Statistical Science
2023, Vol. 38, No. 1, 1–12
https://doi.org/10.1214/21-STS843

The Costs and Benefits of Uniformly Valid Causal Inference with High-Dimensional Nuisance Parameters

Niloofar Moosavi, Jenny Häggström and Xavier de Luna

Abstract. Important advances have recently been achieved in developing procedures yielding uniformly valid inference for a low dimensional causal parameter when high-dimensional nuisance models must be estimated. In this paper, we review the literature on uniformly valid causal inference and discuss the costs and benefits of using uniformly valid inference procedures. Naive estimation strategies based on regularization, machine learning, or a preliminary model selection stage for the nuisance models have finite sample distributions which are badly approximated by their asymptotic distributions. To solve this serious problem, estimators which converge uniformly in distribution over a class of data generating mechanisms have been proposed in the literature. In order to obtain uniformly valid results in high-dimensional situations, sparsity conditions for the nuisance models need typically to be made, although a double robustness property holds, whereby if one of the nuisance model is more sparse, the other nuisance model is allowed to be less sparse. While uniformly valid inference is a highly desirable property, uniformly valid procedures pay a high price in terms of inflated variability. Our discussion of this dilemma is illustrated by the study of a double-selection outcome regression estimator, which we show is uniformly asymptotically unbiased, but is less variable than uniformly valid estimators in the numerical experiments conducted.

Key words and phrases: Double robustness, machine learning, post-model selection inference, regularization, superefficiency.

1. INTRODUCTION

High-dimensional situations, where the number of covariates is larger than the number of observations are common in causal inference applications. Using regularization type estimators such as lasso [54] or other post-model selection estimators are popular strategies in such cases. Important advances have been achieved in developing procedures yielding uniformly valid inference (defined below) for a low dimensional causal parameter when high-dimensional nuisance models must be fitted (e.g., [5, 13, 24, 57, 59]). In this paper, we review the literature on uniformly valid causal inference, and discuss the costs and benefits of using uniformly valid inference procedures. This discussion is important since naive and invalid post-model selection inference is to this day still common in statistical practice.

Leeb and Pötscher [34] demonstrated how a data-driven model selection step can affect the distribution of the estimate of a parameter of interest. Loosely, they show that the scaled ($\sqrt{n}$) bias of a naive two step estimator, which does not take into account the selection step, goes to infinity or stay bounded for a sequence of worst case scenario data generating processes (DGPs), when relying on consistent or conservative model selection, respectively. We say that such a naive estimator is not uniformly unbiased. An estimator with associated uniformly valid inference, on the other hand, is such that its distribution F_n converges uniformly over a set of DGPs $\mathcal{P}$, that is, for any $u \in \mathbb{R}$

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}} \left|F_n(u) - F_P(u)\right| = 0,$$

Niloofar Moosavi is Ph.D. Student, Department of Statistics, USBE, Umeå University, 901 87, Umeå, Sweden (e-mail: niloofar.moosavi@umu.com). Jenny Häggström is Associate Professor, Department of Statistics, USBE, Umeå University, 901 87, Umeå, Sweden (e-mail: jenny.haggstrom@umu.se). Xavier de Luna is Professor, Department of Statistics, USBE, Umeå University, 901 87, Umeå, Sweden (e-mail: xavier.de.luna@umu.se).

where F_P is the cumulative distribution derived from P. Hereafter, we use the terms uniformly valid inference and valid inference interchangeably.

Valid inference for every single parameter in a linear regression model needs careful consideration in high-dimensional settings. Some have considered debiasing lasso for a valid inference that target the true data generating process parameters [31, 56, 61] and others have considered valid inference conditional on the model that has been selected [6, 33]. In a causal inference context, there is typically a low dimensional parameter of interest, for example, the average causal effect of a treatment, and high-dimensional nuisance parameters. Belloni, Chernozhukov and Hansen [5] proposed an estimator with valid inference for a causal parameter in a linear model explaining outcome with a treatment variable and a set of covariates, which can be of high dimension. To achieve uniformly valid inference, they proposed to include the union of two sets of covariates in the model: one obtained by selecting covariates relevant when regressing the outcome on the covariates, and the second by selecting covariates relevant when regressing the treatment on the covariates. Their model implicitly implies a homogeneous causal effect. This can be relaxed to allow for individual heterogeneous effects, using the potential outcome framework [43, 50], and nuisance models for both the potential outcomes and for the treatment assignment given the covariates (propensity score). For this general case, van der Laan and Rubin [59] and van der Laan [57] obtained valid inference for a causal parameter using targeted maximum likelihood estimation, where nuisance models are estimated nonparametrically. Farrell [24] considered the augmented inverse probability of treatment weighting estimator [41], and showed that uniformly valid inference is achieved when using post-lasso estimation for the nuisance models. Similar results were derived in [13] using a double machine learning approach.

In both Belloni, Chernozhukov and Hansen [5] and Farrell [24], approximate sparsity is assumed for the nuisance models. However, in Farrell [24] the outcome model can be less sparse if the propensity score is more sparse and vice versa (called nonparametric double robustness property, [32]). Yet consistency in the nonparametric estimation of all the nuisance models is required, a condition relaxed in van der Laan [57], and in more recent work [2, 52], where one of the nuisance models may be inconsistently estimated.

Procedures yielding uniformly valid inference for a causal parameter, in the general context of heterogeneous treatment effects, allow for the selection of instruments (loosely, variables related to the treatment but not the outcome) in the fit of the propensity score model. This is known to result in possibly large inflation of the variance of the estimators (e.g., [19, 29, 42, 46]). Thus, while uniformly valid inference is a highly desirable property, uniformly valid procedures pay a high price in terms of inflated variability. We discuss and illustrate this dilemma by studying a compromise solution, an outcome regression estimator (e.g., [51]), which we allow to select instruments, but which does not use the fitted propensity score, in contrast with uniformly valid estimators proposed in the literature. The resulting post-model selection estimator is shown to be uniformly asymptotically unbiased under a commonly used product rate condition [24], even though the propensity score is not used in the estimator except for the covariate selection step.

This paper is organised as follows. Section 2 presents a review of the literature on uniformly valid causal inference. Section 3 gives a theoretical discussion of the costs and benefits of uniformly valid inference, by studying a double-selection outcome regression estimator. Section 4 illustrates this discussion with a Monte Carlo study of finite sample properties of a collection of estimators. Section 5 concludes the paper. All proofs are delayed to an Appendix.

2. UNIFORMLY VALID CAUSAL INFERENCE: A REVIEW

This is an extremely active research area. The focus is here on uniformly valid inference on a low dimensional causal parameter after regularization/model selection of high-dimensional nuisance models. We start by introducing some general concepts, and then review first cornerstone work on the homogeneous and then general heterogeneous case. A review of important advances in recent years concludes this section.

The parameter of interest is a causal effect of a binary treatment variable T on an outcome Y as defined below in different contexts. We use the notation X to denote a one dimensional pretreatment covariate and $\boldsymbol{X}$ to denote a set of pretreatment covariates which has dimension p, allowed to grow with n. Note that the set may contain not only the covariates but also transformations of them. We consider a set of identically and independently distributed (i.i.d.) observations, $\{(\boldsymbol{x}_i, y_i, t_i)\}_{i=1}^n$, drawn from a distribution P_n. To study uniformly valid post-model selection inference it is essential that the probability law P_n is allowed to vary with the sample size n. We use further the notations $E_n[w_i] = \frac{1}{n}\Sigma_{i=1}^n w_i$ and $a \vee b = \max\{a, b\}$. Moreover, n_t denotes the number of individuals under treatment.

2.1 Hodges Estimator and Superefficiency

Superefficient estimators of a parameter of a model $\mathcal{M}$ are variants of the well-known Hodges estimator [60] when a model restriction holds; meaning that a model $\mathcal{M}_0 \subset \mathcal{M}$ contains the true data generating process. The asymptotic variance of a superefficient estimator is smaller than the efficiency bound for the class of regular asymptotic linear (RAL) estimators of the parameter under model $\mathcal{M}$, when there is a submodel $\mathcal{M}_0$ under which the same bound is smaller. Superefficiency has a cost in the sense that the asymptotic distribution of a superefficient estimator is valid only pointwise at $\mathcal{M}_0$ instead of uniformly over a larger family of models (uniformly valid inference).

This was highlighted by Leeb and Pötscher [34] in a parametric setting (see below for a detailed exposition), where a consistent model selection step that selects out a "redundant" variable (not part of $\mathcal{M}_0$) before a maximum likelihood fit results in a superefficent estimator. Such an estimator has an oracle property in the sense that it asymptotically (only pointwise at $\mathcal{M}_0$ instead of uniformly over a larger family of models) performs as well as a fictitious orcale estimator which can be constructed by knowing $\mathcal{M}_0$ [23].

2.2 Homogenous Causal Effect

As a primer, consider the parametric regression model $y_i = \alpha t_i + \beta x_i + \epsilon_i$, with $\epsilon_i \sim \mathcal{N}(0, \sigma^2), \sigma^2 > 0$. According to Leeb and Pötscher [34], if we are interested in α, the post-selection estimator which includes a preliminary consistent model selection step on X (i.e., a test whether $\beta = 0$) is more efficient than the simple OLS estimator without this step if $\operatorname{corr}(X, T) \neq 0$. However, if $\beta \neq 0$ and $\operatorname{corr}(X, T) \neq 0$, the finite sample distribution of the post-selection estimate is a mixture of two normal distributions, that is, not well approximated by the normal asymptotic distribution. This is because in the selection step the nonzero coefficient can be detected for some samples and not detected for others. If the resulting omitted variable bias is considerable, the empirical coverage of a naive confidence interval can be far from the nominal coverage. Leeb and Pötscher [34] have shown that the minimal coverage of the naive confidence interval with respect to all possible β values goes to zero as n grows for consistent model selection steps, while the empirical coverage for any fixed β value goes to the nominal one. Their result highlights the importance of uniformly valid inference compared to pointwise asymptotic results.

For α to have a causal interpretation, the linear regression needs to include all confounders as formally defined in next section, and it must be correctly specified as a model for $E(Y \mid T, \boldsymbol{X})$ (in particular implying a homogeneous/constant causal effect). Concerning the former condition, the number of available covariates, hence potential confounders, may be very large when using large observational databases. The latter condition, implicitly requires that series expansions need to be used to approximate $E(Y \mid T, \boldsymbol{X})$ increasingly well with increasing sample sizes. Therefore, these two conditions often yield a high-dimensional setting in practice, that is, where the number of covariates is at least as large as the sample size. Belloni, Chernozhukov and Hansen [5] proposed a strategy for reaching valid inference in such high-dimensional settings [35]. Their suggestion is a two-step lasso-based method, where the union of covariate sets selected by two distinct lasso regressions of Y and T on the covariates, respectively (often called double selection), are utilized in a second step in the main linear model including T as a regressor. Instead of exact sparsity conditions typically used in high-dimensional settings, they consider the following approximate sparsity conditions. Let

$$E(Y|T, \boldsymbol{X}) = \alpha T + \beta_Y' \boldsymbol{X} + R_Y, \tag{1}$$

$$E(T|\boldsymbol{X}) = \beta_T' \boldsymbol{X} + R_T, \tag{2}$$

where R_f is the specification error of using a sparse β_f with only s_f nonzero elements, respectively for $f = Y, T$. The regularity conditions to obtain uniform valid inference include

$$E\big(E_n[R_{f,i}^2]\big)^{1/2} = O(\sqrt{s_f/n}),$$

$$\log^3 p/n = o(1),$$

$$s_f^2 \log^2(p \vee n)/n = o(1).$$

In other words, models (1) and (2) are assumed well approximated by a sparse linear combination of the covariate vector $\boldsymbol{X}$. Under these conditions (and other regularity conditions), Belloni, Chernozhukov and Hansen ([5], Corollary 2) showed that their estimator of α is asymptotically normal uniformly over P_n, thereby uniformly valid inference can be made.

2.3 Heterogeneous Causal Effect

The effect of a binary treatment is now allowed to be heterogeneous using the Neyman–Rubin potential outcome framework [43, 50]. For any unit in the study, denote $Y(1)$ its potential outcome under treatment ($T = 1$), and $Y(0)$ its potential outcome without treatment (or alternative treatment). We assume that $Y = TY(1) + (1 - T)Y(0)$ is the observed outcome, and no interference between units is allowed (stable unit treatment value assumption; [44]). Each unit may have a different causal effect Y(1) - Y(0), and the average causal effect $\tau = E(Y(1) - Y(0))$ is the parameter of interest in the sequel. This parameter is identified given the following assumptions.

ASSUMPTION 1 (No unobserved confounding).

$$Y(1), Y(0) \perp\!\!\!\perp T|\boldsymbol{X}.$$

ASSUMPTION 2 (Overlap).

$$\mathbb{P}(T = t|\boldsymbol{X}) \geq p_{\min} > 0, \quad t = 0, 1.$$

Thus, all confounding covariates are included in $\boldsymbol{X}$ and all units in the study have nonzero probability to be included in both treatment groups. These assumptions are made throughout the article. Note, however, that a sensitivity analysis (see [21], and references therein) should accompany inference based on Assumption 1, even when many covariates are available, since this assumption is not testable without further information (e.g., [18]). Assumption 2 has also important implications as was recently demonstrated in D'Amour et al. [17] for high-dimensional situations, where overlap is linked to the sparsity conditions discussed herein.

Let $E(Y|T = 1, \boldsymbol{X}) = m_1(\boldsymbol{X})$ and $E(Y|T = 0, \boldsymbol{X}) = m_0(\boldsymbol{X})$ denote the outcome models, $E(T|\boldsymbol{X}) = \mathbb{P}(T = 1|\boldsymbol{X}) = e(\boldsymbol{X})$ denotes the propensity score model. One of the earliest proposals that addressed inference on τ when estimating nuisance models nonparametrically is the targeted maximum likelihood estimator (TMLE; [58, 59]). Denote fits of the nuisance models $\mathbb{P}(T = 1|\boldsymbol{X})$, $E(Y(1)|\boldsymbol{X})$ and $E(Y(0)|\boldsymbol{X})$ by $\hat{e}(\boldsymbol{x}_i)$, $\hat{m}^0_1(\boldsymbol{x}_i)$ and $\hat{m}^0_0(\boldsymbol{x}_i)$, respectively. Then, a fluctuated version of the predicted outcome values is used in the following manner:

$$\hat{\tau}_{\mathrm{TMLE}} = E_n\big[\hat{m}^1_1(\boldsymbol{x}_i) - \hat{m}^1_0(\boldsymbol{x}_i)\big],$$

where the fluctuations are found by

$$\operatorname{logit} \hat{m}^1_t(\boldsymbol{x}_i) = \operatorname{logit} \hat{m}^0_t(\boldsymbol{x}_i) + \varepsilon_n h_t(\boldsymbol{x}_i), \quad t \in \{0, 1\},$$

where

$$h_t(\boldsymbol{x}_i) = \frac{\mathbb{1}\{t = 1\}}{\hat{e}(\boldsymbol{x}_i)} - \frac{\mathbb{1}\{t = 0\}}{1 - \hat{e}(\boldsymbol{x}_i)},$$

and ε_n is found by running logistic regression of outcome Y on $h_T(\boldsymbol{X})$ using $\operatorname{logit} \hat{m}^0_T(\boldsymbol{X})$ as intercept. TMLE is consistent if either $e(\cdot)$ or $m_0(\cdot)$ and $m_1(\cdot)$ are consistently estimated. Moreover, it is RAL and semiparametrically efficient if all models are consistently estimated and a product of rate of convergence similar to (9) hold. TMLE can also be constructed by iteratively fluctuating the propensity score and the outcome models, thereby yielding a RAL estimator when at most one of the models is consistently estimated, that is, this TMLE is then not only consistent but also asymptotically normal (so called double robust statistical inference, [57]).

In another major contribution, Farrell [24] showed how to obtain uniformly valid inference for the popular doubly robust augmented inverse probability of treatment weighting (AIPW, [41, 45]) estimator:

$$\begin{aligned} \hat{\tau}_{\mathrm{DR}} = E_n\bigg[&\frac{t_i y_i - (t_i - \hat{e}(\boldsymbol{x}_i))\hat{m}_1(\boldsymbol{x}_i)}{\hat{e}(\boldsymbol{x}_i)} \\ &- \frac{(1 - t_i)y_i + (t_i - \hat{e}(\boldsymbol{x}_i))\hat{m}_0(\boldsymbol{x}_i)}{1 - \hat{e}(\boldsymbol{x}_i)}\bigg]. \end{aligned} \tag{3}$$

The fitted values $\hat{m}_0(\boldsymbol{x}_i)$, $\hat{m}_1(\boldsymbol{x}_i)$ and $\hat{e}(\boldsymbol{x}_i)$ are obtained using, for example, post-lasso estimators. Farrell [24] proposed the use of group-lasso to benefit from grouped sparsity patterns among potential outcomes and different treatment levels. Similar to Belloni, Chernozhukov and Hansen [5], Farrell [24] assumed approximate sparsity, but for logistic propensity score and linear potential outcome models, that is,

$$\begin{aligned} m_t(\boldsymbol{X}, \eta_t) &= \eta_t' \boldsymbol{X} + R^t_Y, \quad t \in \{0, 1\}, \\ e(\boldsymbol{X}, \gamma) &= \operatorname{expit}(\gamma' \boldsymbol{X} + R_T), \end{aligned} \tag{4}$$

where R^t_Y and R_T are approximation errors of estimating the true models with sparse parameters η_t and γ that have s^t_Y and s_T nonzero elements, respectively. Farrell's regularity conditions on the specification errors are slightly different from those in Belloni, Chernozhukov and Hansen [5]. In particular, he assumes

$$\begin{aligned} &E_n\big[(r^t_{Y,i})^2\big]^{1/2} \vee E_{n_t}\big[(r^t_{Y,i})^2\big]^{1/2} \\ &\quad \leq \mathcal{R}^t_Y = O\Big(\sqrt{s^t_Y/n}\Big), \quad t \in \{0, 1\}, \end{aligned} \tag{5}$$

and

$$\begin{aligned} &E_n\big[\big(\operatorname{expit}(\gamma' \boldsymbol{x}_i) - \operatorname{expit}(\gamma' \boldsymbol{x}_i + r_{T,i})\big)^2\big]^{1/2} \\ &\quad \leq \mathcal{R}_T = O(\sqrt{s_T/n}). \end{aligned} \tag{6}$$

More importantly, the sparsity assumption required for each nuisance model separately is weaker compared to the one assumed by Belloni, Chernozhukov and Hansen [5], that is,

$$s_f \log(p \vee n)^{3/2+\delta} = o(n), \quad s_f \in \{s_T, s^0_Y, s^1_Y\}, \tag{7}$$

for some $\delta > 0$, since here we have a multiplicative rate condition

$$s^t_Y s_T \log(p \vee n)^{3+2\delta} = o(n), \quad t \in \{0, 1\}. \tag{8}$$

Thus, if the potential outcome models are more sparse, the propensity score model is allowed to be more dense and the other way around. These sparsity assumptions and other regularity conditions result in the following rates of convergence for the post-lasso estimators of the nuisance models ([24], Section 6):

$$\begin{aligned} &E_n\big[\big(\hat{e}(\boldsymbol{x}_i) - e(\boldsymbol{x}_i)\big)^2\big] = o_{P_n}(1), \\ &E_n\big[\big(\hat{m}_t(\boldsymbol{x}_i) - m_t(\boldsymbol{x}_i)\big)^2\big] = o_{P_n}(1), \\ &E_n\big[\big(\hat{e}(\boldsymbol{x}_i) - e(\boldsymbol{x}_i)\big)^2\big]^{1/2} E_n\big[\big(\hat{m}_t(\boldsymbol{x}_i) - m_t(\boldsymbol{x}_i)\big)^2\big]^{1/2} \\ &\quad = o_{P_n}\big(n^{-1/2}\big), \end{aligned} \tag{9}$$

for $t \in \{0, 1\}$. Under these consistency and product rate conditions (and other regularity assumptions) the estimator (3) is asymptotically normal uniformly over P_n ([24], Corollary 3). This result is not restricted to the post-lasso estimator, but ensures $\sqrt{n}$-consistency of the AIPW estimator of the low dimensional parameter of interest τ for any estimator of nuisance models which fulfills the assumptions. High-dimensional parametric or nonparametric nuisance models fit into this framework even though the estimation cannot be done at the $\sqrt{n}$-rate.

The presentation above has focused on robustness to the danger of including too few covariates (regularization bias), in settings where we believe in sparsity assumptions. Another possible source of bias arises from the danger of overfitting. This is a problem when the nuisance functions are too complex (e.g., cannot be assumed to belong to a Donsker class; e.g., Díaz [20]; Kennedy [32]). A general solution to avoid overfitting error, and thereby obtain valid inference, is to use sample-splitting; see, [7, 11, 12, 62].

2.4 Advances in Recent Years

Recent years have witnessed a great deal of novel results in the field of uniformly valid causal inference. [13] readdressed valid inference for the average causal effect under the framework of double/debiased machine learners in light of the fact that the parameter τ satisfies a Neyman orthogonal moment condition [36, 37]; a moment condition that is not sensitive to local errors in nuisance models and can be derived using the first order influence function of the parameter [8, 55]. They suggest sample splitting which together with the above Neyman orthogonality leads to ignorable remainder terms even when machine learning nuisance estimators are converging relatively slowly. Other works have considered a Neyman orthogonal estimating equation of a l^2-continuous functional of a conditional expectation [9, 15]. In this setting, both the conditional expectation and the Reisz representer of the functional, which is the inverse propensity score in the case of τ, must be estimated. However, in the latter case, the estimation of the inverse propensity score is performed differently compared to Chernozhukov et al. [13]; that is, using the equation which characterize the nuisance model as a Riesz representer.

The notion of double robustness of an estimator has been widely used to indicate that an estimator is consistent if at least one of the nuisance models is estimated consistently, not necessarily both [3, 41]. This property has also been called parametric double robustness [32]. As mentioned in the previous section, van der Laan ([57], Section 4) was the first work which addressed what they called double robust statistical inference (also called model double robustness in Smucler, Rotnitzky and Robins [49]), which indicates that inference on the parameter of interest requires consistent estimation of only one of the nuisance models. The AIPW estimator mentioned in the previous section based on regularized maximum likelihood nuisance estimators has the nonparametric double robust property [32], also called rate double robustness in Smucler, Rotnitzky and Robins [49], whereby weak consistency of the nuisance models is required for uniformly valid inference, but slower convergence in estimating the propensity score can be bought out by faster convergence in the outcome models, and vice versa (9). However, this AIPW estimator does not yield double robust statistical inference. Alternative loss functions have been considered in the estimation of nuisance models, which endow the AIPW estimator double robust statistical inference (e.g., [2, 9, 38, 52]). Smucler, Rotnitzky and Robins [49] extends this area of work by generalizing the property of double robust inference to the estimation of all parameters that belong to what is called the class of bilinear influence function (BIF) functionals. These estimators are specific to sparse settings and employ l_1-regularized estimators. Moreover, they consider a sparsity condition even for the limit of a possibly inconsistent nuisance model estimator.

The BIF class includes important causal parameters such as the average treatment effect and the average treatment effect among treated and cover the classes of parameters studied in Chernozhukov, Newey and Singh [15] and Robins et al. [40]. However, the parameters who enjoy a uniform valid inference are not limited to this class. For example, the continuously differentiable functions of the functionals with bilinear influence function do not belong to the class while validity of inference for those can be directly shown by the delta method [49]. A different technique has been used for constructing a valid confidence interval for a parameter outside the BIF class, a conditional average treatment effect, which is to invert a chi-squared distributed double robust test statistic [22].

Most of the above literature on uniformly valid causal inference is concerned with high-dimensional settings, which can arise both because of a large set of covariates is available, but also because functions/transformations of these covariates are considered in generalized linear (in the parameters) nuisance models with a sparsity property. However, alternative regularity conditions, for example, smoothness, may be considered attractive. For example, neural networks can be used for smooth nuisance functions belonging to Sobolev spaces [25]. Some of the results in the above-mentioned papers are not specific to l_1-regularized estimators and apply to any well-behaved nonparametric nuisance model estimation [13, 24, 59]. In van der Laan and Rubin [59] the choice of a single nonparametric estimator is not considered to be done apriori, but a data-adaptive cross-validated combination of a set of estimators (super learner, ensemble learner) was suggested. Finally, more recently, Cui and Tchetgen Tchetgen [16] suggest two novel selection criteria, where the main focus is on getting smaller bias in the estimation of the target parameter instead of the nuisance ones.

While the literature has focused on situations where the parameter (a causal effect) is of low dimension, Semenova and Chernozhukov [47] recently addressed situations, where the Neyman orthogonal property can be applied to obtain uniformly valid results in nonparametric situations, that is, where the parameter of interest is of infinite dimension. Examples include causal effects conditional on a continuous covariate and causal effects of a continuous valued treatment.

3. COST OF UNIFORMITY AND A DOUBLE-SELECTION OUTCOME REGRESSION ESTIMATOR

One essential component of estimators with uniformly valid inference reviewed in the latter section is that if there are instruments — here covariates that explain T although they are not related to Y conditional on the other covariates included in $\boldsymbol{X}$ — these may be part of the selected set of covariates. This is also obviously true for propensity score centered methods; see, for example, Shortreed and Ertefaie [48]. This is unfortunate because the semiparametric efficiency bound for the average treatment effect is lower if we have knowledge on which variables are instruments [19, 28, 29, 42, 53]. The variance inflation due to instruments can be severe and this has been reported in the literature numerous times, see, Schnitzer, Lok and Gruber [46] and references therein.

In the sequel, we provide a discussion and results which shed new light on this issue by presenting an estimation strategy which seems to yield a compromise between the estimators for which we have a uniformly valid asymptotic distribution (using instruments) and superefficient estimators, where irrelevant instruments are selected away by using the data.

For simplicity, consider as parameter of interest $\tau_1 = E(Y(1))$. However, the results for $\tau_0 = E(Y(0))$ and thereby $\tau = \tau_1 - \tau_0$ are analogous. Let $\boldsymbol{x} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_n]'$, $\boldsymbol{y} = [y_1, \dots, y_n]'$, superscript T denotes subsetting rows that correspond to treated individuals and subscript S denotes subsetting columns using the set S. Then, $P_S^T = \boldsymbol{x}_S^T(\boldsymbol{x}_S^{T\prime}\boldsymbol{x}_S^T)^{-1}\boldsymbol{x}_S^{T\prime}$ is the projection matrix onto the space spanned by $\boldsymbol{x}_S^T$, while $\tilde{P}_S^T = \boldsymbol{x}_S(\boldsymbol{x}_S^{T\prime}\boldsymbol{x}_S^T)^{-1}\boldsymbol{x}_S^{T\prime}$ is the matrix used in predicting $Y(1)$ for all n individuals. Given a selected covariate set S, we define the post-selection outcome regression (OR) estimator as

$$\begin{aligned}\hat{\tau}_{1,\mathrm{OR}}(S) &= E_n\big[\big(\tilde{P}_S^T \boldsymbol{y}^T\big)_i\big]\\ &= \frac{1}{n}\Sigma_{i=1}^n\big[\boldsymbol{x}_{S,i}\big(\boldsymbol{x}_S^{T\prime}\boldsymbol{x}_S^T\big)^{-1}\boldsymbol{x}_S^{T\prime}\boldsymbol{y}^T\big]\\ &:= \frac{1}{n}\Sigma_{i=1}^n\big[\hat{m}_1(\boldsymbol{x}_{S,i})\big].\end{aligned}$$

A classical post-selection OR estimator is $\hat{\tau}_{1,\mathrm{OR}_{XY}} = \hat{\tau}_{1,\mathrm{OR}}(X_Y)$, where X_Y is a set of covariates derived from the $y - \boldsymbol{x}$ association using any covariate selection strategy; for instance, the set of covariates that corresponds to nonzero coefficients in a fitted lasso regression of y versus $\boldsymbol{x}$ may be considered. Instead, we study here theoretically the post-double-selection OR estimator

$$\hat{\tau}_{1,\mathrm{OR}_{DS}} = \hat{\tau}_{1,\mathrm{OR}}(X_Y \cup X_T), \tag{10}$$

where X_T is derived by fitting the $t - \boldsymbol{x}$ association using lasso or any other covariate selection strategy. Note that this can be considered as a generalization of [5] early estimator for the homogeneous case presented in Section 2.2 to the general situation of an heterogeneous treatment effect. Although this estimator is mentioned in the simulation experiments run in [1], no theoretical results are available in the literature up to our knowledge.

Estimator (10) is not asymptotically linear and cannot be shown to yield uniformly valid inference as was the case for the post-selection double robust estimators described in Section 2.3. However, uniform fast rate of decay of the bias can be guaranteed. The conditions used below are of the type used to show uniform validity. In particular, a product convergence rate condition is used.

THEOREM 3.1. *Suppose*

(i)

$$\begin{aligned}&E_n\big[\big(1 - t_i a(\boldsymbol{x}_i)\big)\big(\tilde{P}_S^T m_1(\boldsymbol{x}^T) - m_1(\boldsymbol{x})\big)_i\big]\\ &\quad = o_{P_n}\big(n^{-\upsilon_1}\big),\end{aligned}$$

(ii)

$$\begin{aligned}&E_{n_t}\big[\big(\big(\mathbb{1}_{n_t \times n_t} - P_S^T\big)a\big(\boldsymbol{x}^T\big)\big)_i^2\big]^{1/2}\\ &\qquad \times E_{n_t}\big[\big(\big(\mathbb{1}_{n_t \times n_t} - P_S^T\big)m_1\big(\boldsymbol{x}^T\big)\big)_i^2\big]^{1/2}\\ &\quad = o_p\big(n^{-\upsilon_2}\big),\end{aligned}$$

where $\mathbb{1}_{n_t \times n_t}$ *is the identity matrix of size* n_t, $a(\boldsymbol{X}) = 1/\mathbb{P}(T = 1|\boldsymbol{X})$. *Let* $\upsilon = \min(\upsilon_1, \upsilon_2)$. *Then,*

$$\mathrm{Bias}(\hat{\tau}_{1,\mathrm{OR}}) = E\big(E_n\big[\hat{m}_1(\boldsymbol{x}_{S,i})\big] - \tau_1\big) = o\big(n^{-\upsilon}\big).$$

The first condition (i) requires that the order of the scaled error for all the individuals is equal to the order of the scaled error on the treated individuals weighted by inverse propensity scores, similar to Farrell [24], Assumption 3(c). The second condition (ii) is a multiplicative rate condition similar to the multiplicative rate condition in Farrell [24], Assumption 3(b). However, notice that here the rate must be fulfilled using the same set of covariates in both nuisance models. This necessitates doing double selection to get the double robustness property in terms of bias. The proof of Theorem 3.1 can be found in Appendix A.

To illustrate how a double-selection procedure can benefit in terms of bias, suppose

$$m_1(\boldsymbol{X}, \eta_1) = \eta_1' \boldsymbol{X} + R_Y, \tag{11}$$

and

$$a(\boldsymbol{X}, \gamma) = \gamma' \boldsymbol{X} + R_T. \tag{12}$$

Here, we consider a linear model for the inverse of propensity score model (e.g., [30]), and R_Y and R_t are approximation errors of the true models with sparse coefficients η_1 and γ in the outcome and inverse propensity score models, respectively. The following corollary is a direct consequence of Theorem 3.1 and the asymptotic results for lasso regression in Farrell [24], Section 6.

COROLLARY 1. *Suppose that models* (11) *and* (12), *hold, where the conditions* (5), (6), (7) *and* (8) *are fulfilled. Moreover, assume the regularity conditions in Farrell* [24], *Corollary* 5 *and Appendix F.3. Consider* $\hat{\tau}_{1,\mathrm{OR}_{DS}}$ *in* (10) *where* X_Y *and* X_T *are estimated using lasso regression of the observed potential outcome* (*for treated individuals*) *and the inverse propensity score on* X, *respectively. Then*,

$$\mathrm{Bias}(\hat{\tau}_{1,\mathrm{OR}_{DS}}) = o(n^{-1/2}).$$

The above result shows that root-n decay of the bias can be derived uniformly over a set of DGPs. In this sense, the double-selection OR estimator may be seen as a compromise between single selection estimators (superefficient and no uniformly decaying bias) and the double-selection estimators of Section 2 (with available uniformly valid asymptotic distribution). The implications in terms of finite sample behaviour are studied below with Monte Carlo experiments.

In practice, the inverse propensity score a is not observed and (12) cannot be fitted directly. Instead, we suggest to fit a lasso logistic regression for the propensity score to retrieve the relevant covariates.

4. SIMULATION STUDY

The aim of this simulation study is to illustrate the above theoretical discussion on the cost and benefits of uniformly valid inference. While many simulations studies are available in the literature reviewed above, their focus is either to show the necessity of using uniformly valid procedures in order to avoid regularization bias, or to illustrate the variance inflation due to the use of instruments in the propensity score compared to superefficient procedures. Here, we contrast these two aspects by considering both uniformly valid and superefficient post-selection strategies, as well as the double-selection outcome regression estimator (10).

4.1 Simulation Design

We use 500 replicates in all situations, and consider sample sizes $n = 500$, 1000, 1500, 2000. The covariate vector $\boldsymbol{X}$ is generated from a multivariate normal distribution with zero mean and identity covariance matrix. The dimension p of the covariate vector equals n. Results for low dimensional settings, $p \ll n$, portray a similar general picture and are available from the authors upon request. Data generation and all computations are performed with the software R [39].

4.1.1 *High-dimensional setting.* We consider the following models:

$$\begin{aligned} Y(0) &= m_0(\boldsymbol{X}) + \epsilon_0 = 1 + \eta_0' \boldsymbol{X} + \epsilon_0, \\ Y(1) &= m_1(\boldsymbol{X}) + \epsilon_1 = 2 + \eta_1' \boldsymbol{X} + \epsilon_1, \\ \mathbb{P}(T = 1 | \boldsymbol{X}) &= e(\boldsymbol{X}) = \mathrm{expit}(\gamma' \boldsymbol{X}), \end{aligned}$$

indexed by the parameter vectors

$$\begin{aligned} \eta_0 = \frac{k}{2} \cdot (&1, 1/2, 1/3, 1/4, 1/5, \\ &0, 0, 0, 0, 0, \\ &1, 1/2, 1/3, 1/4, 1/5, \\ &0, \dots, 0), \\ \eta_1 = k \cdot (&1, 1/2, 1/3, 1/4, 1/5, \\ &0, 0, 0, 0, 0, \\ &1, 1/2, 1/3, 1/4, 1/5, \\ &0, \dots, 0), \\ \gamma = (&1, 1/2, 1/3, 1/4, 1/5, \\ &1, 1, 1, 1, 1, \\ &0, \dots, 0), \end{aligned}$$

where $k \in \{0.1, 0.4, 0.8, 1.2\}$ and the error terms, $\epsilon_t, t = 0, 1$, are generated from a normal distribution with $E(\epsilon_t | \boldsymbol{X}) = 0$ and $\mathrm{Var}(\epsilon_t | \boldsymbol{X}) = (1 + p)^{-1}(1 + \iota' \cdot \boldsymbol{X}^2)$ where ι is the vector of ones. The parameter k determines the strength of the association between the outcome and the covariates.

4.1.2 *Post-selection estimators of* τ. We use lasso as implemented in the R package hdm [14] to estimate the nuisance models $m_t(\boldsymbol{X})$ and $e(\boldsymbol{X})$. We denote the sets of variables which corresponds to nonzero coefficients in the estimated sparse linear outcome models and logistic propensity score by $X_Y = X_{Y_0} \cup X_{Y_1}$, the union of the sets estimated by each of the two potential outcome models, and X_T, respectively. The lasso penalty parameter is selected as $\lambda = 2.2\sqrt{n}\Phi^{-1}([\log(n) - 0.1][2p\log(n)]^{-1})$ [4].

With the purpose of estimating τ, using different combinations of the above covariate sets, we compare two versions of the OR estimator, three versions of the doubly robust AIPW estimator and one version of the doubly robust TMLE estimator. Specifically, OR$_{X_Y}$ and OR$_{DS}$ use X_Y and the union $X_T \cup X_Y$ in the outcome models refitting steps, respectively. AIPW$_{X_Y}$ uses X_Y in the propensity score and outcome model refitting steps, AIPW$_{DS}$ uses the union $X_T \cup X_Y$ in both the propensity score and outcome models (as in [5], Section 5) and AIPW$_{X_Y,X_T}$ uses X_T in the propensity score model and X_Y in the outcome models (as in [24]). TMLE$_{X_Y,X_T}$ uses X_T in the propensity score model and X_Y in the outcome models.

Given the same consistency and product rate conditions on the initial estimators of nuisance models as in Farrell [24], TMLE$_{X_Y,X_T}$ has the same uniformly valid asymptotic distribution as AIPW$_{X_Y,X_T}$ ([58], Chapter 27). Hence, if the refitted models used in AIPW$_{X_Y,X_T}$ are used as initial model estimates in TMLE$_{X_Y,X_T}$ we would expect similar results for large samples. In summary, AIPW$_{X_Y,X_T}$, AIPW$_{DS}$ and TMLE$_{X_Y,X_T}$ have uniformly valid asymptotic distributions, OR$_{X_Y}$ and AIPW$_{X_Y}$ have no such uniform validity, select away instruments and are superefficient, while OR$_{DS}$ has uniformly decaying bias (Corollary 1). For TMLE, we use the R package `tmle` [27] and do not truncate the estimated propensity scores, that is, `gbound = c(0,1)`. For the other estimators, we use own written R code as well as the R package `ui` [26] for the variances.

4.2 Results

From 500 replicates, we compute empirical biases, standard errors, root mean squared errors (RMSE), and empirical coverages. We also compute mean estimated standard errors. Table 1 presents results for $k = 0.4$. Results for all values of k are given for bias, RMSE and coverages in the Appendix, Tables 2–4.

We see that estimators selecting away instruments (OR$_{X_Y}$ and AIPW$_{X_Y}$) have lower Monte Carlo standard error but at the cost of larger bias and poor empirical coverages (clearly lower than nominal level). As expected the other estimators, which all have uniformly decaying bias, show low bias, but at the cost of larger standard error. This cost is, however, smallest for OR$_{DS}$, and AIPW$_{X_Y,X_T}$, AIPW$_{DS}$ and TMLE$_{X_Y,X_T}$ have standard errors roughly up to five times as large as OR$_{X_Y}$, while for OR$_{DS}$ the increase in standard error is not as severe (up to 1.22 times the standard error of OR$_{X_Y}$). All the low-bias estimators have good empirical coverages, at least for sample sizes 1000 and higher, although we do not have such theoretical guarantee for OR$_{DS}$.

On a side note, we observe that the estimated standard errors of AIPW$_{X_Y,X_T}$, AIPW$_{DS}$ and TMLE$_{X_Y,X_T}$ are distinctively smaller than the Monte Carlo standard errors, but the difference is reduced when increasing sample size. For OR$_{DS}$ no such underestimation of the variance is observed. Bootstrap estimation of the standard errors might remedy the finite sample underestimation we see here. Cai and van der Laan [10] recently proposed a consistent bootstrap method for TMLE, but the validity of bootstrap methods for post-selection estimators in general is yet to be investigated.

TABLE 1

Results of 500 *simulation replicates for estimators of* τ, *for varying sample sizes* n, *number of covariates* $p = n$ *and* $k = 0.4$. *RMSE, root mean-squared error*; *Bias*; *SE, Monte Carlo standard deviation*; *ESE, estimated standard error* (*influence curve based estimates*, *ignoring the variability in the selection step*); *CP, empirical coverage probability of* 95% *confidence intervals*

n	Estimator	RMSE	Bias	SE	ESE	CP
500	OR$_{X_Y}$	0.26	−0.19	0.18	0.16	0.74
	AIPW$_{X_Y}$	0.26	−0.19	0.18	0.16	0.74
	OR$_{DS}$	0.22	−0.05	0.22	0.21	0.94
	AIPW$_{X_Y,X_T}$	0.95	−0.06	0.95	0.35	0.93
	AIPW$_{DS}$	0.88	−0.06	0.88	0.34	0.91
	TMLE$_{X_Y,X_T}$	0.29	−0.08	0.28	0.24	0.91
1000	OR$_{X_Y}$	0.19	−0.14	0.14	0.11	0.72
	AIPW$_{X_Y}$	0.19	−0.14	0.14	0.11	0.73
	OR$_{DS}$	0.16	−0.02	0.16	0.15	0.94
	AIPW$_{X_Y,X_T}$	0.38	−0.00	0.38	0.25	0.97
	AIPW$_{DS}$	0.37	−0.00	0.37	0.25	0.95
	TMLE$_{X_Y,X_T}$	0.21	−0.04	0.20	0.18	0.95
1500	OR$_{X_Y}$	0.15	−0.11	0.11	0.09	0.76
	AIPW$_{X_Y}$	0.15	−0.11	0.11	0.09	0.76
	OR$_{DS}$	0.12	−0.02	0.12	0.12	0.95
	AIPW$_{X_Y,X_T}$	0.22	−0.03	0.22	0.18	0.95
	AIPW$_{DS}$	0.24	−0.02	0.24	0.18	0.95
	TMLE$_{X_Y,X_T}$	0.15	−0.04	0.15	0.14	0.92
2000	OR$_{X_Y}$	0.13	−0.09	0.10	0.08	0.80
	AIPW$_{X_Y}$	0.13	−0.09	0.10	0.08	0.80
	OR$_{DS}$	0.10	−0.01	0.10	0.10	0.95
	AIPW$_{X_Y,X_T}$	0.19	−0.02	0.19	0.16	0.94
	AIPW$_{DS}$	0.19	−0.02	0.19	0.16	0.93
	TMLE$_{X_Y,X_T}$	0.15	−0.03	0.15	0.13	0.92

5. DISCUSSION

To obtain an unbiased estimator of the average causal effect of a treatment, we need to control for all confounders. Knowing that some covariates are related to the outcome, but not to the treatment does not change the semiparametric efficiency bound. However, knowledge on the existence of instruments has implications on the asymptotic variance that can be achieved by unbiased estimators [19, 29, 42]. In practice, we typically do not have such apriori (to data) knowledge. Naively selecting away covariates (here instruments) using the data at hand, by, for example, regularization yields inference that is not uniformly valid (may translate into large bias and incorrect coverage rates). On the other hand, using methods which yield uniformly valid inference may yield large variability compared to these naive (often superefficient) methods. In this paper, we have reviewed the literature on uniformly valid causal inference, and have discussed the costs and benefits of uniformly valid inference. The latter discussion has been illustrated by studying a double-selection outcome regression estimator, which is shown to be uniformly asymptotically unbiased under a product rate condition. This seems to translate into finite sample properties which are a compromise between uniformly valid and superefficient estimators. The good properties of the double-selection OR estimator may arguably be due to the designs considered in our simulations, for which the outcome regression models can be consistently fitted. Consistency of the outcome regression model is indeed assumed to show uniformly decaying bias of the double-selection estimator. This assumption is also used by estimators which have the nonparametric double robust property. On the other hand, procedures that yield double robust statistical inference [2, 52, 57], allow for one of the nuisance models to be inconsistent (converges to an incorrect function of the covariates). This alternative limit need, however, to be assumed sparse in the covariates, a rather high-level assumption [49].

APPENDIX A: PROOF OF THEOREM 3.1

The bias of post selection outcome regression estimator can be expressed as follows:

$$
\begin{aligned}
&\mathrm{Bias}(\hat{\tau}_{1,OR})\\
&\quad = E\big(E_n[\hat{m}_1(\boldsymbol{x}_{S,i})] - \tau_1\big)\\
&\quad = E\big(E_n[\hat{m}_1(\boldsymbol{x}_{S,i}) - m_1(\boldsymbol{x}_i)]\big) + E\big(E_n[m_1(\boldsymbol{x}_i)] - \tau_1\big)\\
&\quad = E\big(E_n[\hat{m}_1(\boldsymbol{x}_{S,i}) - m_1(\boldsymbol{x}_i)]\big),
\end{aligned}
$$

where the last equality follows by Assumption 1. To show that the scaled bias term is asymptotically negligible we show that it is negligible conditional on S and $\{t_i, x_i\}_{i=1}^n$:

$$
\begin{aligned}
&n^v \,\mathrm{Bias}(\hat{\tau}_{1,\mathrm{OR}}|S, \{t_i, x_i\}_{i=1}^n)\\
&\quad := n^v E\big(E_n[\hat{m}_1(\boldsymbol{x}_{S,i}) - m_1(\boldsymbol{x}_i)]|S, \{t_i, x_i\}_{i=1}^n\big)\\
&\quad = n^v E_n\big[\boldsymbol{x}_{S,i}(\boldsymbol{x}_S^{T\prime}\boldsymbol{x}_S^T)^{-1}\boldsymbol{x}_S^{T\prime} m_1(\boldsymbol{x}^T) - m_1(\boldsymbol{x}_i)\big]\\
&\quad \approx n^v E_n\big[t_i a(\boldsymbol{x}_i)(\boldsymbol{x}_{S,i}(\boldsymbol{x}_S^{T\prime}\boldsymbol{x}_S^T)^{-1}\boldsymbol{x}_S^{T\prime} m_1(\boldsymbol{x}^T)\\
&\qquad - m_1(\boldsymbol{x}_i))\big]\\
&\quad = \frac{n^v}{n} a(\boldsymbol{x}^T)'(P_S^T - \mathbb{1}_{n_t\times n_t}) m_1(\boldsymbol{x}^T)\\
&\quad = \frac{n^v}{n} ((P_S^T - \mathbb{1}_{n_t\times n_t}) a(\boldsymbol{x}^T))'(P_S^T - \mathbb{1}_{n_t\times n_t}) m_1(\boldsymbol{x}^T),
\end{aligned}
$$

where $\approx$ means that both side have the same limits which holds by 3.1(i):

$$
\begin{aligned}
&\big|n^v \,\mathrm{Bias}(\hat{\tau}_{1,\mathrm{OR}}|S, \{t_i, x_i\}_{i=1}^n)\big|\\
&\quad \le n^v E_{n_t}\big[\big(\big(P_S^T - \mathbb{1}_{n_t\times n_t}\big)a(\boldsymbol{x}^T)\big)_i^2\big]^{1/2}\\
&\qquad \times E_{n_t}\big[\big(\big(P_S^T - \mathbb{1}_{n_t\times n_t}\big)m_1(\boldsymbol{x}^T)\big)_i^2\big]^{1/2}\\
&\quad = o_{P_n}(1),
\end{aligned}
$$

where the last equality holds by 3.1(ii). The statement in the theorem follows by the above result on the order of decay of the conditional expectation and uniform integrability of this conditional expectation.

APPENDIX B: PROOF OF COROLLARY 1

By construction $X_Y \subseteq S$. Therefore by Farrell [24], Appendix F.3, (5) and the sparsity condition (7) we have $v_1 = 1/2$. Moreover, using Farrell [24], Corollary 5, $X_Y \subseteq S$, $X_T \subseteq S$, and conditions (5), (6) and (8) we have $v_2 = 1/2$.

APPENDIX C: SIMULATION RESULTS

TABLE 2

$\sqrt{n}$-bias based on 500 simulation replicates, for estimators of τ, with varying sample sizes n and values for k, and number of covariates $p = n$

		k			
n	Estimator	0.1	0.4	0.8	1.2
500	OR_{X_Y}	−1.25	−4.16	−3.44	−2.97
	AIPW_{X_Y}	−1.25	−4.17	−3.44	−2.96
	OR_{DS}	−0.26	−1.08	−2.09	−2.56
	AIPW_{X_Y,X_T}	−0.42	−1.26	−2.57	−3.12
	AIPW_{DS}	−0.35	−1.45	−2.49	−2.44
	TMLE_{X_Y,X_T}	−0.45	−1.79	−2.55	−2.84
1000	OR_{X_Y}	−1.81	−4.32	−3.14	−2.28
	AIPW_{X_Y}	−1.81	−4.31	−3.15	−2.27
	OR_{DS}	−0.17	−0.76	−1.40	−1.63
	AIPW_{X_Y,X_T}	0.59	−0.06	−0.86	−1.12
	AIPW_{DS}	0.56	−0.03	−0.88	−1.08
	TMLE_{X_Y,X_T}	−0.30	−1.23	−1.76	−1.83
1500	OR_{X_Y}	−2.50	−4.10	−3.20	−2.08
	AIPW_{X_Y}	−2.50	−4.09	−3.21	−2.08
	OR_{DS}	−0.50	−0.91	−1.44	−1.47
	AIPW_{X_Y,X_T}	−0.66	−1.02	−1.36	−1.55
	AIPW_{DS}	−0.42	−0.89	−1.39	−1.48
	TMLE_{X_Y,X_T}	−0.75	−1.38	−1.71	−1.59
2000	OR_{X_Y}	−2.83	−3.91	−2.91	−1.56
	AIPW_{X_Y}	−2.82	−3.92	−2.89	−1.53
	OR_{DS}	−0.29	−0.60	−0.91	−0.87
	AIPW_{X_Y,X_T}	−0.54	−0.91	−1.21	−1.02
	AIPW_{DS}	−0.44	−0.81	−1.08	−0.85
	TMLE_{X_Y,X_T}	−0.68	−1.15	−1.33	−1.04

TABLE 3

RMSE based on 500 *simulation replicates, for estimators of* τ, *with varying sample sizes* n *and values for* k, *and number of covariates* $p = n$

		k			
n	Estimator	0.1	0.4	0.8	1.2
500	OR_{X_Y}	0.17	0.26	0.25	0.24
	AIPW_{X_Y}	0.17	0.26	0.25	0.24
	OR_{DS}	0.21	0.22	0.24	0.26
	AIPW_{X_Y,X_T}	0.85	0.95	0.95	0.88
	AIPW_{DS}	0.79	0.88	0.88	0.76
	TMLE_{X_Y,X_T}	0.28	0.29	0.31	0.32
1000	OR_{X_Y}	0.12	0.19	0.17	0.15
	AIPW_{X_Y}	0.12	0.19	0.17	0.15
	OR_{DS}	0.15	0.16	0.17	0.17
	AIPW_{X_Y,X_T}	0.37	0.38	0.37	0.38
	AIPW_{DS}	0.36	0.37	0.38	0.38
	TMLE_{X_Y,X_T}	0.20	0.21	0.22	0.22
1500	OR_{X_Y}	0.11	0.15	0.13	0.11
	AIPW_{X_Y}	0.11	0.15	0.13	0.11
	OR_{DS}	0.12	0.12	0.13	0.13
	AIPW_{X_Y,X_T}	0.22	0.22	0.24	0.24
	AIPW_{DS}	0.22	0.24	0.25	0.25
	TMLE_{X_Y,X_T}	0.15	0.15	0.16	0.16
2000	OR_{X_Y}	0.10	0.13	0.11	0.10
	AIPW_{X_Y}	0.10	0.13	0.11	0.10
	OR_{DS}	0.10	0.10	0.11	0.11
	AIPW_{X_Y,X_T}	0.18	0.19	0.19	0.19
	AIPW_{DS}	0.18	0.19	0.19	0.19
	TMLE_{X_Y,X_T}	0.15	0.15	0.15	0.15

TABLE 4

Empirical coverage probability of 95% *confidence intervals based on* 500 *simulation replicates, for estimators of* τ, *with varying sample sizes* n *and values for* k, *and number of covariates* $p = n$

		k			
n	Estimator	0.1	0.4	0.8	1.2
500	OR_{X_Y}	0.94	0.74	0.83	0.86
	AIPW_{X_Y}	0.94	0.74	0.82	0.87
	OR_{DS}	0.93	0.94	0.92	0.92
	AIPW_{X_Y,X_T}	0.95	0.93	0.92	0.91
	AIPW_{DS}	0.92	0.91	0.90	0.90
	TMLE_{X_Y,X_T}	0.93	0.91	0.89	0.90
1000	OR_{X_Y}	0.93	0.72	0.84	0.89
	AIPW_{X_Y}	0.93	0.73	0.84	0.90
	OR_{DS}	0.93	0.94	0.92	0.92
	AIPW_{X_Y,X_T}	0.97	0.97	0.95	0.95
	AIPW_{DS}	0.96	0.95	0.94	0.93
	TMLE_{X_Y,X_T}	0.95	0.95	0.94	0.94
1500	OR_{X_Y}	0.90	0.76	0.83	0.91
	AIPW_{X_Y}	0.90	0.76	0.84	0.92
	OR_{DS}	0.96	0.95	0.96	0.95
	AIPW_{X_Y,X_T}	0.96	0.95	0.95	0.94
	AIPW_{DS}	0.95	0.95	0.94	0.93
	TMLE_{X_Y,X_T}	0.95	0.92	0.93	0.92
2000	OR_{X_Y}	0.88	0.80	0.86	0.92
	AIPW_{X_Y}	0.88	0.80	0.87	0.92
	OR_{DS}	0.95	0.95	0.94	0.95
	AIPW_{X_Y,X_T}	0.94	0.94	0.94	0.95
	AIPW_{DS}	0.94	0.93	0.94	0.94
	TMLE_{X_Y,X_T}	0.92	0.92	0.92	0.93

ACKNOWLEDGMENTS

We acknowledge comments from Tetiana Gorbach, Mohammad Ghasempour and two anonymous referees that have improved the paper.

FUNDING

This work was supported by a grant from the Marianne and Marcus Wallenberg Foundation.

REFERENCES

[1] ATHEY, S., IMBENS, G. W. and WAGER, S. (2018). Approximate residual balancing: Debiased inference of average treatment effects in high dimensions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 597–623. MR3849336 https://doi.org/10.1111/rssb.12268

[2] AVAGYAN, V. and VANSTEELANDT, S. (2021). High-dimensional inference for the average treatment effect under model misspecification using penalized bias-reduced double-robust estimation. *Biostatistics & Epidemiology* 1–18. https://doi.org/10.1080/24709360.2021.1898730

[3] BANG, H. and ROBINS, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61** 962–972. MR2216189 https://doi.org/10.1111/j.1541-0420.2005.00377.x

[4] BELLONI, A., CHEN, D., CHERNOZHUKOV, V. and HANSEN, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica* **80** 2369–2429. MR3001131 https://doi.org/10.3982/ECTA9626

[5] BELLONI, A., CHERNOZHUKOV, V. and HANSEN, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Rev. Econ. Stud.* **81** 608–650. MR3207983 https://doi.org/10.1093/restud/rdt044

[6] BERK, R., BROWN, L., BUJA, A., ZHANG, K. and ZHAO, L. (2013). Valid post-selection inference. *Ann. Statist.* **41** 802–837. MR3099122 https://doi.org/10.1214/12-AOS1077

[7] BICKEL, P. J. (1982). On adaptive estimation. *Ann. Statist.* **10** 647–671. MR0663424

[8] BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, New York. Reprint of the 1993 original. MR1623559

[9] BRADIC, J., WAGER, S. and ZHU, Y. (2019). Sparsity double robust inference of average treatment effects. ArXiv preprint. Available at arXiv:1905.00744.

[10] CAI, W. and VAN DER LAAN, M. (2019). Nonparametric bootstrap inference for the targeted highly adaptive LASSO estimator. ArXiv preprint. Available at arXiv:1905.10299.

[11] CATTANEO, M. D., JANSSON, M. and MA, X. (2019). Two-step estimation and inference with possibly many included covariates. *Rev. Econ. Stud.* **86** 1095–1122. MR3945564 https://doi.org/10.1093/restud/rdy053

[12] CATTANEO, M. D., JANSSON, M. and NEWEY, W. K. (2018). Inference in linear regression models with many covariates and heteroscedasticity. *J. Amer. Statist. Assoc.* **113** 1350–1361. MR3862362 https://doi.org/10.1080/01621459.2017.1328360

[13] 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. *Econom. J.* **21** C1–C68. MR3769544 https://doi.org/10.1111/ectj.12097

[14] CHERNOZHUKOV, V., HANSEN, C. and SPINDLER, M. (2016). hdm: High-dimensional metrics. ArXiv preprint. Available at arXiv:1608.00354.

[15] CHERNOZHUKOV, V., NEWEY, W. and SINGH, R. (2020). De-Biased Machine Learning of Global and Local Parameters Using Regularized Riesz Representers.

[16] CUI, Y. and TCHETGEN TCHETGEN, E. (2019). Selective machine learning of doubly robust functionals. ArXiv preprint. Available at arXiv:1911.02029.

[17] D'AMOUR, A., DING, P., FELLER, A., LEI, L. and SEKHON, J. (2021). Overlap in observational studies with high-dimensional covariates. *J. Econometrics* **221** 644–654. MR4215042 https://doi.org/10.1016/j.jeconom.2019.10.014

[18] DE LUNA, X. and JOHANSSON, P. (2014). Testing for the unconfoundedness assumption using an instrumental assumption. *J. Causal Inference* **2** 187–199. MR4289420 https://doi.org/10.1515/jci-2013-0011

[19] DE LUNA, X., WAERNBAUM, I. and RICHARDSON, T. S. (2011). Covariate selection for the nonparametric estimation of an average treatment effect. *Biometrika* **98** 861–875. MR2860329 https://doi.org/10.1093/biomet/asr041

[20] DÍAZ, I. (2020). Machine learning in the estimation of causal effects: Targeted minimum loss-based estimation and double/debiased machine learning. *Biostatistics* **21** 353–358. MR4132549 https://doi.org/10.1093/biostatistics/kxz042

[21] DÍAZ, I., LUEDTKE, A. R. and VAN DER LAAN, M. J. (2018). Sensitivity analysis. In *Targeted Learning in Data Science*. *Springer Ser. Statist.* 511–522. Springer, Cham. MR3820742

[22] DUKES, O. and VANSTEELANDT, S. (2021). Inference for treatment effect parameters in potentially misspecified high-dimensional models. *Biometrika* **108** 321–334. MR4259134 https://doi.org/10.1093/biomet/asaa071

[23] FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581 https://doi.org/10.1198/016214501753382273

[24] FARRELL, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *J. Econometrics* **189** 1–23. MR3397349 https://doi.org/10.1016/j.jeconom.2015.06.017

[25] FARRELL, M. H., LIANG, T. and MISRA, S. (2021). Deep neural networks for estimation and inference. *Econometrica* **89** 181–213. MR4220387 https://doi.org/10.3982/ecta16901

[26] GENBÄCK, M. and DE LUNA, X. (2019). Causal inference accounting for unobserved confounding after outcome regression and doubly robust estimation. *Biometrics* **75** 506–515. MR3999174 https://doi.org/10.1111/biom.13001

[27] GRUBER, S. and VAN DER LAAN, M. J. (2012). tmle: An R package for targeted maximum likelihood estimation. *J. Stat. Softw.* **51** 1–35. https://doi.org/doi:10.18637/jss.v051.i13

[28] HAHN, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* **66** 315–331. MR1612242 https://doi.org/10.2307/2998560

[29] HAHN, J. (2004). Functional restriction and efficiency in causal inference. *Rev. Econ. Stat.* **86** 73–76.

[30] IMBENS, G. W., NEWEY, W. K. and RIDDER, G. (2005). Mean-square-error calculations for average treatment effects. IEPR Working Paper 05.34.

[31] JAVANMARD, A. and MONTANARI, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.* **15** 2869–2909. MR3277152

[32] KENNEDY, E. H. (2016). Semiparametric theory and empirical processes in causal inference. In *Statistical Causal Inferences and Their Applications in Public Health Research* (H. He, P. Wu and D.-G. Chen, eds.) *ICSA Book Ser. Stat.* 141–167. Springer, Cham. MR3617956

[33] LEE, J. D., SUN, D. L., SUN, Y. and TAYLOR, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.* **44** 907–927. MR3485948 https://doi.org/10.1214/15-AOS1371

[34] LEEB, H. and PÖTSCHER, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory* **21** 21–59. MR2153856 https://doi.org/10.1017/S0266466605050036

[35] LEEB, H. and PÖTSCHER, B. M. (2008). Sparse estimators and the oracle property, or the return of Hodges' estimator. *J. Econometrics* **142** 201–211. MR2394290 https://doi.org/10.1016/j.jeconom.2007.05.017

[36] NEYMAN, J. (1959). Optimal asymptotic tests of composite statistical hypotheses. In *Probability and Statistics*: *The Harald Cramér Volume* (U. Grenander, ed.) 416–444. Wiley, New York. MR0112201

[37] NEYMAN, J. (1979). $C(\alpha)$ tests and their use. *Sankhyā Ser. A* **41** 1–21. MR0615037

[38] NING, Y., PENG, S. and IMAI, K. (2020). Robust estimation of causal effects via a high-dimensional covariate balancing propensity score. *Biometrika* **107** 533–554. MR4138975 https://doi.org/10.1093/biomet/asaa020

[39] R CORE TEAM (2019). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

[40] ROBINS, J., LI, L., TCHETGEN 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. IMS.

[41] ROBINS, J. M., ROTNITZKY, A. and ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89** 846–866. MR1294730

[42] ROTNITZKY, A. and SMUCLER, E. (2020). Efficient adjustment sets for population average causal treatment effect estimation in graphical models. *J. Mach. Learn. Res.* **21** Paper No. 188. MR4209474

[43] RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** 688.

[44] RUBIN, D. B. (1990). Formal modes of statistical inference for causal effects. *J. Statist. Plann. Inference* **25** 279–292.

[45] SCHARFSTEIN, D. O., ROTNITZKY, A. and ROBINS, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J. Amer. Statist. Assoc.* **94** 1096–1146. MR1731478 https://doi.org/10.2307/2669923

[46] SCHNITZER, M. E., LOK, J. J. and GRUBER, S. (2016). Variable selection for confounder control, flexible modeling and collaborative targeted minimum loss-based estimation in causal inference. *Int. J. Biostat.* **12** 97–115. MR3505689 https://doi.org/10.1515/ijb-2015-0017

[47] SEMENOVA, V. and CHERNOZHUKOV, V. (2021). Debiased machine learning of conditional average treatment effects and other causal functions. *Econom. J.* **24** 264–289. MR4281225 https://doi.org/10.1093/ectj/utaa027

[48] SHORTREED, S. M. and ERTEFAIE, A. (2017). Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics* **73** 1111–1122. MR3744525 https://doi.org/10.1111/biom.12679

[49] SMUCLER, E., ROTNITZKY, A. and ROBINS, J. M. (2019). A unifying approach for doubly-robust \ell_1 regularized estimation of causal contrasts. ArXiv preprint. Available at arXiv:1904.03737.

[50] SPLAWA-NEYMAN, J. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statist. Sci.* **5** 465–472. Translated from the Polish and edited by D. M. D'abrowska and T. P. Speed. MR1092986

[51] TAN, Z. (2007). Comment: Understanding OR, PS and DR [MR2420458]. *Statist. Sci.* **22** 560–568. MR2420461 https://doi.org/10.1214/07-STS227A

[52] TAN, Z. (2020). Model-assisted inference for treatment effects using regularized calibrated estimation with high-dimensional data. *Ann. Statist.* **48** 811–837. MR4102677 https://doi.org/10.1214/19-AOS1824

[53] TANG, D., KONG, D., PAN, W. and WANG, L. (2020). Outcome model free causal inference with ultra-high dimensional covariates. ArXiv preprint. Available at arXiv:2007.14190.

[54] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

[55] TSIATIS, A. A. (2006). *Semiparametric Theory and Missing Data. Springer Series in Statistics*. Springer, New York. MR2233926

[56] VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42** 1166–1202. MR3224285 https://doi.org/10.1214/14-AOS1221

[57] VAN DER LAAN, M. J. (2014). Targeted estimation of nuisance parameters to obtain valid statistical inference. *Int. J. Biostat.* **10** 29–57. MR3208072 https://doi.org/10.1515/ijb-2012-0038

[58] VAN DER LAAN, M. J. and ROSE, S. (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data. Springer Series in Statistics*. Springer, New York. MR2867111 https://doi.org/10.1007/978-1-4419-9782-1

[59] VAN DER LAAN, M. J. and RUBIN, D. (2006). Targeted maximum likelihood learning. *Int. J. Biostat.* **2** Art. 11. MR2306500 https://doi.org/10.2202/1557-4679.1043

[60] VAN DER VAART, A. W. (1997). Superefficiency. In *Festschrift for Lucien Le Cam* (D. Pollard, E. Torgersen and G. L. Yang, eds.) 397–410. Springer, New York. MR1462961

[61] ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 217–242. MR3153940 https://doi.org/10.1111/rssb.12026

[62] ZHENG, W. and VAN DER LAAN, M. J. (2011). Cross-validated targeted minimum-loss-based estimation. In *Targeted Learning. Springer Ser. Statist.* 459–474. Springer, New York. MR2867139 https://doi.org/10.1007/978-1-4419-9782-1_27