8), with $g(.)$ the identity link and without making the assumption of conditionally independent exposures. We generated a 10-dimensional covariate $L \sim N(0, \Sigma)$, where Σ was (once) randomly generated as before. The exposure was generated as in the previous section, and the outcome as $Y \sim N(3/(1+\exp(L_3-L_2))+A/(1+\exp(L_1-L_2)), 1)$. This data-generating mechanism is inspired by Nie and Wager (2017), but made more complicated by means of a non-randomised exposure A.

Our aim was to assess evidence for modification of the effect of A by L_3. Since such effect modification is absent, we here studied the performance of different estimation methods w.r.t. their ability to retrieve zero effect modification (thus also giving us a different perspective than in the previous section, where we contrasted each estimator with its limit in experiments 3 and 4). The simulation results in Table 2 demonstrate favourable results for the proposal, based on random forests (via the 'grf' package described in Athey et al. (2019)) as compared to OLS based on a linear model that includes all main effects along with the interaction between A and L_3. In particular, we observe smaller bias and better coverage at the expense of an increase in standard errors (around 30% larger).

In a second set of simulation experiments, we made the data-generating mechanism even more challenging by changing the outcome model to $Y \sim N(3/(1+\exp(L_3-L_2))+A/(1+\exp(L_1-L_2))+5AL_6, 1)$. The inclusion of an interaction between A and L_6 now makes it increasingly difficult to demonstrate the absence of effect modification between A and L_3 (which has a correlation of -0.54 with L_6). The simulation results demonstrate drastically favourable results for the proposal with a much smaller bias as well as standard errors (up to 4 times smaller than for OLS), resulting in much better coverage.

To demonstrate the behaviour under conditions where the linear regression model is correctly specified, we additionally generated a continuous exposure $A \sim N(\gamma^T L, 1)$, where γ is the d-dimensional unit vector scaled by $1/\sqrt{40}$, and the outcome as $Y \sim N(\gamma^T L + 5AL_3, 1)$. Both methods give good performance in this setting, with the proposal not surprisingly delivering larger standard errors (roughly up to 2.5 times larger) in view of the poorer ability of random forest regression to pick up linear trends. Here, better performance can be expected with the use of ensemble learners.

Table 2 about here.

6.3. High-dimensional variable selection, continuous exposure

In a third set of simulation experiments, we considered inference for the main effect estimand (5) with $g(.)$ the logit link in the presence of high-dimensional covariates using the data-generating mechanism in Belloni et al. (2013). In particular, we generated a 250-dimensional covariate $L \sim N(0, \Sigma)$, where Σ is an autoregressive correlation matrix with correlation parameter 0.5. The exposure was normally distributed with mean given by $\sum_{j=1}^{10} L_j/j$ and unit residual variance. The outcome was dichotomous with mean given by expit $\left[0.2A + \sum_{j=1}^{5} L_j/(2j) + \sum_{j=11}^{15} L_j/\{2(j-10)\}\right]$.

We evaluated the performance of the standard lasso and ϵ-net estimators under a main effect logistic regression model, as well as the post-lasso (P-lasso) estimator obtained by refitting that model using the selected variables. In each case, the penalty was chosen as the largest value for which the cross-validated prediction error is within 1 standard error of the minimum. We moreover evaluated the proposed assumption-lean procedure (AL) based on these fitting strategies for both the outcome and exposure, assuming that these obey main effect logistic and linear models, respectively. We finally also included a plug-in estimator (SL) and the proposed estimator (AL SL) based on SuperLearner fits for $E(Y|A, L)$, for $E\left[g\left\{E(Y|A, L)\right\}\right]$ and for $E(A|L)$. The SuperLearner library included two lasso procedures and two ϵ-net procedures, using penalties equal to either the above suggested penalty or the one that minimises the cross-validated prediction error. It additionally included a screening procedure based on running lasso on only the variables selected in a first lasso run.

Table 3 shows that the post-lasso estimator was heavily biased with downwardly biased standard error estimators (given by the default model-based standard errors) as a result of ignoring variable selection uncertainty. The proposal based on post-lasso reduced bias in the estimator, as well its variability, but did not result in a convincing improvement in standard errors. Much better results were found with standard use of the lasso and ϵ-net, where the proposal was able to remove bias completely, while also reducing variability further relative to the use of post-lasso. It moreover provided unbiased standard error estimators (which are not available for standard lasso and ϵ-net procedures in view of the complex distribution of the estimators they return), leading to nominal coverage of the Wald confidence intervals being attained. The standard lasso and ϵ-net estimators were less variable, but this is largely due to shrinkage bias, with coefficients often being set to zero. The use of SuperLearner worsened performance. While less bias was observed with the plug-in estimator, standard errors were very poorly estimated resulting in poor coverage of confidence intervals. Results indicate that larger sample sizes are needed for the proposed estimator based on SuperLearner to perform well in settings with such high-dimensional covariates. In the Supplementary Materials, we provide additional simulation results under complex data-generating mechanisms with misspecified link function, in which we also study the performance of cross-fitting.

Table 3 about here.

6.4. Complementary log-log link function

In a final set of simulations, we considered the same main effect estimand (5) with a logit link as in Section 6.1, but now included a complementary log-log link in the data-generating model. The exposure A and L were generated as in Section 6.1 and Y was generated in 4 different ways: 1) $Y \sim \text{Bern}\bigl(1-\exp\bigl(-\exp(0.3A+\delta^T L_{[1:5]})\bigr)\bigr]$ where δ is a 5-dimensional unit vector scaled by 1/10; 2) $Y \sim \text{Bern}\bigl(1-\exp\bigl(-\exp(0.3A+\delta^T L_{[1:5]} -0.025L_1^2)\bigr)\bigr)$; 3) $Y \sim \text{Bern}\bigl(1-\exp\bigl(-\exp(0.1L_1A+\delta^T L_{[1:5]})\bigr)\bigr)$; and 4) $Y \sim \text{Bern}\bigl(1-\exp\bigl(-\exp\bigl(0.025/\left(1+\exp(0.1L_3-0.1L_2)\right)+0.075A/\left(1+\exp(-0.1L_2)\right) +0.125AL_6-0.025L_1^2\bigr)\bigr)\bigr)$.

For each setting, we fitted a generalised linear model with a complementary log-log link function, that was linear (on the complementary log-log scale) in A and the covariates L; the maximum likelihood estimator of the main effect of A is referred to as 'MLE-cloglog'. Only in the first setting was this model correctly specified; the maximum likelihood estimator is inconsistent for the parameter β indexing a (correctly specified) partially linear complementary log-log model $\text{cloglog}\{E(Y|A,L)\} = \beta A + \omega(L)$ in the second setting. In the third and fourth, this estimator converges to a population limit which was approximated via simulation, and which may not be easily interpretable.

We also implemented the same estimator from Section 6.1, developed for the estimand (5) with a logit link. We emphasise that although the link function for generating the data was different to the one used in the considered estimand, nevertheless the estimand remains well defined. For comparison, we also considered the maximum likelihood estimator of the main effect for A in a logistic regression model that was linear (on the logit scale) in A and L (MLE-logit). The logistic model was misspecified in each of the experiments, so bias and coverage of the maximum likelihood estimator were again reported relative to the estimator's population limit (approximated via simulation). From the results in Table 4, one can see that even when the complementary log-log link was used in the data-generating model, our estimator continues to infer a weighted average of the conditional association of interest (on the log-odds scale) with relatively low bias, and with confidence intervals that possess close to their advertised coverage. This is reassuring, given that often data analysts may prefer to report results on the log-odds (rather than complementary log-log) scale, since the interpretation may be more familiar. While also the results for the other estimators appear favourable because bias is defined relative to their population limit for these estimators, a key drawback of these estimators is that it is not well understood what their population limit represents.

Table 4 about here.

7. Data analysis

The First Steps program was set up in 1989 in Washington State, United States, in order to serve low-income pregnant women and children. A specific goal was to reduce the risk of low birth weight. Using data obtained from birth certificates from 2,500 children born in King County, Washington in 2001, we sought to evaluate the effects of the First

Steps program on infant birthweight, as well as its association with maternal age. We were also interested in the possible interaction between the two exposures considered.

We first carried out a more traditional analysis using parametric models. Specifically, we fit a linear model for infant birth weight (in grams), with an indicator of participation on the First Steps program and maternal age as predictors, as well other baseline covariates (child's sex, mother's age, race (asian, black, hispanic, white or other), number of previous live born infants, weight prior to pregnancy, education, smoking status and marital status). This model yielded estimates of -13.57 (95% CI: -76.34 to 49.20) for First Steps participation. Assuming that we have adjusted for all common causes of First Steps participation and birth weight, and additionally that the linear model is correctly specified, then the first regression coefficient suggests that participation in the program led to an average reduction of -13.57 grams in birth weight (although the confidence interval contained the null). For comparison, fitting a linear model unadjusted for covariates yielded an estimate of -66.18 (95% CI: -125.79 to -6.57), such that ignoring confounding gives the impression that the intervention was harmful. We then refit the linear model with an interaction term; it was estimated that the association between program participation and birth weight increased by 2.7 units per year increase in maternal age (95% CI: -6.99 to 12.33). We fit a separate linear model, adjusted for all other covariates except program participation, to assess the effect of age which was estimated as 0.037 (95% CI: -4.40 to 4.47). We did not adjust for participation given that it was an externally introduced factor that may be predicted by age.

We repeated this analysis after dichotomising the outcome (an infant was considered to have low birth weight if they weighed $< 2,500g$). The estimated log-odds ratios for low birth weight were -0.038 (95% CI: -0.55 to 0.45) for First Steps participation and 0.037 for age (95% CI: 0.00, 0.07), again taken from separate models.

We re-analysed the data using the methods proposed in this article; first we estimated the propensity-overlap weighted effect of First Steps participation on birth weight using the influence function-based estimator in (15). The nuisance functionals $E(A|L)$ and $E(Y|L)$ (along with all others described in the section) were estimated using the SuperLearner. The SuperLearner library included a generalised linear model with main effects only, a generalised linear model with main effects and pairwise interactions, random forests regression, support vector machines, k-nearest neighbours andthe default generalised additive model procedures as well as 8 additional variants of it with degrees of freedom fixed at 3, ..., 10. We obtained an estimate of -5.91 (95% CI: -85.12 to 73.31), which was smaller in magnitude than in the previous analysis, and reflects our a priori belief that program participation is unlikely to lead to a strong decrease in infant birth weight. In looking at the weighted effect of maternal age, we again did not adjust for program participation. The proposal yielded an estimate of -1.39 (95% CI: -6.32 to 3.54). By construction, these can be interpreted as the main effects of First Steps participation and age, regardless of the presence of possible interactions. In a subsequent analysis, we also re-estimated the interaction between First Steps participation and maternal age without making assumptions about possible dependencies between these exposures, and found the interaction to be more pronounced. We obtained an estimate of 6.96 (95% CI: -6.90 to 20.82) based on SuperLearner. Repeating this analysis for the weighted average difference of log-odds of low birthweight gave the effect of program participation as 0.01

(95% CI: -0.43 to 0.44) and maternal age as 0.055 (95% CI: 0.03 to 0.08).

8. Discussion

We have emphasised that most data analyses rely on modelling assumptions in more intricate ways than we may realise. They extract information from those assumptions, rather than from the data alone. This may result in estimators for, for instance, a conditional association that are not guaranteed to summarise that association well (e.g. that cannot be viewed as a weighted average of covariate-specific conditional association measures) when those modelling assumptions fail. It may moreover deliver overly optimistic uncertainty assessments, even when based on sandwich standard errors, that are only justified when those modelling assumptions hold. With others, we therefore recommend that the starting point of a data analysis becomes the choice of an estimand, as opposed to the choice of a model. This ensures that the analysis' aim is unambiguously clear at all times, regardless of issues of model misspecification. It moreover assures that uncertainty assessments, by virtue of being obtained under the nonparametric model, reflect solely the information that is contained in the data. To prevent this rendering interpretation more complicated, we have chosen to focus on estimands that can be interpreted as familiar regression parameters when corresponding models hold, but continue to capture what these parameters aim to summarise when these models are misspecified. The proposal thereby addresses the usual tension between the need for possibly complex models versus the desire to obtain easy-to-communicate results (Breiman, 2001).

The idea of starting the analysis with the choice of an estimand, has become well integrated in causal inference research (Hernan and Robins, 2020). Here, estimands are typically chosen with a view on specific interventions, whose impact one aims to assess. This literature has primarily focused on the average causal effect, $E\left(Y^1 - Y^0\right)$, which expresses how different the expected outcome would be if all subjects in the population were treated versus untreated. This is useful - in fact, often more useful than the estimands we consider - if such interventions can be conceived. For a continuous exposure, contrasts of $E\left(Y^a\right)$ for different exposure levels a are arguably less meaningful as interventions that force each one's exposure to take on level a may not be realistic (consider e.g. the effect of fixing everyone's BMI at 25) and demand enormous extrapolations. Continuous exposures moreover demand a greater need to summarise, which is naturally done by means of so-called marginal structural models in the causal inference literature (Robins et al., 2000), such as

$$E(Y^a) = \alpha + \beta a,$$

for all a. Weighted least squares regression of Y on A, using so-called stabilised weights $f(A)/f(A|L)$, then delivers an estimator for β whose probability limit equals

$$\frac{\int \{a - E(A)\} E(Y|A = a, L = l) f(a) f(l) da dl}{\text{Var}\,(A)}.$$

This expression shows that while the starting point of a causal analysis is often an explicit estimand, also here, the desire to summarise high-dimensional information leads one to working with estimands that are implicitly defined by the estimation procedure. In particular, adjustment for baseline covariates C is rather common in marginal structural models and has lead one to consider projection estimands for the parameters indexing models like

$$E(Y^a|C) = \alpha + \beta a + \gamma C.$$

These have for instance been defined, for dichotomous exposure, as the minimiser to

$$E\left[(Y^1 - \alpha - \beta - \gamma C)^2\right] + E\left[(Y^0 - \alpha - \gamma C)^2\right]$$

(Neugebauer and van der Laan, 2007). When stabilised weights are used or a non-linear link function is involved, then this raises similar concerns as in Section 2 when the dependence of Y^a on C is misspecified, for then the minimiser for β may no longer be guaranteed to capture the exposure effect on outcome.

In causal inference applications, this explicit need for summarisation can be avoided by focussing on estimands that depend on the natural value of treatment (Hubbard and Van der Laan, 2008; Muñoz and van der Laan, 2012; Young et al., 2014), for instance, that consider the effect of shifting the exposure with one unit:

$$E\left(Y^{A+1} - Y^A\right).$$

This estimand, which also reduces to β in model (4) with identity link when that model is correctly specified, is directly relevant if interest lies in the effect of interventions that aim to increase the exposure by one unit. In such settings, it is easier to interpret than the estimand (5). It has the drawback, however, that such specific interventions may be rare in practice and that the estimand is very specific to the chosen intervention. In particular, since $E\left(Y^{A+2} - Y^A\right)$ will not generally equal twice $E\left(Y^{A+1} - Y^A\right)$, a need to summarise the effects $E\left(Y^{A+a} - Y^A\right)/a$ for different levels of a may remain when there is no convincing reason to consider $a = 1$. In this paper, we have therefore opted to work with more generic estimands, that are also relevant when no specific interventions are considered (e.g. when describing the association of an outcome with age, when measuring time trends, ...), and whose efficient influence function under the nonparametric model does not involve inverse weighting by the conditional density of A, given L. Such inverse weighting complicates the use of flexible data-adaptive procedures when e.g. the exposure is continuous (it may require the need for binning, as in Muñoz and van der Laan (2012)), as conditional density estimation is a difficult problem which has received little attention in the machine learning literature. Inverse weighting also reflects a change of measure, and thus signals extrapolations being made (e.g., the fact that a one-unit increase in exposure may be very unlikely for subjects in certain covariate strata) and thus estimators that rely on it may exhibit erratic behaviour, even when the density is known. We have therefore focussed on estimands with a generic definition (regardless of whether the exposure is discrete or continuous, and regardless of whether one aims to answer a causal question or not), for which inference can be developed in a generic way (regardless of whether the exposure is discrete or continuous). Such generic estimands are important to enable broadly accessible data analyses. Nevertheless, we acknowledge that in specific circumstances, other estimands may be of greater interest.

Arguably, a drawback of the estimands considered in this paper is that they depend on the exposure distribution, as is for instance seen in (6). This may be considered undesirable (in a similar way that the partial likelihood estimator of the hazard ratio under a Cox model has been criticised for its limit depending on the censoring distribution in a complicated manner (van der Laan and Rose, 2011)); however, it is the unavoidable consequence of working with estimands that eschew inverse weighting and thus avoid strong extrapolations away from the observed exposure distribution.

In our attempt to come up with generic estimands for regression parameters, we have experienced a need for clear principles for choosing estimands, as opposed to letting them be mere projection parameters (Buja et al., 2019b). In the considered context, we have found it useful to start from the premise that $E(Y|A, L)$ is known for all levels of A and L, and to consider how to best summarise this information when the aim is parsimony. This is best done with some regression model in mind, to ensure that the estimand coincides with a familiar regression parameter when that model is correctly specified, and thus remains well interpretable. To prevent that the assumptions embodied in the entire regression model dominate the choice of estimand, we have focussed on (generalised) partially linear models, which merely specify the conditional association or effect modification term of interest. The population limit of semiparametric estimators under such model may then serve as a template for a choice of estimand. Such choice is non-unique. In our work, we have aimed for simplicity, realising that other estimands (e.g. that involve inverse weighting by the conditional outcome variance) can be inferred more efficiently. For instance, when $g(.)$ is the logit link, it may be advantageous to define the main effect estimand instead as

$$\frac{E\left(\sigma^2(A,L)\left[A - \frac{E\{\sigma^2(A,L)A|L\}}{E\{\sigma^2(A,L)|L\}}\right] g\left\{E(Y|A,L)\right\}\right)}{E\left(\sigma^2(A,L)\left[A - \frac{E\{\sigma^2(A,L)A|L\}}{E\{\sigma^2(A,L)|L\}}\right] A\right)},$$

for $\sigma^2(A, L) = E(Y|A, L)\{1 - E(Y|A, L)\}$. Further work is needed to develop inference for this estimand, and insight in its interpretation.

In future work, we will make similar developments for parameters indexing proportional hazard models for time-to-event data and marginal models for repeated measures data. We will moreover study how the dependence of $E(Y|A = a, L = l)$ for continuous a can be described in a less restrictive way by focussing on the estimand

$$\frac{E\left[\mathrm{Var}\,(A|L)\, g\left\{E(Y|A=a,L)\right\}\right]}{E\left\{\mathrm{Var}\,(A|L)\right\}}$$

as an unrestricted function of a.

Acknowledgement

The authors would like to thank Mark van der Laan for inspiring discussions that have influenced this work, and Vanessa Didelez, Oliver Hines and Pawel Morzywolek for useful feedback. We gratefully acknowledge support from BOF Grants BOF.01P08419 and BOF.24Y.2017.0004.01.

References

Angrist, J. D. and Krueger, A. B. (1999) Empirical Strategies in Labor Economics. In *Handbook of Labor Economics*, vol. 3, 1277–1366. Elsevier.

Angrist, J. D. and Pischke, J.-S. (2009) *Mostly harmless econometrics: an empiricist's companion*. Princeton: Princeton University Press. OCLC: ocn231586808.

Aronow, P. M. and Samii, C. (2016) Does Regression Produce Representative Estimates of Causal Effects? *American Journal of Political Science*, **60**, 250–267.

Athey, S., Tibshirani, J., Wager, S. et al. (2019) Generalized random forests. *The Annals of Statistics*, **47**, 1148–1178.

Belloni, A., Chernozhukov, V. and Wei, Y. (2013) Honest confidence regions for a regression parameter in logistic regression with a large number of controls. *Tech. rep.*, cemmap working paper.

Berk, R., Brown, L., Buja, A., Zhang, K. and Zhao, L. (2013) Valid post-selection inference. *The Annals of Statistics*, **41**, 802–837.

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*, vol. 4. Johns Hopkins University Press Baltimore.

Breiman, L. (2001) Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, **16**, 199–231.

Buja, A., Brown, L., Berk, R., George, E., Pitkin, E., Traskin, M., Zhang, K. and Zhao, L. (2019a) Models as Approximations I: Consequences Illustrated with Linear Regression. *Statistical Science*, **34**, 523–544.

Buja, A., Brown, L., Kuchibhotla, A. K., Berk, R., George, E., Zhao, L. et al. (2019b) Models as approximations ii: A model-free theory of parametric regression. *Statistical Science*, **34**, 545–565.

Buja, A., Kuchibhotla, A. K., Berk, R., George, E., Tchetgen Tchetgen, E. and Zhao, L. (2019c) Models as approximations—rejoinder. *Statistical Science*, **34**, 606–620.

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.

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. *The Econometrics Journal*, **21**, C1–C68.

Crump, R. K., Hotz, V. J., Imbens, G. W. and Mitnik, O. A. (2006) Moving the goalposts: Addressing limited overlap in the estimation of average treatment effects by changing the estimand. *Tech. rep.*, National Bureau of Economic Research.

Freedman, D. A. (2006) On The So-Called "Huber Sandwich Estimator" and "Robust Standard Errors". *The American Statistician*, **60**, 299–302.

Graham, B. S. and Pinto, C. C. d. X. (2018) Semiparametrically efficient estimation of the average linear regression function. *arXiv:1810.12511 [econ]*. URL: `http://arxiv.org/abs/1810.12511`. ArXiv: 1810.12511.

Hernan, M. A. and Robins, J. M. (2020) *Causal inference: What If.* Boca Raton: Chapman & Hall/CRC.

Hubbard, A. E. and Van der Laan, M. J. (2008) Population intervention models in causal inference. *Biometrika*, **95**, 35–47.

Kennedy, E. H., Lorch, S. and Small, D. S. (2019) Robust causal inference with continuous instruments using the local instrumental variable curve. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **81**, 121–143.

Kosorok, M. R. (2007) *Introduction to empirical processes and semiparametric inference.* Springer Science &amp; Business Media.

Van der Laan, M. J., Polley, E. C. and Hubbard, A. E. (2007) Super learner. *Statistical applications in genetics and molecular biology*, **6**.

van der Laan, M. J. and Rose, S. (2011) *Targeted Learning.* Springer Series in Statistics. New York, NY: Springer New York.

van der Laan, M. J. and Rubin, D. (2006) Targeted Maximum Likelihood Learning. *The International Journal of Biostatistics*, **2**.

Lin, W. (2013) Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *The Annals of Applied Statistics*, **7**, 295–318.

Muñoz, I. D. and van der Laan, M. (2012) Population intervention causal effects based on stochastic interventions. *Biometrics*, **68**, 541–549.

Neugebauer, R. and van der Laan, M. (2007) Nonparametric causal effects based on marginal structural models. *Journal of Statistical Planning and Inference*, **137**, 419–434.

Newey, W. K. and Robins, J. R. (2018) Cross-Fitting and Fast Remainder Rates for Semiparametric Estimation. *arXiv:1801.09138 [math, stat].* URL: `http://arxiv.org/abs/1801.09138`. ArXiv: 1801.09138.

Nie, X. and Wager, S. (2017) Quasi-oracle estimation of heterogeneous treatment effects. *arXiv preprint arXiv:1712.04912.*

Pfanzagl, J. (1990) Estimation in semiparametric models. In *Estimation in Semiparametric Models*, 17–22. Springer.

Robins, J., Li, L., Tchetgen, E., van der Vaart, A. et al. (2008) Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and statistics: essays in honor of David A. Freedman*, 335–421. Institute of Mathematical Statistics.

Robins, J. M., Hernan, M. A. and Brumback, B. (2000) Marginal structural models and causal inference in epidemiology.

Robins, J. M., Mark, S. D. and Newey, W. K. (1992) Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, 479–495.

Robinson, P. M. (1988) Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, 931–954.

Rotnitzky, A. and Robins, J. (1997) Analysis of semi-parametric regression models with non-ignorable non-response. *Statistics in medicine*, **16**, 81–102.

Rotnitzky, A., Robins, J. M. and Scharfstein, D. O. (1998) Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the american statistical association*, **93**, 1321–1339.

Scharfstein, D. O., Rotnitzky, A. and Robins, J. M. (1999) Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, **94**, 1096–1120.

Słoczyński, T. (2020) Interpreting OLS estimands when treatment effects are heterogeneous: Smaller groups get larger weights. *The Review of Economics and Statistics*.

Tan, Z. (2019) On doubly robust estimation for logistic partially linear models. *Statistics and Probability Letters*, **155**, 108577.

Tchetgen Tchetgen, E. J. (2013) On a closed-form doubly robust estimator of the adjusted odds ratio for a binary exposure. *American journal of epidemiology*, **177**, 1314–1316.

Tchetgen Tchetgen, E. J., Robins, J. M. and Rotnitzky, A. (2010) On doubly robust estimation in a semiparametric odds ratio model. *Biometrika*, **97**, 171–180.

Vansteelandt, S. and Daniel, R. M. (2014) On regression adjustment for the propensity score. *Statistics in medicine*, **33**, 4053–4072.

Vansteelandt, S., Joffe, M. et al. (2014) Structural nested models and g-estimation: the partially realized promise. *Statistical Science*, **29**, 707–731.

Vansteelandt, S., VanderWeele, T. J., Tchetgen, E. J. and Robins, J. M. (2008) Multiply robust inference for statistical interactions. *Journal of the American Statistical Association*, **103**, 1693–1704.

Wasserman, L. (2014) Discussion: "A significance test for the lasso". *The Annals of Statistics*, **42**, 501–508.

White, H. (1980) A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica*, **48**, 817.

Whitney, D., Shojaie, A. and Carone, M. (2019) Comment: Models as (deliberate) approximations. *Statistical Science*, **34**, 591–598.

Young, J. G., Hernán, M. A. and Robins, J. M. (2014) Identification, estimation and approximation of risk under interventions that depend on the natural value of treatment using observational data. *Epidemiologic methods*, **3**, 1–19.

Zheng, W. and van der Laan, M. J. (2011) Cross-Validated Targeted Minimum-Loss-Based Estimation. In *Targeted Learning*, 459–474. New York, NY: Springer New York.

Table 1. Simulation results on main effects: empirical bias (Bias) and standard deviation (SD), sample average of the estimated influence-function based standard errors (SE), and coverage of 95% Wald confidence intervals (Cov). Bias and coverage taken w.r.t. the truth 0.3 in experiments (Exp.) 1 and 2, and w.r.t. the limiting values of each estimator in experiment 3 (0.33 (MLE) 0.43 (ES), 1.00 (DR) and 0.50 (AL)) and in experiment 4 (-0.08 (MLE), 0.19 (ES) 0.37 (DR) and 0.21 (AL)).

Exp.	Est.	$n = 500$				$n = 1000$				$n = 2000$			
		Bias	SD	SE	Cov	Bias	SD	SE	Cov	Bias	SD	SE	Cov
1	MLE	0.00	0.21	0.21	95	0.00	0.15	0.15	95	0.00	0.11	0.10	94
	ES	0.04	0.20	0.19	93	0.03	0.15	0.14	92	0.02	0.11	0.10	92
	DR	0.06	0.21	0.23	96	0.05	0.15	0.16	96	0.03	0.11	0.11	95
	AL	0.02	0.19	0.20	95	0.02	0.14	0.14	95	0.01	0.10	0.10	94
2	MLE	-0.59	0.22	0.22	26	-0.59	0.15	0.16	3	-0.59	0.11	0.11	0
	ES	-0.14	0.21	0.20	86	-0.04	0.16	0.14	90	-0.02	0.12	0.10	91
	DR	-0.11	0.22	0.24	90	-0.01	0.17	0.18	96	0.01	0.12	0.13	95
	AL	-0.17	0.20	0.21	89	-0.07	0.15	0.15	92	-0.04	0.11	0.11	93
3	MLE	0.01	0.26	0.23	92	0.01	0.18	0.16	94	0.00	0.13	0.12	94
	ES	0.04	0.28	0.23	88	0.05	0.22	0.18	88	0.02	0.17	0.14	88
	DR	-0.60	0.21	0.19	16	-0.50	0.16	0.14	7	-0.40	0.13	0.10	6
	AL	-0.05	0.28	0.23	88	0.00	0.22	0.19	91	0.01	0.17	0.15	92
4	MLE	-0.01	0.20	0.21	95	0.00	0.15	0.15	95	0.00	0.10	0.10	96
	ES	-0.11	0.20	0.19	91	-0.06	0.16	0.14	90	-0.04	0.12	0.11	91
	DR	-0.29	0.21	0.21	69	-0.25	0.16	0.15	61	-0.22	0.11	0.11	48
	AL	-0.12	0.19	0.20	92	-0.07	0.15	0.14	91	-0.04	0.12	0.11	92

Table 2. Simulation results on effect modification: empirical bias (Bias) and standard deviation (Emp SD), sample average of the estimated influence-function based standard errors (Mean SE), and coverage of 95% Wald confidence intervals (Cov).

Exp.	Est.	$n = 500$				$n = 1000$				$n = 2000$			
		Bias	SD	SE	Cov	Bias	SD	SE	Cov	Bias	SD	SE	Cov
1	OLS	-0.047	0.051	0.051	84	-0.046	0.037	0.036	76	-0.046	0.027	0.025	55
	AL	-0.034	0.067	0.073	95	-0.016	0.051	0.050	93	-0.015	0.036	0.035	92
2	OLS	-2.92	0.24	0.23	0	-2.92	0.17	0.16	0	-2.92	0.11	0.11	0
	AL	-0.31	0.15	0.16	49	-0.12	0.077	0.085	77	-0.057	0.044	0.052	88
3	OLS	0.00	0.015	0.015	94	0.00	0.010	0.010	95	0.00	0.007	0.007	97
	AL	0.019	0.042	0.043	93	0.013	0.027	0.029	95	0.002	0.018	0.021	97

Table 3. Simulation results on variable selection: empirical bias (Bias) and standard deviation (Emp SD), sample average of the estimated influence-function based standard errors (Mean SE), and coverage of 95% Wald confidence intervals (Cov) for post-lasso (P-lasso), Lasso, ϵ-net and SuperLearner (SL), and the proposed variants thereof (AL).

Est.	$n = 200$				$n = 400$			
	Bias	SD	SE	Cov	Bias	SD	SE	Cov
P-lasso	0.15	0.21	0.13	65	0.095	0.14	0.099	73
AL P-lasso	0.072	0.21	0.13	75	0.039	0.13	0.093	83
Lasso	-0.031	0.088			0.010	0.072		
AL Lasso	-0.00011	0.15	0.14	93	-0.00096	0.10	0.099	94
ϵ-net	-0.073	0.051			-0.042	0.041		
AL ϵ-net	-0.0085	0.14	0.14	93	-0.011	0.098	0.096	95
SL	-0.19	0.17	0.019	30	0.0048	0.078	0.00033	0.4
AL SL	0.42	0.22	0.11	13	0.096	0.11	0.096	79

Table 4. Simulation results with a misspecified link function: empirical bias (Bias) and standard deviation (SD), sample average of the estimated influence-function based standard errors (SE), and coverage of 95% Wald confidence intervals (Cov). Bias and coverage taken w.r.t. the limiting values of each estimator in experiment 1 (0.3 (MLE-cloglog), 0.51 (MLE-logit), 0.55 (AL)), experiment 3 (0.04 (MLE-cloglog), 0.03 (MLE-logit) and 0.1 (AL)) and experiment 4 (0.29 (MLE-cloglog), 0.42 (MLE-logit) and 0.12 (AL)); in experiment 2, bias/coverage were taken w.r.t. the truth for MLE-cloglog and w.r.t the population limits for the other estimators (0.83 (MLE-logit), 0.52 (AL))

Exp.	Est.	$n = 500$				$n = 1000$				$n = 2000$			
		Bias	SD	SE	Cov	Bias	SD	SE	Cov	Bias	SD	SE	Cov
1	MLE-cloglog	0.01	0.14	0.13	94	0.01	0.10	0.09	93	0.00	0.07	0.07	96
	MLE-logit	0.02	0.25	0.24	94	0.01	0.17	0.17	94	0.00	0.12	0.12	96
	AL	-0.02	0.21	0.20	94	-0.03	0.16	0.15	92	-0.03	0.11	0.11	93
2	MLE-cloglog	0.24	0.14	0.14	58	0.23	0.10	0.10	33	0.23	0.07	0.07	8
	MLE-logit	0.03	0.24	0.23	94	0.01	0.17	0.16	95	0.00	0.11	0.11	95
	AL	0.10	0.21	0.20	90	0.06	0.16	0.14	88	0.03	0.12	0.10	91
3	MLE-cloglog	0.00	0.15	0.14	95	0.00	0.10	0.10	94	0.00	0.07	0.07	95
	MLE-logit	0.00	0.24	0.23	95	0.00	0.16	0.16	95	-0.01	0.11	0.11	95
	AL	0.02	0.21	0.20	94	0.01	0.15	0.14	94	0.01	0.11	0.10	93
4	MLE-cloglog	0.00	0.13	0.14	96	0.01	0.09	0.10	95	0.00	0.07	0.07	95
	MLE-logit	0.01	0.20	0.21	96	0.01	0.14	0.15	95	0.00	0.10	0.10	95
	AL	0.10	0.19	0.20	93	0.06	0.14	0.14	93	0.02	0.10	0.10	94