Biometrika (2021), **108**, 2, *pp.* 321–334
Printed in Great Britain
doi: 10.1093/biomet/asaa071
Advance Access publication 18 November 2020

Inference for treatment effect parameters in potentially misspecified high-dimensional models

By OLIVER DUKES and STIJN VANSTEELANDT

Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Krijgslaan 281 (S9), 9000 Ghent, Belgium

oliver.dukes@ugent.be stijn.vansteelandt@ugent.be

Summary

Eliminating the effect of confounding in observational studies typically involves fitting a model for an outcome adjusted for covariates. When, as often, these covariates are high-dimensional, this necessitates the use of sparse estimators, such as the lasso, or other regularization approaches. Naïve use of such estimators yields confidence intervals for the conditional treatment effect parameter that are not uniformly valid. Moreover, as the number of covariates grows with the sample size, correctly specifying a model for the outcome is nontrivial. In this article we deal with both of these concerns simultaneously, obtaining confidence intervals for conditional treatment effects that are uniformly valid, regardless of whether the outcome model is correct. This is done by incorporating an additional model for the treatment selection mechanism. When both models are correctly specified, we can weaken the standard conditions on model sparsity. Our procedure extends to multivariate treatment effect parameters and complex longitudinal settings.

Some key words: Causal inference; Doubly robust estimation; High-dimensional inference; Post-selection inference.

1. Introduction

This article focuses on the problem of constructing confidence intervals for a low-dimensional component in a high-dimensional conditional mean model. In epidemiological studies, this component may correspond to the effect of a discrete-valued exposure A on an outcome Y, conditional on a set of baseline covariates L. When the dimension of the covariates is large relative to the sample size, data-adaptive model selection methods such as the lasso are typically used to select a final regression model, on the basis of which inference on the conditional treatment effect is performed. However, standard inferential techniques ignore the additional uncertainty induced by the selection process. A more serious issue is that they can also fail to be uniformly valid; there may be no sample size at which a given procedure is guaranteed to attain its nominal coverage or size. In particular, the treatment effect estimator may have a complex, nonnormal distribution due to uncertainty in the model selection step, even when the sample size is very large. The series of models considered may also fail to contain the true model for the conditional mean of Y given A and L.

Broadly speaking, there have been two main approaches to obtaining valid inferences on a low-dimensional parameter that depends on a high-dimensional regression adjustment. The first is based on doubly robust estimating equations (Robins et al., 1994), where a working propensity score model is postulated for the treatment selection mechanism and for the outcome. Doubly robust estimators are unbiased when at least one of these working models is correctly specified.

It was originally proposed by van der Laan & Rubin (2006) to combine this framework with flexible data-adaptive estimation of nuisance parameters. In the context of marginal treatment effects, Farrell (2015) showed that uniformly valid inferences can be obtained in high-dimensional settings by fitting both working models using the lasso. If sample splitting is used, uniformly valid inferences are available under simple and generic conditions (Chernozhukov et al., 2018). Essentially, the predictions from each regression must converge to the truth, and the product of the ℓ_2-norms of the prediction errors must shrink as $o_{\mathrm{p}}(n^{-1/2})$. In high-dimensional parametric models, the latter condition requires that the product of the number of nonzero coefficients in each model be small relative to the sample size.

The second strand of work focuses instead on estimating the target parameter by debiasing or desparsifying an initial lasso-based estimate (van de Geer et al., 2014) or score equation (Ning & Liu, 2017). In this approach the bias in the penalized estimator is typically corrected via a single iteration of a Newton–Raphson-style scheme. The bias correction term also depends on an additional regression adjustment, but one which does not necessarily correspond to a meaningful model for treatment selection. Instead, its role is purely to mitigate the bias incurred by estimating the parameter via the lasso; after the debiasing, under certain conditions the updated estimator is uniformly consistent and asymptotically normal. The doubly robust and debiased approaches may coincide in certain problems and are available in considerable generality.

In this article we show how to construct confidence intervals for parameters in high-dimensional linear and log-linear regression models which are uniformly valid, regardless of whether the outcome model is correctly specified. This is achieved by using specific sparse estimators of nuisance parameters, which are similar to those proposed by Smucler et al. (2019), who targeted estimands with the mixed bias property that include the parameters discussed in Hirshberg & Wager (2019) and Chernozhukov et al. (2020). Their work generalizes the methodology of Avagyan & Vansteelandt (2017), Ning et al. (2018), Bradic et al. (2019) and Tan (2020), who proposed closely related approaches to nuisance parameter estimation in the context of average treatment effects under a nonparametric model. However, these works do not consider the parameter ψ_0 indexing the semiparametric model $\mathcal{M}$; nor do they give confidence intervals via the inversion of score tests. As we will see in § 3, focusing on conditional effect parameters allows one to weaken the assumptions on sparsity described in Avagyan & Vansteelandt (2017), for example, while avoiding the use of sample splitting, which is recommended in Smucler et al. (2019). Moreover, when the model $\mathcal{M}$ fails to hold, in certain cases our estimators target overlap-weighted treatment effects and thus continue to be useful.

2. PROPOSAL

2.1. *Model and motivation*

We consider the model $\mathcal{M}$ defined by the restriction

$$g\{E(Y \mid A = a, L = l)\} - g\{E(Y \mid A = 0, L = l)\} = \psi_0 a,$$

where $g(\cdot)$ is a known link function and ψ_0 is an unknown parameter. We consider a binary treatment $A \in \{1, 0\}$; extensions to categorical treatments are given in the Supplementary Material. For continuous Y, one could use the identity link $g(x) = x$ so that ψ_0 encodes the mean difference; or, if Y is restricted to taking only positive values, the log link $g(x) = \log(x)$ could be used so that $\exp(\psi_0)$ is a ratio of expectations. We focus on these two choices of link function in this paper. Model $\mathcal{M}$ assumes that there is no treatment heterogeneity with respect to L on the scale

determined by the link function; we will weaken this restriction in the Supplementary Material. If, along with other standard conditions in the causal inference literature, one is willing to assume that L is sufficient to adjust for confounding, then ψ_0 can be interpreted, on either the additive or the multiplicative scale, as the average causal effect of removing a unit of treatment on the mean of Y, conditional on L.

Since ψ_0 can be expressed as a functional of conditional expectations, it is tempting to estimate it and construct confidence intervals based on postulating a parametric model $\mathcal{B}$ for the conditional mean of the outcome. For example, consider the model $E(Y \mid A = 0, L) = m(L; \beta_0)$, where $m(L; \beta)$ is a known function that is smooth in β and β_0 is an unknown finite-dimensional parameter. Typically, some dimension reduction is needed when the number of covariates is large relative to the sample size. Estimating β_0 via the lasso (Tibshirani, 1996) is convenient because it enforces a sparse solution; components of the estimate of β_0 will likely be set to zero. Alternatively, these estimators could be an intermediate step in selecting covariates to be included in a final model.

However, this raises two concerns. The first is that in finite samples, the distribution of a sparse estimator $\tilde{\beta}$ is typically complex. One cannot in general rule out the existence of covariates that weakly predict the outcome, but are strongly associated with the exposure, such that β_0 contains components that are close, but not equal to zero. The estimator of these entries may be forced to be zero in certain samples, but not in others, and the resulting estimator of ψ_0 based on $\tilde{\beta}$ will tend to inherit this nonregular behaviour. The consequence is that standard confidence intervals based on the normal approximation will not be uniformly valid, in the sense that for any finite n there exist parts of the parameter space for which the interval coverage may be poor (Leeb & Pötscher, 2005). The second concern is that the true model for $E(Y \mid A = 0, L)$ may not be nested within the series of regressions considered during the selection process. When L is high-dimensional, specification of a correct model for Y is especially challenging, particularly in observational studies where the distribution of the covariates differs greatly between treatment groups. In this case, model $\mathcal{B}$ will tend to extrapolate to regions outside the observed data range, and small changes in the model may greatly affect conclusions on the treatment effect.

2.2. *Doubly robust scores for conditional treatment effects under sparsity*

Although our focus is on obtaining valid confidence intervals, for ease of exposition we will begin by considering the problem of testing the hypothesis $\psi = \psi_0$. In § 2.4 we will link back to the construction of confidence intervals.

To construct our test we require two regression adjustments; the first is based on model $\mathcal{B}$ for the conditional mean of the outcome. The second is a model $\mathcal{A}$ for the conditional mean of the exposure, i.e., the propensity score when A is binary, namely $E(A \mid L) = \pi(L; \gamma_0)$, where $\pi(L; \gamma)$ is smooth in γ and γ_0 is an unknown finite-dimensional parameter. For binary A, one typically uses a logistic model, such as $\pi(L; \gamma_0) = \text{expit}(\gamma_0^{\mathrm{T}} L)$. Our test statistic will then be based on the score (Robins et al., 1992)

$$U(\psi, \eta) = d(L; \psi, \gamma)\{A - \pi(L; \gamma)\}\{H(\psi) - m(L; \beta)\}.$$

Here, $\eta = (\gamma^{\mathrm{T}}, \beta^{\mathrm{T}})^{\mathrm{T}}$ and $H(\psi) = Y \exp(-\psi A)$ if $g(\cdot)$ is the log link; otherwise $H(\psi) = Y - \psi A$. Also, $d(L; \psi, \gamma)$ is chosen for efficiency; if $g(\cdot)$ is the log link then $d(L; \psi, \gamma) = \exp(\psi)/\{\pi(L; \gamma)\exp(\psi) + 1 - \pi(L; \gamma)\}$, and otherwise $d(L; \psi, \gamma) = 1$. We refer to the Supplementary Material for further details on efficiency.

In what follows, we allow one of the models to be misspecified, such that γ_0 or β_0 no longer agrees with the truth. Even in that case, by using the law of iterated expectation one can show

that $E\{U(\psi_0, \eta_0)\} = 0$ if either $E(Y \mid A = 0, L) = m(L; \beta_0)$ or $E(A \mid L) = \pi(L; \gamma_0)$; this property of double robustness will be key to obtaining uniformly valid inference even when model $\mathcal{B}$ is misspecified. To obtain guarantees on the performance of the proposed estimators and confidence intervals, we require conditions on the sparsity of the models $\mathcal{A}$ and $\mathcal{B}$. Let us define the active set of variables as $S_\gamma = \text{support}(\gamma_0)$ and $S_\beta = \text{support}(\beta_0)$. Furthermore, let s_γ denote the cardinality $|S_\gamma|$, and likewise $s_\beta = |S_\beta|$. We will assume, for instance, that the pseudo-true parameter vector β_0 indexing the potentially misspecified model $\mathcal{B}$ includes many components equal to zero, or that s_β is much smaller than the dimension p of the covariates. Making such an assumption is necessary as our nuisance parameter estimators will enforce a sparse solution, so we want to ensure that they converge to a stable limit, regardless of whether we believe model $\mathcal{B}$ to be correct. Not doing so could adversely affect inference on the treatment effect.

Nevertheless, the interpretation of these conditions is subtle; even if a correct model is sparse, this property does not necessarily carry over if we misspecify that model. Given that in practice all statistical models are to some extent misspecified, if sparsity were plausible only under correct models then its usefulness would be limited. Bühlmann & van de Geer (2015) presented some results for linear models; in particular, if the covariates are multivariate Gaussian, then the support of the incorrect model will be a subset of the support of the correct model. For nonlinear models, let us partition L as $L = (L^{*\mathrm{T}}, X^{\mathrm{T}})^{\mathrm{T}}$, where L^* includes the true confounders of the A-Y association and X is independent of Y and A. If $X \perp\!\!\!\perp (L^*, A, Y)$, then no matter how we misspecify $E(Y \mid A, L)$, as long as the working model allows parameter values such that $E(Y \mid A, L)$ can be a function of L^* alone, the fitted mean of Y will not depend on X. A general theory is more challenging to develop when $X \perp\!\!\!\perp Y \mid (A, L^*)$, but X depends on A and L^*. Nonetheless, in the Supplementary Material, we give an example where the true and misspecified models are both sparse as well as a setting in which a misspecified model is sparse, but the true model is not.

Once we have obtained estimates $\hat{\eta}$ of η_0, we can construct a test of $\psi = \psi_0$ based on the statistic

$$T_n(\psi_0, \hat{\eta}) = \hat{V}\{\psi_0, \hat{\eta}(\psi_0)\}^{-1/2} \frac{1}{n^{1/2}} \sum_{i=1}^{n} U_i\{\psi_0, \hat{\eta}(\psi_0)\}, \tag{1}$$

where $\hat{V}\{\psi_0, \hat{\eta}(\psi_0)\}$ is the sample variance of $U(\psi_0, \hat{\eta})$ and the notation $\hat{\eta}(\psi_0)$ makes it explicit that η_0 is estimated at the fixed value ψ_0. In the following section, we will propose specific estimators $\hat{\eta}(\psi_0)$ of η_0 under the assumption that $\psi = \psi_0$. In § 3 we discuss the conditions under which the statistic (1) is uniformly asymptotically normal.

2.3. *Estimation of the nuisance parameter η*

For the moment, we will work under the model $\mathcal{M} \cap \mathcal{A}$; in other words, we assume that in addition to the semiparametric model $\mathcal{M}$, the propensity score model for the exposure holds. We will postulate a logistic model for the exposure, $\pi(L; \gamma_0) = \text{expit}(\gamma_0^{\mathrm{T}} L)$. Because our setting is high-dimensional, we will estimate γ_0 by fitting this model with a lasso penalty; for example, we solve the minimization problem

$$\hat{\gamma} = \arg\min_{\gamma} \frac{1}{n} \sum_{i=1}^{n} \log\{1 + \exp(\gamma^{\mathrm{T}} L_i)\} - A_i(\gamma^{\mathrm{T}} L_i) + \lambda_1 \|\gamma\|_1 \tag{2}$$

(Tibshirani, 1996), where $\lambda_1 > 0$ is the penalty parameter and $\|\cdot\|_1$ denotes the ℓ_1-norm. To improve finite-sample performance, in practice we recommend refitting this model with the selected covariates adjusted for using maximum likelihood. Unfortunately, in the asymptotic distribution of $n^{-1/2}\sum_{i=1}^{n} U_i\{\psi, \hat{\eta}(\psi)\}$, terms like

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\gamma}U_i\{\psi_0, \hat{\eta}(\psi_0)\}n^{1/2}(\gamma_0 - \hat{\gamma})$$

are problematic for inference, because the distribution of $\gamma_0 - \hat{\gamma}$ can be complex and difficult to approximate well.

Let $Q_i(\psi, \beta)$ denote the contribution of an individual to the negative loglikelihood associated with the proposed regression model for the transformed outcome $H(\psi_0)$. Also, let $w(L; \psi, \gamma) = \text{expit}(\gamma^{\mathrm{T}}L)\{1 - \text{expit}(\gamma^{\mathrm{T}}L)\}d(L; \psi, \gamma)$. Then we propose estimating β as the solution to the minimization problem

$$\hat{\beta}(\psi_0) = \arg\min_{\beta}\frac{1}{n}\sum_{i=1}^{n} w(L_i; \psi_0, \hat{\gamma})Q_i(\psi_0, \beta) + \lambda_2\|\beta\|_1, \tag{3}$$

where $\lambda_2 > 0$. We also formally define β_0 now as the solution to the population analogue of the above minimization problem without the penalty term that corresponds to the truth when model $\mathcal{M} \cap \mathcal{B}$ is correct. When $d(L; \psi, \gamma) = 1$, noting that $\partial w(L; \psi_0, \hat{\gamma})Q\{\psi_0, \hat{\beta}(\psi_0)\}/\partial\beta = \partial U\{\psi_0, \hat{\eta}(\psi_0)\}/\partial\gamma$, it follows that estimating β_0 as described above ensures that

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\gamma}U_i\{\psi_0, \hat{\eta}(\psi_0)\}\right\|_{\infty} \leqslant \lambda_2 \tag{4}$$

by virtue of the Karush–Kuhn–Tucker conditions; the proof of our Theorem 1 also extends to other choices of $d(L; \psi, \gamma)$. So for penalty terms satisfying the standard condition $\lambda_2 = O[\{\log(p \vee n)/n\}^{1/2}]$, where $a \vee b$ denotes the maximum of a and b, assuming that $\log(p \vee n) = o(n)$, it follows that the ℓ_∞-norm of the gradient term asymptotically goes to zero.

By using the average gradient with respect to γ as a penalized estimating equation for β, our proposal is derived from bias-reduced doubly robust estimation (Vermeulen & Vansteelandt, 2015); an estimator of β that satisfies (4) is said to possess the high-dimensional bias reduction property. In obtaining doubly robust inference on the average treatment effect under a nonparametric model, Avagyan & Vansteelandt (2017) and Tan (2020) proposed ℓ_1-penalized m-estimators for the nuisance parameter which ensure that the relevant gradients are dominated by the penalty, similar to (4). Smucler et al. (2019) generalized this framework to estimands with a bilinear influence function. Members of this class have an influence function that subtracts the target parameter from a term that depends only on η. However, the efficient influence function for the parameter ψ_0 in the semiparametric model $\mathcal{M}$ does not possess this property, and the relevant gradients may depend on the unknown target parameter. The results of Vermeulen & Vansteelandt (2015) for low-dimensional models do, however, carry over to score tests of the null hypothesis $\theta = \theta_0$, motivating our proposal to obtain confidence intervals by inverting tests.

In what follows and in the proofs given in the Supplementary Material, we focus on estimating β_0 using a lasso penalty. However, there exist other estimators which also satisfy the constraint in (4), such as a Dantzig selector-type approach or a linear program similar to that of Zhu & Bradic (2018); additional constraints may indeed be advantageous with respect to relaxing the sparsity conditions.

2.4. *Inverting the score test*

By plugging in estimates $\hat{\eta}$ of η_0 and scaling $U\{\psi_0, \hat{\eta}(\psi_0)\}$, one can obtain a statistic $T_n\{\psi_0, \hat{\eta}(\psi_0)\}$. Given the conditions discussed in the following section, we will argue that from the form of the score equation and the choice of estimators of η_0 it follows that under model $\mathcal{M} \cap \mathcal{A}$, $T_n^2\{\psi_0, \hat{\eta}(\psi_0)\} \xrightarrow{d} \chi_1^2$ in probability, where χ_1^2 is a chi-squared distribution with one degree of freedom. Hence $T_n\{\psi_0, \hat{\eta}(\psi_0)\}$ can be used to straightforwardly test the hypothesis that $\psi = \psi_0$.

We can adapt this reasoning to construct a $(1-\alpha)100\%$ confidence interval for ψ_0 as

$$[\hat{l}_s, \hat{u}_s] = \left(\psi_0 : \left[\frac{1}{n} \sum_{i=1}^{n} U_i\{\psi_0, \hat{\eta}(\psi_0)\} \right]^2 - \frac{\chi_1^2(\alpha)}{n} \hat{V}\{\psi_0, \hat{\eta}(\psi_0)\} \leqslant 0 \right), \tag{5}$$

where $\chi_1^2(\alpha)$ is the $1-\alpha$ percentile of the χ_1^2 distribution. In practice, we search over a grid of values of ψ to find the values l_s and u_s that satisfy the above inequality; note that β_0 will be re-estimated for each value of ψ considered. Furthermore, using the same reasoning, we can obtain a point estimate of ψ_0 as $\hat{\psi} = \arg\min_\psi T_n^2\{\psi, \hat{\eta}(\psi)\}$.

In the next section, we will discuss the theoretical properties of the intervals given above and indicate the specific benefits of inverting the score test as proposed.

3. ASYMPTOTIC PROPERTIES

3.1. *Main results*

Let $\mathcal{P}'$ be the class of laws that obey the intersection submodel $\mathcal{M} \cap \mathcal{A}$ and satisfy Assumptions 1–7 below; we are interested in convergence under a sequence of laws $P_n \in \mathcal{P}'$. We allow p to increase with n; we also allow the values of the population parameters ψ_0, γ_0 and β_0 to depend on n, and hence also models $\mathcal{A}$ and $\mathcal{B}$, although in the notation the dependence will be suppressed for convenience. At a given n, we assume the existence of a sparse parameter β_0 that is the solution to the unpenalized population analogue of (3). We use $\mathbb{P}_{P_n}[\cdot]$ to denote probability and $\mathbb{E}_{P_n}[\cdot]$ to denote expectation taken with respect to the local data-generating process P_n; we write $\mathbb{E}_n[\cdot]$ for sample expectations. Similarly, $\mathrm{var}_{P_n}[\cdot]$ denotes the variance with respect to the local data-generating process P_n.

We make the following assumptions.

Assumption 1 (Moment conditions). For some constants $0 < c < C < \infty$ and $4 < r < \infty$, the following hold:

(i) $\mathbb{E}_{P_n}[\{H(\psi_0) - m(L;\beta_0)\}^4 \mid A, L] < C$ with probability approaching 1;
(ii) $\mathbb{E}_{P_n}[|H(\psi_0) - m(L;\beta_0)|^r] < C$;
(iii) $c < \mathbb{E}_{P_n}[\{A - \pi(L;\gamma_0)\}^2 \mid L]$ and $c < \mathbb{E}_{P_n}[\{H(\psi_0) - m(L;\beta_0)\}^2 \mid A, L]$ with probability approaching 1.

Assumption 2 (Concentration bounds). We have that

$$\left\| \frac{1}{n^{1/2}} \sum_{i=1}^{n} \frac{\partial U_i(\psi_0, \eta_0)}{\partial \eta} \right\|_\infty = O_{P_n}\big[\{\log(p \vee n)\}^{1/2}\big].$$

Assumption 3 (Rates of convergence of the parameter estimators). For each sequence P_n and with $s^* = s_\gamma \vee s_\beta$, the estimators $\hat{\gamma}$ and $\hat{\beta}(\psi_0)$ satisfy the following:

(i) $\|\hat{\gamma} - \gamma_0\|_1 = O_{P_n}[s_\gamma\{\log(p \vee n)/n\}^{1/2}]$;
(ii) $\|\hat{\gamma} - \gamma_0\|_2 = O_{P_n}[\{s_\gamma \log(p \vee n)/n\}^{1/2}]$;
(iii) $\|\hat{\beta}(\psi_0) - \beta_0\|_1 = O_{P_n}[s^*\{\log(p \vee n)/n\}^{1/2}]$;
(iv) $\|\hat{\beta}(\psi_0) - \beta_0\|_2 = O_{P_n}[\{s^* \log(p \vee n)/n\}^{1/2}]$.

Assumption 4 (Rates of convergence for the predictions). For a given sequence P_n we have that:

(i) $\mathbb{E}_n\{(\gamma_0^{\mathrm{T}} L_i - \hat{\gamma}^{\mathrm{T}} L_i)^2\} = O_{P_n}\{s_\gamma \log(p \vee n)/n\}$;
(ii) $\mathbb{E}_n\big(\big[\{m(L_i; \beta_0) - m\{L_i; \hat{\beta}(\psi_0)\}\big]^2\big) = O_{P_n}\{s^* \log(p \vee n)/n\}$.

Assumption 5 (Restrictions on the parameter space). Let Θ denote the parameter space for ψ, so that $\psi \in \Theta$; then Θ is compact.

Assumption 6 (Dependency on estimated weights). If $m(L; \beta_0)$ is linear in β_0, then for a given sequence P_n we have that

$$\left\|\hat{\beta}(\psi_0, \gamma_0) - \hat{\beta}(\psi_0, \hat{\gamma})\right\|_1 = O_{P_n}\left[\max_{i \leqslant n} \left|H_i(\psi_0) - m(L_i; \beta_0)\right| \left\{\frac{s_\gamma s^* \log(p \vee n)}{n}\right\}^{1/2}\right].$$

Assumption 7 (Regularity conditions on the errors). We have that

$$\max_{i \leqslant n} \left|H_i(\psi_0) - m(L_i; \beta_0)\right| (s_\gamma s^*)^{1/2} \log(p \vee n) = o(n^{1/2})$$

with probability approaching 1.

Assumption 1 places mild moment conditions on the residuals. Upon taking the derivative with respect to β, one can use the moderate deviation theory of self-normalized sums to show that Assumption 2 holds by invoking Assumptions 3 and 4, making use of the fact that the model for the exposure is correctly specified, and assuming that the covariates have bounded support (Farrell, 2015). Taking the derivative with respect to γ, Assumption 2 is required in order to show consistency of the proposed estimator of β_0; the ℓ_∞-norm of the estimating equations evaluated at β_0 should be dominated by the penalty λ_2, which is assumed to be $O_{P_n}[\{\log(p \vee n)/n\}^{1/2}]$. The rates in Assumptions 3 and 4 are known to hold for several sparse estimators, including the lasso, post-lasso and weighted lasso (Belloni et al., 2016; Ning & Liu, 2017). That the rate in Assumption 4(i) holds for lasso logistic regression follows from Farrell (2015), and the rate in Assumption 4(ii) holds for weighted lasso and post-lasso estimators (Belloni et al., 2016). Assumption 5 is needed only when $H(\psi_0) = Y \exp(-\psi_0 A)$. The property in Assumption 6 has been shown to hold for the proposed estimators of β_0 in Dukes et al. (2020), so long as the model is linear; see Lemma 1 in the appendix of that paper, and we also refer to Dukes et al. (2020) for a list of primitive conditions required for the property to hold. Assumption 7 allows us to trade off restrictions on the distribution of the errors against stronger sparsity conditions; it holds automatically when the errors are bounded.

THEOREM 1. *Suppose Assumptions 1–4 hold and that also Assumption 5 holds if* $H(\psi_0) = Y \exp(-\psi_0 A)$. *If, in addition,*

(i) $(s_\gamma^2 + s_\beta^2)\log^2(p \vee n) = o(n)$

holds, then, using the estimators $\hat{\gamma}$ *and* $\hat{\beta}(\psi)$ *defined in* (2) *and* (3)*, we have*

$$\lim_{n\to\infty} \sup_{P_n \in \mathcal{P}'} \left| \mathbb{P}_{P_n}\left(\psi_0 \in [\hat{l}_s, \hat{u}_s]\right) - (1-\alpha) \right| = 0 \tag{6}$$

under model $\mathcal{M} \cap \mathcal{A}$.

This result shows that under ultra-sparse regimes where $s_\gamma \ll n^{1/2}$ and $s_\beta \ll n^{1/2}$, one can construct a uniformly valid interval for ψ_0 without requiring a correct outcome model $\mathcal{B}$. Fitting the working model for Y in the specific way proposed above helps to correct for the regularization bias incurred by using the sparse estimate $\hat{\gamma}$, similar to the work on debiasing the lasso (Belloni et al., 2016; Ning & Liu, 2017). Indeed, the ultra-sparsity condition in that literature is standard if one restricts to estimation via the lasso or the Dantzig selector. The key difference is that we do not require a correct model for $\mathcal{B}$. One might object to using sparsity assumptions in causal inference settings, because in many studies all measured variables are somewhat related to the exposure and/or outcome. It may nevertheless be plausible that after adjusting for a small number of key variables, the rest will have a minor contribution, and our results should extend to approximately sparse models (Farrell, 2015).

Stronger results are available on robustness to misspecification when the working model for the outcome is linear.

COROLLARY 1. *Suppose that* $m(L;\beta)$ *is linear with respect to* β*. Then under the conditions of Theorem* 1*, the confidence interval* $[\hat{l}_s, \hat{u}_s]$ *is uniformly valid as in* (6) *under the union model* $\mathcal{M} \cap (\mathcal{A} \cup \mathcal{B})$.

In this case, the resulting intervals are uniformly doubly robust, in the sense that they asymptotically contain the true parameter with probability determined by the nominal α-level when either model $\mathcal{A}$ or model $\mathcal{B}$ is correct, uniformly over the parameter space. The result follows upon noting that the estimator of γ proposed in (2) satisfies the constraint that $\|n^{-1}\sum_{i=1}^n \partial U_i\{\psi_0, \hat{\eta}(\psi_0)\}/\partial\beta\|_\infty \leqslant \lambda_1$. In principle, uniformly doubly robust confidence intervals could be constructed when the outcome model is nonlinear. However, this is challenging computationally as estimating γ_0 now requires weights dependent on $\hat{\beta}(\psi_0)$ such that iteration is required or $\hat{\gamma}$ and $\hat{\beta}(\psi_0)$ have to be simultaneously estimated. This must be done over all values of ψ considered in solving (5).

If all models are correct and a particular location-shift condition holds, then one can weaken the corresponding assumptions on model sparsity.

THEOREM 2. *Let us restrict our attention to the class of laws* $\mathcal{P}$ *that obey the intersection submodel* $\mathcal{M} \cap \mathcal{A} \cap \mathcal{B}$*. Moreover, suppose that*

(i) $(s_\gamma + s_\beta)\log(p \vee n) = o(n)$,
(ii) $(s_\gamma s^*)\log^2(p \vee n) = o(n)$, *and*
(iii) $H(\psi_0) \perp\!\!\!\perp A \mid L$

Then, if $m(L;\beta_0)$ *is linear in* β_0*, under Assumptions* 1, 2, 4, 6 *and* 7 *and the conditions* (i)–(iii) *above, the confidence interval* $[\hat{l}_s, \hat{u}_s]$ *is uniformly valid as in* (6).

For general models for $m(L;\beta_0)$*, the same result holds if* γ_0 *and* β_0 *are estimated from a subsample of the data separate from the one used to construct the interval, without requiring condition* (iii)*, but requiring Assumption* 5.

It follows from Theorem 2 that for $m(L;\beta_0) = \beta_0^{\mathrm{T}} L$, if both of the models $\mathcal{A}$ and $\mathcal{B}$ are correct in addition to model $\mathcal{M}$ and if model $\mathcal{A}$ is ultra-sparse, one can allow model $\mathcal{B}$ to be dense and vice versa. Hence we describe our confidence intervals as sparsity adaptive. Condition (iii) would hold under the semiparametric location-shift model

$$Y = \psi_0 A + \epsilon, \tag{7}$$

where $\epsilon \perp\!\!\!\perp A \mid L$. If L is sufficient to adjust for confounding of the effect of A on Y, we can rephrase model (7) as a linear structural distribution model (Robins, 1997).

With nonlinear $m(L;\beta_0)$, we revert to sample splitting to relax the sparsity assumptions, although we conjecture that uniform validity under weakened conditions is also possible here. This is partly because results in Dukes et al. (2020) imply that confidence intervals obtained via our procedure without weighting are valid under the intersection submodel if conditions (i)–(iii) in Theorem 2 hold; see also the corollary below. It also follows from the proofs in the Supplementary Material that without sample splitting, so long as all models are correct, we merely require $s_\gamma \log(p \vee n) = o(n)$ and $s^{*2} \log^2(p \vee n) = o(n)$; hence we have sparsity robustness in one direction, in that we can potentially weaken the conditions for the estimation of γ.

When $H(\psi_0) = Y - \psi_0 A$, Chernozhukov et al. (2018) arrived at conditions (i) and (ii) without requiring (iii) via the use of sample splitting. Moreover, as long as their recommended cross-fitting scheme is used, asymptotically there should be little or no efficiency loss. Nevertheless, the benefits of sample splitting are currently apparent only when estimators of both β_0 and γ_0 converge to the truth so that the sparsity conditions can be weakened. When their score equations are not linear in the target parameter, the regularity conditions in Chernozhukov et al. (2018) are more complicated even when combined with sample splitting, whereas the confidence intervals proposed in the present article are valid under simpler conditions regardless of whether $H(\psi_0)$ is linear in ψ_0, largely by virtue of inverting a score test.

The sparsity adaptivity property does not appear to be available for the following two-stage estimator, where first a model for $E(Y \mid A, L)$ is fitted to obtain estimates $\check{\psi}$ and $\check{\beta}(\check{\psi})$ of ψ_0 and β_0, then $\check{\psi}$ is discarded, and ψ_0 is estimated at the second stage via the solution to the estimating equations $\sum_{i=1}^n U_i\{\psi, \hat{\gamma}, \check{\beta}(\check{\psi})\} = 0$; Wald-based tests and intervals are then straightforward to construct. The estimates $\check{\beta}(\check{\psi})$ of β are dependent on $(A_i)_{i=1}^n$, whereas $\hat{\beta}(\psi_0)$ is allowed only to depend on the exposure data via the transformed outcome $H(\psi_0)$. In the proof of Theorem 2 we exploit this property, in combination with the conditional independence of $H(\psi_0)$ and A from (iii), to emulate settings where β is estimated in a separate sample.

In a final corollary, we indicate the consequences of these results for a broader class of machine learning algorithms. This result is implied by the proofs of Theorem 2.

COROLLARY 2. *Suppose that we obtain unweighted estimators $\hat{\pi}(L)$ and $\hat{m}(L;\psi_0)$ of $E(A \mid L) = \pi(L)$ and $E(Y \mid A = 0, L) = m(L)$, respectively, via machine learning and repeat the above steps, inverting the test statistic $T_n\{\psi_0, \hat{\pi}, \hat{m}(\psi_0)\}$ to obtain an interval $[\check{l}_{\mathrm{s}}, \check{u}_{\mathrm{s}}]$. Furthermore, we assume that the estimators satisfy $n^{-1}\sum_{i=1}^n\{\hat{\pi}(L_i) - \pi(L_i)\}^2 = o_{P_n}(1)$, $n^{-1}\sum_{i=1}^n\{\hat{m}(L_i;\psi_0) - m(L_i)\}^2 = o_{P_n}(1)$, and*

$$\left[n^{-1}\sum_{i=1}^n\{\hat{\pi}(L_i) - \pi(L_i)\}^2\right]^{1/2}\left[n^{-1}\sum_{i=1}^n\{\hat{m}(L_i;\psi_0) - m(L_i)\}^2\right]^{1/2} = o_{P_n}(n^{-1/2}).$$

Then, provided that Assumption 1 and condition (iii) *hold, under the class of laws $\mathcal{P}$ that obey the intersection submodel $\mathcal{M} \cap \mathcal{A} \cap \mathcal{B}$, $[\check{l}_{\mathrm{s}}, \check{u}_{\mathrm{s}}]$ is a uniformly valid interval as defined above.*

By inverting a score test and utilizing the location-shift condition, one can use arbitrary machine learning estimators to construct the interval without having to either employ sample splitting or invoke strong Donsker-type conditions. This applies only when estimators converge to the truth.

3.2. *Practical considerations*

The previous theoretical results have implications for statistical practice. Firstly, given that data-adaptive methods such as penalized estimators tend to perform better with more data, at small-to-moderate sample sizes we recommend implementing the proposed procedure without sample splitting or cross-fitting. Under homoscedasticity there appears to be no theoretical gain in using sample splitting. When the location-shift assumption is violated, our procedure remains valid under stronger assumptions, and one may still observe better performance due to the variable selection involved being more stable at larger n. Cross-fitting may also become cumbersome in more complex causal inference settings; see the Supplementary Material.

While the sparsity adaptivity property appears to be specific to inverted score test confidence intervals, these are nevertheless computationally more demanding to implement than Wald-based intervals. Although this is perhaps less of an issue for the simple semiparametric models considered here, it may become a more prominent concern when the target parameter is higher-dimensional, as in the settings considered in the Supplementary Material. Important directions for further work therefore include an empirical comparison of full-sample versus sample-splitting/cross-fitting methods and development of computationally feasible methods for estimating vector-valued ψ_0 along with confidence intervals.

In terms of efficiency, if $g(x)$ is the identity link function and $\text{var}\{Y - \psi_0 A - m(L;\beta_0) \mid A, L\} = \text{var}\{Y - \psi_0 A - m(L;\beta_0)\}$, then the asymptotic variance of $\hat{\psi}$ attains the semiparametric efficiency bound under the partially linear model of Robinson (1988) at the intersection submodel. For $g(x) = \log(x)$, our estimator is also efficient under the partially log-linear model when all working models are correct. Although choosing between semiparametric versus nonparametric inference is challenging, our estimators tend to be much more efficient than those of the average treatment effect under the nonparametric model, as is borne out in the simulation studies in § 4. In very high-dimensional models or under violations of sparsity, our empirical results also suggest that imposing the additional parametric structure yields estimators with more stable behaviour. Nevertheless, even when model $\mathcal{M}$ does not hold, our estimator may still retain a meaningful interpretation. It is known that when $g(x)$ is the identity link, $d(L;\psi,\gamma) = 1$ and model $\mathcal{A}$ holds, the proposed estimator $\hat{\psi}$ converges to the overlap-weighted average of the stratum-specific treatment effects $\psi_0^*(L)$,

$$\frac{E\{\text{var}(A \mid L)\psi_0^*(L)\}}{E\{\text{var}(A \mid L)\}}$$

(Crump et al., 2009). For $g(x) = \log(x)$, it follows from Dukes & Vansteelandt (2018) that when the exposure is normally distributed and our score is weighted by the reciprocal of a consistent estimator of $E(Y \mid L)$, the corresponding estimator converges to the overlap-weighted log-risk ratio. Unfortunately, a similar result is not currently available for binary exposures. Nevertheless, in simulations we see that our proposed method performs well at inferring the overlap-weighted effect; see the Supplementary Material.

Our analysis so far is specific to postulating high-dimensional models for the conditional means of A and Y. Sparse parametric estimators may be preferable to their more flexible nonparametric counterparts when covariates are high-dimensional. Nonetheless, it is an interesting open question as to whether the high-dimensional bias reduction property holds for a general

class of nonparametric estimators. In the nonparametric setting, Benkeser et al. (2017) described how to obtain doubly robust inference in settings where one of the data-adaptive estimators does not converge to the truth. However, it is currently unclear whether their approach can be made uniformly valid.

4. Simulation study

In each of the 1000 simulations in our Experiment 1, we created a dataset with $n = 200$ observations. We generated the covariates L^* from a multivariate normal distribution $N(0, \Sigma)$ where Σ is a Toeplitz matrix with $\Sigma_{j,k} = 2^{-|j-k|-1}$; then we created L by including an additional column for the intercept. The dimension of L is $p = 200$. Further, A was taken to be a Bernoulli random variable with conditional expectation $E(A \mid L) = \text{expit}(\gamma_0^{\mathrm{T}} L)$, and Y was generated from the normal distribution $N(0.3A + \beta_0^{\mathrm{T}} L, 1)$ where $\beta_0 = \tau\{-1, 1, -1, 2^{-\rho}, \ldots, (p-2)^{-\rho}\}$. As in Farrell (2015), we used τ to vary the signal strength, with 1 indicating a stronger signal and 0.4 a weaker one, and we used ρ to control the sparsity, with 2 indicating a sparser model and 0.5 a denser one. In this and all subsequent experiments, $\gamma_0 = 1, -1, 1, -2^{-2}, \ldots, -(p-2)^{-2}$, except in Experiment 2 where $\gamma_0 = 1, 1, 1, -2^{-2}, \ldots, -(p-2)^{-2}$. In Experiment 2, we created the covariates $X_1 = |\log(10 + L_1^*)|$, $X_2 = \{L_2^* \exp(L_1^*)\}/5$ and $X_3 = (L_2^* + L_3^*)^2$. A matrix X was constructed by binding X_1, X_2 and X_3 along with columns 4 to p of L^* plus a column corresponding to the intercept. Then we generated $N(0.3A + \bar{\beta}_0^{\mathrm{T}} X, 1)$, where $\bar{\beta}_0$ is the same as β_0 except that its leading three entries are equal to 1. Experiment 3 is similar to Experiment 1, except that now $Y \sim N\{0.3A + \beta_0^{\mathrm{T}} L, \sigma^2(A, L)\}$ where $\sigma^2(A_i, L_i) = \{n^{-1} \sum_{j=1}^{n} (0.3A_j + \beta_0^{\mathrm{T}} L_j)^2\}^{-1} (0.3A_i + \beta_0^{\mathrm{T}} L_i)^2$. Each experiment was repeated with $p = 250$ while varying τ and ρ.

We first considered a naïve post-selection approach, where Y was regressed on A and L using a lasso penalty, forcing the exposure into the model. The final model was then refitted and adjusted for A and the selected covariates, yielding the estimate $\hat{\psi}_{\mathrm{OLS}}$. We compared this with the post-double-selection method described in Belloni et al. (2016) and implemented in the `R` (R Development Core Team, 2021) package `hdm` (Chernozhukov et al., 2016), as well as the partialling-out method in the same `R` package, yielding the estimators $\hat{\psi}_{\mathrm{PDS}}$ and $\hat{\psi}_{\mathrm{PO}}$. Both approaches were implemented using the penalties selected by the `hdm` package; we also present results for post-double-selection and partialling-out using penalties obtained via cross-validation instead; let $\hat{\psi}_{\text{PDS-CV}}$ and $\hat{\psi}_{\text{PO-CV}}$ denote the respective estimators. For $\hat{\psi}_{\mathrm{OLS}}$, we used the standard model-based variance estimators; for $\hat{\psi}_{\mathrm{PDS}}$, $\hat{\psi}_{\mathrm{PO}}$, $\hat{\psi}_{\text{PDS-CV}}$ and $\hat{\psi}_{\text{PO-CV}}$, variance estimation was implemented as in the `hdm` package.

For our approach, each working model was adjusted for L. The lasso penalties for the working models for exposure and outcome were both selected using 20-fold cross-validation, choosing λ_1 to be the value that minimized the expected cross-validated error, and likewise for λ_2. In the case of the outcome model, cross-validation was done under the null $\psi = 0$. If too many covariates were selected such that refitting model $\mathcal{A}$ using maximum likelihood failed, an increment of 0.005 was added to λ_1. In Experiment 1 both models were correctly specified, whereas in Experiment 2 only the logistic model for the exposure was correct. In Experiment 3, both models were correct again; however, condition (iii) in Theorem 2 was violated due to the dependence of the residual variance on the exposure. A point estimate of the treatment effect, denoted by $\hat{\psi}_{\mathrm{HDBR}}$, and confidence intervals were otherwise obtained as in § 2.4. To compare the efficiency of $\hat{\psi}_{\mathrm{HDBR}}$ relative to the other estimators in the ASE column of the tables, we evaluated the sample standard error of the score function $U\{\psi, \hat{\eta}(\psi)\}$ with ψ held fixed at the true value.

The results for Experiments 1 and 2 are reported in Tables 1 and 2, respectively; results for Experiment 3 can be found in the Supplementary Material. They indicate that even in highly sparse, strong-signal settings where the model for Y is correctly specified, the naïve approach still

Table 1. *Simulation results from Experiment* 1 *with* $n = 200$

		$p = 200$				$p = 250$			
ρ, τ	Est	Bias	MCSD	ASE	Cov	Bias	MCSD	ASE	Cov
2, 1	$\hat{\psi}_{\text{OLS}}$	−0.77	2.0	1.6	85.5	−0.97	1.9	1.5	82.7
	$\hat{\psi}_{\text{PDS}}$	−0.50	1.9	1.7	90.6	−0.60	1.9	1.7	91.3
	$\hat{\psi}_{\text{PO}}$	−0.82	1.7	1.6	91.1	−0.91	1.7	1.6	91.3
	$\hat{\psi}_{\text{PDS-CV}}$	−0.73	2.0	1.8	90.2	−0.83	1.9	1.8	91.4
	$\hat{\psi}_{\text{PO-CV}}$	−0.90	1.8	1.6	86.9	−1.02	1.7	1.6	86.2
	$\hat{\psi}_{\text{HDBR}}$	−0.78	2.3	1.8	91.6	−0.92	2.3	1.8	89.4
0.5, 1	$\hat{\psi}_{\text{OLS}}$	−1.15	3.1	2.0	78.8	−1.67	3.2	2.1	73.7
	$\hat{\psi}_{\text{PDS}}$	−5.71	3.0	2.8	44.0	−5.81	3.1	2.8	43.6
	$\hat{\psi}_{\text{PO}}$	−5.71	2.9	2.7	44.8	−5.82	3.0	2.8	43.0
	$\hat{\psi}_{\text{PDS-CV}}$	−1.18	2.4	2.3	90.6	−1.29	2.6	2.4	89.7
	$\hat{\psi}_{\text{PO-CV}}$	−1.71	1.8	1.5	67.4	−1.72	2.0	1.7	70.4
	$\hat{\psi}_{\text{HDBR}}$	−1.20	3.2	2.1	92.1	−1.59	3.5	2.4	90.2
2, 0.4	$\hat{\psi}_{\text{OLS}}$	−1.55	2.1	1.5	74.4	−1.77	2.3	1.5	66.4
	$\hat{\psi}_{\text{PDS}}$	−2.78	1.6	1.6	59.6	−2.90	1.7	1.7	57.4
	$\hat{\psi}_{\text{PO}}$	−2.78	1.6	1.7	60.9	−2.91	1.7	1.7	57.6
	$\hat{\psi}_{\text{PDS-CV}}$	−0.68	2.0	1.8	91.5	−0.70	2.1	1.9	89.7
	$\hat{\psi}_{\text{PO-CV}}$	−0.77	1.9	1.8	89.9	−0.80	2.0	1.8	87.6
	$\hat{\psi}_{\text{HDBR}}$	−0.69	2.2	1.9	92.5	−0.69	2.2	2.0	90.8
0.5, 0.4	$\hat{\psi}_{\text{OLS}}$	−2.16	2.3	1.7	68.5	−2.33	2.2	1.7	64.7
	$\hat{\psi}_{\text{PDS}}$	−2.95	1.9	1.8	62.9	−3.00	1.8	1.8	61.1
	$\hat{\psi}_{\text{PO}}$	−2.95	1.9	1.8	63.7	−3.01	1.8	1.8	61.7
	$\hat{\psi}_{\text{PDS-CV}}$	−1.06	2.2	2.0	89.1	−1.09	2.2	2.1	89.5
	$\hat{\psi}_{\text{PO-CV}}$	−1.16	2.1	1.9	86.0	−1.19	2.1	2.0	86.4
	$\hat{\psi}_{\text{HDBR}}$	−1.04	2.4	2.1	90.8	−1.07	2.4	2.2	91.4

Est, estimator; Bias, Monte Carlo bias multiplied by 10; MCSD, Monte Carlo standard deviation multiplied by 10; ASE, average estimated standard error multiplied by 10; Cov, coverage probability multiplied by 100.

has a large bias with standard errors that do not adequately reflect the uncertainty induced by the lasso procedure. The post-double-selection and partialling-out methods performed better in this situation, but often failed to attain the nominal coverage level either under denser models or when the signal was weaker. Part of the poorer performance was due to the choice of penalty terms; use of cross-validation improved results considerably. This was particularly true of the post-double-selection method, which performed well in Experiment 3. The results for these methods were comparable or worse under misspecification of the outcome model, indicating that performance is very sensitive to the data-generating mechanism. In contrast, our proposed confidence intervals came close to attaining their nominal coverage across the majority of settings. They performed poorest in dense settings where the signal was strong, as well as when errors were heteroscedastic, as predicted by the theory. However, they still yielded a general improvement over alternatives. Experiments 1–3 were repeated with $n = p = 400$ and with $n = 500$ and $p = 2000$, for which superior coverage was observed across all settings; see the Supplementary Material, where we also compare our estimators with those proposed by Athey et al. (2018) and Ning et al. (2018) for the average treatment effect. We see that our approach performs better in terms of bias and interval coverage, but unlike those other methods it relies on the stronger assumption of the semiparametric model $\mathcal{M}$ being correct.

Table 2. *Simulation results from Experiment* 2 *with* $n = 200$

		$p = 200$				$p = 250$			
ρ, τ	Est	Bias	MCSD	ASE	Cov	Bias	MCSD	ASE	Cov
2, 1	$\hat{\psi}_{\text{OLS}}$	−1.53	3.8	2.7	82.6	−1.50	4.0	2.6	79.8
	$\hat{\psi}_{\text{PDS}}$	−3.60	2.9	2.9	79.5	−3.60	3.0	2.9	78.9
	$\hat{\psi}_{\text{PO}}$	−3.60	2.9	3.1	84.3	−3.60	3.0	3.1	83.6
	$\hat{\psi}_{\text{PDS-CV}}$	−3.46	3.2	3.1	78.8	−3.57	3.3	3.1	80.8
	$\hat{\psi}_{\text{PO-CV}}$	−3.44	3.1	3.1	78.9	−3.57	3.2	3.1	80.0
	$\hat{\psi}_{\text{HDBR}}$	0.25	3.0	2.8	94.5	0.13	3.1	2.8	93.8
0.5, 1	$\hat{\psi}_{\text{OLS}}$	−2.64	6.4	5.3	89.0	−2.23	6.3	5.3	88.9
	$\hat{\psi}_{\text{PDS}}$	−7.17	5.9	5.9	78.2	−6.81	5.8	5.9	81.2
	$\hat{\psi}_{\text{PO}}$	−7.17	5.9	6.4	84.0	−6.81	5.8	6.5	85.6
	$\hat{\psi}_{\text{PDS-CV}}$	−6.87	6.5	6.4	81.4	−6.24	6.7	6.6	83.9
	$\hat{\psi}_{\text{PO-CV}}$	−6.79	6.3	6.6	83.5	−6.17	6.6	6.8	86.2
	$\hat{\psi}_{\text{HDBR}}$	−0.38	6.2	5.6	95.7	0.02	6.7	5.8	95.0
2, 0.4	$\hat{\psi}_{\text{OLS}}$	0.14	2.3	1.7	84.2	0.19	2.3	1.7	86.1
	$\hat{\psi}_{\text{PDS}}$	−1.41	2.0	2.0	89.0	−1.46	2.1	2.0	88.2
	$\hat{\psi}_{\text{PO}}$	−1.41	2.0	2.1	90.8	−1.47	2.1	2.1	89.9
	$\hat{\psi}_{\text{PDS-CV}}$	−1.53	2.2	2.1	88.2	−1.55	2.2	2.2	87.2
	$\hat{\psi}_{\text{PO-CV}}$	−1.56	2.1	2.1	86.8	−1.59	2.2	2.1	86.1
	$\hat{\psi}_{\text{HDBR}}$	0.02	2.4	2.2	94.2	0.04	2.5	2.2	94.4
0.5, 0.4	$\hat{\psi}_{\text{OLS}}$	−0.77	2.9	2.5	90.2	−0.80	2.9	2.5	91.2
	$\hat{\psi}_{\text{PDS}}$	−2.73	2.8	2.9	85.4	−2.84	2.9	2.9	83.6
	$\hat{\psi}_{\text{PO}}$	−2.73	2.8	3.1	88.8	−2.84	2.9	3.1	87.0
	$\hat{\psi}_{\text{PDS-CV}}$	−2.72	3.1	3.1	87.0	−2.89	3.3	3.2	83.8
	$\hat{\psi}_{\text{PO-CV}}$	−2.71	3.0	3.2	87.3	−2.88	3.3	3.2	84.9
	$\hat{\psi}_{\text{HDBR}}$	0.02	3.4	3.0	94.2	−0.06	3.5	3.0	95.1

Est, estimator; Bias, Monte Carlo bias multiplied by 10; MCSD, Monte Carlo standard deviation multiplied by 10; ASE, average estimated standard error multiplied by 10; Cov, coverage probability multiplied by 100.

5. Discussion

In the Supplementary Material, we extend our proposed method to settings with effect heterogeneity, time-varying exposures and confounders. This extension is nontrivial, given that the theory of bias-reduced doubly robust estimation was previously only well developed for scalar target parameters. Our proposal would require some adaptation for the conditional causal odds ratio, since no doubly robust estimator of this parameter currently exists under the union model $\mathcal{M} \cap (\mathcal{A} \cup \mathcal{B})$ when $g(\cdot)$ is the logit link, in part due to the noncollapsibility of the parameter. The same is true of the hazard ratio.

Acknowledgement

Both authors were supported by the Ghent University Special Research Fund and the Research Foundation of Flanders. Vansteelandt is also affiliated with the Department of Medical Statistics at the London School of Hygiene and Tropical Medicine.

Supplementary material

Supplementary material available at *Biometrika* online includes details on effect heterogeneity and categorical exposures, discussion of controlled direct effects, examples of sparse misspecified models, proofs of Theorems 1 and 2, and additional simulation results.

REFERENCES

ATHEY, S., IMBENS, G. W. & WAGER, S. (2018). Approximate residual balancing: Debiased inference of average treatment effects in high dimensions. *J. R. Statist. Soc.* B **80**, 597–623.

AVAGYAN, V. & VANSTEELANDT, S. (2017). Honest data-adaptive inference for the average treatment effect under model misspecification using penalised bias-reduced double-robust estimation. *arXiv:* 1708.03787.

BELLONI, A., CHERNOZHUKOV, V. & WEI, Y. (2016). Post-selection inference for generalized linear models with many controls. *J. Bus. Econ. Statist.* **34**, 606–19.

BENKESER, D., CARONE, M., LAAN, M. J. V. D. & GILBERT, P. B. (2017). Doubly robust nonparametric inference on the average treatment effect. *Biometrika* **104**, 863–80.

BRADIC, J., WAGER, S. & ZHU, Y. (2019). Sparsity double robust inference of average treatment effects. *arXiv:* 1905.00744.

BÜHLMANN, P. & VAN DE GEER, S. (2015). High-dimensional inference in misspecified linear models. *Electron. J. Statist.* **9**, 1449–73.

CHERNOZHUKOV, V., CHETVERIKOV, D., DEMIRER, M., DUFLO, E., HANSEN, C., NEWEY, W. & ROBINS, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Economet. J.* **21**, C1–68.

CHERNOZHUKOV, V., NEWEY, W. & SINGH, R. (2020). Double/de-biased machine learning of global and local parameters using regularized Riesz representers. *arXiv:* 1802.08667v4.

CHERNOZHUKOV, V., HANSEN, C. & SPINDLER, M. (2016). hdm: High-dimensional metrics. *R Journal* **8**, 185–99.

CRUMP, R. K., HOTZ, V. J., IMBENS, G. W. & MITNIK, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika* **96**, 187–99.

DUKES, O., AVAGYAN, V. & VANSTEELANDT, S. (2020). Doubly robust tests of exposure effects under high-dimensional confounding. *Biometrics* **76**, 1190–200.

DUKES, O. & VANSTEELANDT, S. (2018). A note on G-estimation of causal risk ratios. *Am. J. Epidemiol.* **187**, 1079–84.

FARRELL, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *J. Economet.* **189**, 1–23.

HIRSHBERG, D. A. & WAGER, S. (2019). Augmented minimax linear estimation. *arXiv:* 1712.00038v5.

LEEB, H. & PÖTSCHER, B. M. (2005). Model selection and inference: Facts and fiction. *Economet. Theory* **21**, 21–59.

NING, Y. & LIU, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *Ann. Statist.* **45**, 158–95.

NING, Y., PENG, S. & IMAI, K. (2018). Robust estimation of causal effects via high-dimensional covariate balancing propensity score. *arXiv:* 1812.08683.

R DEVELOPMENT CORE TEAM (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org.

ROBINS, J. M. (1997). Causal inference from complex longitudinal data. In *Latent Variable Modeling and Applications to Causality*, M. Berkane, ed., Lecture Notes in Statistics. New York: Springer, pp. 69–117.

ROBINS, J. M., MARK, S. D. & NEWEY, W. K. (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics* **48**, 479–95.

ROBINS, J. M., ROTNITZKY, A. & ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Am. Statist. Assoc.* **89**, 846–66.

ROBINSON, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica* **56**, 931–54.

SMUCLER, E., ROTNITZKY, A. & ROBINS, J. M. (2019). A unifying approach for doubly-robust ℓ_1 regularized estimation of causal contrasts. *arXiv:* 1904.03737v3.

TAN, Z. (2020). Model-assisted inference for treatment effects using regularized calibrated estimation with high-dimensional data. *Ann. Statist.* **48**, 811–37.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B **58**, 267–88.

VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. & DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42**, 1166–202.

VAN DER LAAN, M. J. & RUBIN, D. (2006). Targeted maximum likelihood learning. *Int. J. Biostatist.* **2**, article no. 11.

VERMEULEN, K. & VANSTEELANDT, S. (2015). Bias-reduced doubly robust estimation. *J. Am. Statist. Assoc.* **110**, 1024–36.

ZHU, Y. & BRADIC, J. (2018). Significance testing in non-sparse high-dimensional linear models. *arXiv:* 1610.02122v4.

[*Received on* 13 *March* 2019. *Editorial decision on* 13 *July* 2020]