Uniformly valid confidence intervals for conditional treatment effects in misspecified high-dimensional models

Oliver Dukes
Department of Applied Mathematics, Computer Science and Statistics
Ghent University, Belgium

and Stijn Vansteelandt
Department of Applied Mathematics, Computer Science and Statistics
Ghent University, Belgium
and Department of Medical Statistics
London School of Hygiene and Tropical Medicine, U.K.

email: oliver.dukes@ugent.be

Abstract

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 regularisation 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 sample size, correctly specifying a model for the outcome is non-trivial. In this work, we deal with both of these concerns simultaneously, delivering 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.

1 Introduction

We focus on the problem of constructing confidence intervals for a low-dimensional component in a high-dimensional conditional mean model. In epidemiologic 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. More seriously, they may also fail to be *uniformly valid*; there may be no sample size at which a given procedure is guaranteed to attain its nominal coverage/size. In particular, the treatment effect estimator may have a complex, non-normal 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 for 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 model for the treatment-selection mechanism is postulated (a.k.a. the propensity score model), as well as for the outcome. Doubly robust estimators are unbiased when at least one of these working models is correctly specified. van der Laan and Rubin (2006) originally proposed combining this framework with flexible data-adaptive estimation of nuisance parameters. In the context of marginal treatment effects, Farrell (2015) shows that uniformly valid inferences can be obtained in high-dimensional settings by fitting both working models using the Lasso. Chernozhukov et al. (2018) extend this work in many directions, for example allowing for a wide class of machine learning methods to estimate the nuisance parameters indexing the working models. 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 $o_p(n^{-1/2})$. In high-dimensional parametric models, the latter condition requires that the product of the number of non-zero coefficients in each model must be small relative to the sample size.

The second strand of work instead focuses on estimating the target parameter by 'de-biasing' or 'de-sparsifying' an initial Lasso-based estimate (Javanmard and Montanari, 2014; van de Geer et al., 2014; Zhang and Zhang, 2014) or score equation (Chernozhukov et al., 2015; Ning and Liu, 2017). Here, the bias in the penalised 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 de-biasing, (under certain conditions) the updated estimator is uniformly consistent and asymptotically normal. An advantage of this approach is its generality; in contrast, doubly robust estimators are only known to exist for a limited class of parameters. However, much stronger conditions are required on the sparsity of both the model for Y and/or the de-biasing regression adjustment (unless certain tailored sparse estimators are used). In some settings, both approaches may coincide e.g. under the partially linear regression model.

In this work, 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. We work instead under a correct model for the exposure, since in many epidemiologic studies, clinicians may have some knowledge on which variables affect the decision to give treatment and how they do so. Moreover, in complex settings (with time-varying treatments and covariates), coherent specification of multiple models for the outcome is non-trivial, so one typically prefers to do inference based on the propensity score (Robins, 1997). Hence a different perspective is taken to recent research focusing on settings where a propensity score is misspecified or difficult to estimate well (Athey et al., 2018). In linear models, our intervals are *uniformly doubly robust*; valid if either a model for the exposure or outcome is correctly specified. We achieve this robustness by considering specific penalised estimators of the nuisance parameters. In comparison, the inferences in Farrell (2015) and Chernozhukov et al. (2018) are not generally robust to misspecification of either working model. When both models are correct (and a location-shift assumption holds), our proposal is valid under weakened conditions on model sparsity, without requiring sample-splitting. As we will discuss, our results here have wider ramifications for the use of machine learning algorithms in estimating causal effects. In the nonparametric setting, Benkeser et al. (2017) describe how to obtain doubly robust inference in settings where one of the data-adaptive estimators does not converge to the truth. Their focus is however on marginal treatment effects, and their results are not strictly applicable in high-dimensional settings, since they are developed under Donsker conditions which prohibit the dimension of the covariates growing with the sample size. We hence take an alternative approach.

Our work generalises the proposal of Dukes et al. (2019), where the focus was on hypothesis tests under the causal null hypothesis of no treatment effect. In comparison, we show how to construct uniformly valid confidence intervals and tests away from the null, describe a more general class of estimators for the nuisance parameters and extend the proposal to multivariate treatment effect parameters.

2 Proposal

2.1 Model and motivation

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

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

where $g(\cdot)$ is a known link function and ψ_0 is an unknown parameter. For continuous Y, one might use the identity link $g(x) = x$ so that ψ_0 encodes the mean difference, or if Y is constrained to only taking positive values, the log link $g(x) = \log(x)$ (so $\exp(\psi_0)$ is a ratio of expectations). These two choices of link function are the focus of 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 Section 5. If one is willing to assume that L is sufficient to adjust for confounding (along with the other standard conditions in the causal inference literature), then ψ_0 can be interpreted (either on the additive or 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 $E(Y|A = 0, L) = m(L; \beta_0)$, where $m(L; \beta)$ is a known function 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) or the Dantzig selector (Candes and Tao, 2007) is convenient because they enforce a sparse solution; components of the estimate of β_0 will likely be set to zero. Alternatively, these estimators could be an intermediate step for 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 (Knight and Fu, 2000). 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 zero in certain samples but not others, and the resulting estimator of ψ_0 based on $\tilde{\beta}$ will tend to inherit this non-regular behaviour. The consequence is that standard confidence intervals (based on the normal approximation) are not 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 and Pötscher, 2005). The second concern is that the true model for $E(Y|A = 0, L)$ 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 of the observed data range, and small changes in the model may greatly impact conclusions on the treatment effect.

2.2 Doubly robust scores for conditional treatment effects

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

In order to construct our test, we will 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 (a.k.a. the propensity score when A is binary); namely $E(A|L) = \pi(L; \gamma_0)$ 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 e.g. $\pi(L; \gamma_0) = \text{expit}(\gamma_0^T L)$. Our test statistic will then be based on the score

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

(Robins et al., 1992). Here, $\eta = (\gamma^T, \beta^T)^T$ and $H(\psi) = Y\text{expit}(-\psi A)$ if $g(\cdot)$ is the log link; otherwise $H(\psi) = Y - \psi A$. In what follows, we will 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, using the law of iterated expectation one can show that $E\{U(\psi_0, \eta_0)\} = 0$ if either $E(Y|A = 0, L) = m(L; \beta_0)$ or $E(A|L) = \pi(L; \gamma_0)$; this property of double robustness will be key for obtaining uniformly valid inference (even when model $\mathcal{B}$ is misspecified).

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

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

where $\hat{V}$ is the empirical estimate of the variance of $U(\psi_0, \hat{\eta})$ and $\hat{\eta}(\psi_0)$ makes 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 Section 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 will 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 $\pi(L; \gamma_0) = \text{expit}(\gamma_0^T L)$; because our setting is high-dimensional, we will estimate γ_0 by fitting this model with a Lasso penalty e.g. we solve the minimization problem:

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

where $\lambda_\gamma > 0$ is the penalty parameter and $\|.\|_1$ denotes the ℓ_1 norm. To improve finite sample performance, in practice we recommend refitting this model adjusted for the selected covariates using maximum likelihood. Unfortunately, in the asymptotic distribution of the score $U(\psi,\eta)$, terms like

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\gamma}U_i\{\psi_0,\hat{\eta}(\psi_0)\}\sqrt{n}(\gamma_0-\hat{\gamma}) \tag{3}$$

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

We therefore recommend constructing an estimator $\hat{\beta}(\psi_0)$ at which:

$$0=\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\gamma}U_i\{\psi_0,\hat{\eta}(\psi_0)\}+\lambda_\beta\delta|\hat{\beta}(\psi_0)|^{\delta-1}\circ\text{sign}\{\hat{\beta}(\psi_0)\} \tag{4}$$

(Fu, 2003). Here $\lambda_\beta>0$, $\delta\geq 1$, $\circ$ is the Hadamard product operator and for a vector $a\in\mathbb{R}^p$, $\text{sign}(a)$ is a vector of elements $\text{sign}(a_j)$ (for $j=1,...,p$). Also, $\delta|\beta|^{\delta-1}\circ\text{sign}(\beta)$ refers to the partial derivative of $||\beta||_\delta^\delta$ with respect to β; the ℓ_δ norm is defined as $||a||_\delta\equiv\left(\sum_{i=1}^p|a_i|^\delta\right)^{1/\delta}$. Then we define β_0 as the solution to the population analogue of the above estimating equations (without the penalty term) that corresponds with the truth when model $\mathcal{B}$ is correct. Above, the gradient $\partial U(\psi_0,\hat{\eta}(\psi_0)/\partial\gamma$ is used as an *estimating function* for β_0, so as to ensure that $\sum_{i=1}^n\partial U_i(\psi_0,\hat{\eta}(\psi_0)/\partial\gamma$ is close to zero at the estimator of the nuisance parameter. Specifically, letting $w(L;\gamma)=\text{expit}(\gamma^TL)\{1-\text{expit}(\gamma^TL)\}$, we are proposing to estimate β_0 as the solution to

$$0=-\frac{1}{n}\sum_{i=1}^{n}w(L_i;\hat{\gamma})\{H_i(\psi_0)-m(L_i;\beta)\}L_i+\lambda_\beta\delta|\beta|^{\delta-1}\circ\text{sign}(\beta). \tag{5}$$

We will let δ converge to 1, in order that a sparse solution will be returned and the procedure can be implemented using software for the Lasso.

Estimating β_0 as described above ensures that

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\gamma}U_i\{\psi_0,\hat{\eta}(\psi_0)\}\right\|_\infty=||\lambda_\beta\delta|\hat{\beta}(\psi_0)|^{\delta-1}\circ\text{sign}\{\hat{\beta}(\psi_0)\}||_\infty\leq\delta\lambda_\beta$$

since $||\delta|\hat{\beta}(\psi_0)|^{\delta-1}\circ\text{sign}\{\hat{\beta}(\psi_0)\}||_\infty\leq 1$ for $\delta\to 1+$. So for penalty terms satisfying the standard condition that $\lambda_\beta=O(\sqrt{\log(p\vee n)/n})$ (where $a\vee b$ denotes the maximum of a

and b) and assuming $\log(p \vee n) = o(n)$, it follows that the ℓ_∞ norm of the gradient term asymptotically goes to zero.

More generally, we require estimators of β that have the property that

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \gamma} U_i\{\psi_0, \hat{\eta}(\psi_0)\} \right\|_\infty \leq C\lambda_\beta \tag{6}$$

where C is a constant and λ_β is a positive tuning parameter that converges to zero as $n \to \infty$. Since this idea is developed from the theory of bias-reduced doubly robust estimation (Vermeulen and Vansteelandt, 2015), we describe this as the 'high-dimensional bias reduction' property. Whilst in the previous work, the procedure was motivated by preventing the inflation of asymptotic bias under misspecification of one or both working models, here we use it to minimise the impact of using a non-regular estimator of γ_0.

In what follows (and in the proofs in Appendix A), we will focus on estimating β_0 via a bridge penalty, letting $\delta \to 1+$. However, there exist other estimators which also obtain the 'high-dimensional bias reduction' property. If one is happy to postulate a linear outcome model for $\mathcal{B}$ e.g. $m(L;\beta) = \beta^T L$, then one can use a Dantzig-based estimator of β:

$$\hat{\beta}(\psi_0) = \arg\min_\beta \|\beta\|_1 \quad \text{s.t.} \quad \left\| \frac{1}{n} \sum_{i=1}^{n} w(L_i;\hat{\gamma})\{H_i(\psi_0) - \beta^T L_i\} L_i \right\|_\infty \leq \lambda_\beta \tag{7}$$

Similar to (5), this also sets the relevant gradient approximately to zero. It is a generalization of the proposal of Ning and Liu (2017), where they use a similar approach in fitting a model for the exposure. If we are not willing to assume ultra-sparsity in a model for Y, then one can also adapt the methodology of Zhu and Bradic (2016), basing estimation of

β_0 on the solution path of the linear program:

$$\begin{aligned}\hat{\beta}(\psi_0) = \arg\min \|\beta\|_1 \quad \text{s.t.} \quad & \left\| \frac{1}{n}\sum_{i=1}^{n} w(L_i;\hat{\gamma})\{H_i(\psi_0) - \beta^T L_i\}L_i \right\|_\infty \leq \lambda_\beta \\ & \left\| \sum_{i=1}^{n} \{H_i(\psi_0) - \beta^T L_i\} \right\|_\infty \leq \kappa \\ & \frac{1}{n}\sum_{i=1}^{n} w(L_i;\hat{\gamma})\{H_i(\psi_0) - \beta^T L_i\}H_i(\psi_0) \geq \bar{\kappa}\end{aligned}$$

where κ and $\bar{\kappa}$ are positive tuning parameters. Compared with (7), this approach includes two extra constraints; these allow for the outcome model $\mathcal{B}$ to be dense by controlling the relevant remainder terms in the distribution of ψ_0. We do not consider either of the estimation approaches described in this paragraph any further in this paper, since it's currently unclear how feasible they are for non-linear models.

We close this section by noting that in low-dimensional settings, when estimating the variance of the doubly robust estimator, gradient terms like (3) are often ignored. The motivation is that if an efficient estimator of γ_0 is used, then pretending the propensity score is known will generally yield conservative inferences when model $\mathcal{B}$ is misspecified (Robins et al., 1992). However, as far as we are aware, this result currently has no analogue in the high-dimensional setting (where sparse estimation of γ_0 is required) and thus using estimators of β_0 that lack the 'high-dimensional bias reduction' property may not be guaranteed to yield intervals that exceed their nominal coverage level. Moreover, we would expect reduced bias for an estimator of ψ_0 based on $\hat{\beta}(\psi_0)$ when model $\mathcal{B}$ is grossly misspecified, relative to using an arbitrary sparse estimator of β (given the results in Vermeulen and Vansteelandt (2015)). This is because the weights $w(L;\gamma)$ will make the resulting estimator less prone to extrapolating outside of the observed data range (since regions of low overlap in L will be given weights close to zero).

2.4 Inverting the score test

Plugging in estimates $\hat{\eta}$ of η_0 and scaling $U\{\psi_0, \hat{\eta}(\psi_0)\}$, one can now obtain a statistic $T_n\{\psi_0, \hat{\eta}(\psi_0)\}$. Given the conditions discussed in the following section, we will argue that by the form of the score equation and the choice of estimators of η_0, it follows that under model $\mathcal{M} \cap \mathcal{A}$, $T_n^2\{\psi_0, \hat{\eta}(\psi_0)\} \xrightarrow{p} \chi_1^2$ where χ_1^2 denotes a chi-squared distribution on 1 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

$$[\hat{l}_s, \hat{u}_s] = \left(\psi_0 : \left[\frac{1}{n}\sum_{i=1}^{n} U_i\{\psi_0, \hat{\eta}(\psi_0)\}\right]^2 - \frac{\chi_1^2(\alpha)}{n}\hat{V} \leq 0\right) \tag{8}$$

where $\chi_1^2(\alpha)$ is the critical value of χ_1^2 corresponding to the significance level α. In practice, we will search over a grid of values of ψ in order to find the values l_s and u_s that satisfy the above inequality; note that β_0 will be re-estimated under each considered ψ. 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 following 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

Let $\mathcal{P}'$ be the class of laws that obey the intersection submodel $\mathcal{M} \cap \mathcal{A}$; then we are interested in convergence under a sequence of laws $P_n \in \mathcal{P}'$. We will allow for p to increase with n, and for 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 the notation will be suppressed for convenience). Note that at a given n, we will assume the existence of a sparse parameter

β_0 that is the solution to the unpenalised population analogue of the equations in (4). We use $\mathbb{P}_{P_n}[]$ to denote a probability taken with respect to the local data generating process P_n. 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 use the following result to show that $\hat{l}_s$ and $\hat{u}_s$ in (8) form a uniformly valid confidence interval; the proofs of all results are left to Appendix A.

Theorem 1. *If, in addition to Assumptions 1-4 in Appendix A,*
(i) $(s_\gamma^2 + s_\beta^2)\log^2(p \vee n) = o(n)$
holds, then - using estimators $\hat{\gamma}$ *and* $\hat{\beta}(\psi_0)$ *defined in (2) and (4) - we have*

$$\lim_{n\to\infty} \sup_{P_n \in \mathcal{P}'} \left|\mathbb{P}_{P_n}\left(\psi_0 \in [\hat{l}_s, \hat{u}_s]\right) - (1-\alpha)\right| = 0 \tag{9}$$

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

This result shows that under 'ultra-sparse' regimes ($s_\gamma << \sqrt{n}$ and $s_\beta << \sqrt{n}$), 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 regularisation bias incurred via the sparse estimate $\hat{\gamma}$, similar to the literature on de-biasing the Lasso (Belloni et al., 2016; Ning and 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 $\mathcal{B}$.

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 same conditions as Theorem 1, the confidence interval* $[\hat{l}_s, \hat{u}_s]$ *is uniformly valid as in (9) 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 should contain the true parameter with probability determined by the nominal α-level when either model $\mathcal{A}$ or $\mathcal{B}$ is correct, uniformly over the parameter space. Stronger conditions on s_γ and s_β are not required. In principle, uniformly doubly robust confidence intervals could be constructed when the outcome model is non-linear. However, this is challenging computationally as estimating γ_0 now requires weights dependent on $\hat{\beta}(\psi_0)$ such that iteration is required. This would have to be done over all values of ψ considered in solving (8).

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 consideration to the class of laws $\mathcal{P}$ that obey the intersection sub-model $\mathcal{M} \cap \mathcal{A} \cap \mathcal{B}$. We also suppose that*

(ii) $(s_\gamma + s_\beta)\log(p \vee n) = o(n)$

(iii) $(s_\gamma s^*)\log^2(p \vee n) = o(n)$

(iv) $H(\psi_0) \perp\!\!\!\perp A|L$

hold, where $s^* = s_\gamma \vee s_\beta$. *Then if* $m(L;\beta_0)$ *is linear in* β_0*, under Assumptions 1, 2, 4, 5 and 6 in Appendix A and the conditions (ii)-(iv), the confidence interval* $[\hat{l}_s, \hat{u}_s]$ *is uniformly valid as in (9). For general models for* $m(L;\beta_0)$*, the same result holds if* β_0 *is estimated from a subsample of the data separate to the one used to construct the interval, without requiring condition (iv).*

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

$$Y = \psi_0 A + \epsilon \tag{10}$$

where $\epsilon \perp\!\!\!\perp A|L$. If L is sufficient to adjust for confounding of the effect of A on Y, we can rephrase model (10) as a *linear structural distribution model* (Robins, 1997).

With non-linear $m(L;\beta_0)$, 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. (2019) imply that confidence intervals obtained via our procedure without weighting are valid under the intersection submodel if conditions (ii), (iii) and (iv) hold (see also the corollary below). It also follows from the proofs in Appendix A that without sample splitting, so long as all models are correct, ultra-sparsity is only required in model $\mathcal{B}$ a.k.a we require $s_\gamma \log(p \vee n) = o(n)$ and $s_\beta^2 \log^2(p \vee n) = o(n)$.

When $H(\psi_0) = Y - \psi_0 A$, Chernozhukov et al. (2018) arrive at conditions (ii) and (iii) without requiring (iv) 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 only apparent when estimators of both β_0 and γ_0 converge to the truth, and it may become infeasible with limited sample sizes or in more complex causal inference settings (see section 5). When their score equations are not linear in the target parameter, the regularity conditions in Chernozhukov et al. (2018) are less intuitive even when combined with sample splitting, whereas the confidence intervals proposed above 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 Wald-based intervals. To see why, note that for testing $\psi = \psi_0$, a Wald test based on the score $U(\psi, \eta)$ would require fitting a model for $E(Y|A, L)$ and discarding an initial estimate $\check{\psi}$ of ψ. The resulting estimates $\check{\beta}$ of β are then dependent on $(A_i)_{i=1}^n$, whereas $\hat{\beta}(\psi_0)$ (as proposed above) can only depend on the exposure data via the transformed outcome $H(\psi_0)$. In the proof of Theorem 2, we exploit this property - combined with the conditional independence of $H(\psi_0)$ and A from (iv) - to emulate settings where β is estimated in a separate sample. We note that the construction of intervals in high-dimensional models by inverting tests has been described before e.g. in Chernozhukov et al. (2015), although there it was used more as a pedagogic device rather than specifically to weaken sparsity assumptions.

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 $E(A|L) = \pi(L)$ and $E(Y|A = 0, L) = m(L)$ respectively via machine learning estimators and we repeat the above steps, inverting the test statistic $T_n\{\psi_0, \hat{\pi}, \hat{m}(\psi_0)\}$ (replacing $T_n\{\psi_0, \hat{\eta}(\psi_0)\}$) to obtain an interval $[\check{l}_s, \check{u}_s]$. Furthermore, we will assume 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 so long as Assumption 1 and (iv) holds, under the class of laws $\mathcal{P}$, $[\check{l}_s, \check{u}_s]$ is a uniformly valid interval as defined above.

By inverting a score test and utilising the location-shift condition, one can use arbitrary machine learning estimators in constructing the interval without having to either use sample splitting or invoke strong Donsker-type conditions. For a discussion of which estimators meet these conditions, see Chernozhukov et al. (2018). We emphasise this only applies so long as both estimators converge to the truth.

4 Simulation study

In each of the 1,000 simulations comprising our Experiment 1, we created a dataset with $n = 200$ observations. We generated the covariates L^* from a multivariate normal distribution $\mathcal{N}(0, \Sigma)$, where Σ is a Toeplitz matrix with $\Sigma_{j,k} = 2^{-|j-k|}$; then we created L by including an additional column for the intercept. The dimension of L was $p = 200$. Further, A was a Bernoulli random variable with conditional expectation $E(A|L) = \text{expit}(\gamma_0^T L)$ and Y was generated from the normal distribution, $\mathcal{N}(0.3A + \beta_0^T L, 1)$, where $\beta_0 = \tau(-1, 1, -1, 2^\rho, ..., (p-2)^\rho)$. 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 ρ to control sparsity (with 2 indicating a sparser model and 0.5 more dense). In this and all subsequent experiments, $\gamma_0 = 1, -1, 1, -2^{-2}, ..., (p-2)^{-2}$. In Experiment 2, we created the covariates $X_1 = |\log(5+L_1^*)|$, $X_2 = L_2^* \exp(L_1^*)$ and $X_3 = -(L_2^* + L_3^*)^2$. A matrix X was created 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 $N(0.3A + \bar{\beta}_0^T X, 1)$, where $\bar{\beta}_0$ was the same as β_0 except the leading three entries were equal to 1. Experiment 3 was similar to Experiment 1, except now $Y \sim N\{0.3A + \beta_0^T L, \sigma(A, L)\}$, where $\sigma(A_i, L_i) = \{n^{-1} \sum_i^n (0.3A_i + \beta_0^T L_i)^2\}^{-1/2}(0.3A_i + \beta_0^T L_i)$. Each experiment was repeated with $p = 250$, varying τ and ρ.

We first considered a naïve post-selection approach, where Y was regressed on A and L using a Lasso-penalty (selected via 20-fold cross-validation), forcing the exposure into the model. The final model was then refit, adjusted for A and the selected covariates, yielding the estimate $\hat{\psi}_{OLS}$. We compared this with the 'post-double selection' method, as described in Belloni et al. (2014), and implemented using the 'hdm' package in R

(Chernozhukov et al., 2016), as well the 'partialling out' method in the same package (yielding the estimators $\hat{\psi}_{PDS}$ and $\hat{\psi}_{PO}$ respectively). 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 instead obtained via cross-validation (let $\hat{\psi}_{PDS-CV}$ and $\hat{\psi}_{PO-CV}$ denote the respective estimators). For the estimators $\hat{\psi}_{OLS}$, $\hat{\psi}_{PDS}$, $\hat{\psi}_{PO}$, $\hat{\psi}_{PDS-CV}$ and $\hat{\psi}_{PO-CV}$, we used 'model-based' variance estimators given by the software.

For our approach, each working model was adjusted for L. Lasso penalties for the working models for exposure and outcome were both selected using 20-fold cross-validation, choosing λ_γ as the value that minimised the expected cross-validated error (and likewise for λ_β). In the case of the outcome model, cross-validation was done under the null $\psi = 0$, which we would expect to generally yield anti-conservative penalties. If too many covariates were selected such that refitting model $\mathcal{A}$ using maximum likelihood failed, a small increment was added to λ_γ. 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 (iv) 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 $\hat{\psi}_{HDBR}$) and confidence intervals were otherwise obtained as in Section 2.4. To compare the efficiency of $\hat{\psi}_{HDBR}$ with the other estimators in the 'MSE' 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.

Results for the three experiments are given in Tables 1-3. They indicate that even in highly sparse, strong signal settings where the model for Y is correctly specified, the naïve approach still 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

Table 1: Simulation results from Experiment 1 ($n = 200$). Estimators considered (Est); Monte Carlo bias multiplied by 10 (Bias); Monte Carlo standard deviation multiplied by 10 (MCSD); Mean standard error multiplied by 10 (MSE); coverage probability multiplied by 100 (Cov).

		$p = 200$				$p = 250$			
ρ, τ	Est	Bias	MCSD	MSE	Cov	Bias	MCSD	MSE	Cov
2,1	$\hat{\psi}_{OLS}$	-0·78	2	1·6	85·4	-0·94	1·9	1·5	83·8
	$\hat{\psi}_{PDS}$	-0·5	1·9	1·7	90·6	-0·6	1·9	1·7	91·3
	$\hat{\psi}_{PO}$	-0·82	1·7	1·6	91·1	-0·91	1·7	1·6	91·3
	$\hat{\psi}_{PDS-CV}$	-0·73	1·9	1·8	91·9	-0·81	1·8	1·8	92·1
	$\hat{\psi}_{PO-CV}$	-0·91	1·8	1·6	88	-1·05	1·7	1·6	86·3
	$\hat{\psi}_{HDBR}$	-0·72	2·1	1·9	92·2	-0·86	2·1	1·9	90·7
0·5,1	$\hat{\psi}_{OLS}$	-1·14	3·1	2	78·7	-1·69	3·2	2·1	73·8
	$\hat{\psi}_{PDS}$	-5·71	3	2·8	44	-5·81	3·1	2·8	43·6
	$\hat{\psi}_{PO}$	-5·71	2·9	2·7	44·8	-5·82	3	2·8	43
	$\hat{\psi}_{PDS-CV}$	-1·18	2·4	2·3	90·2	-1·28	2·6	2·4	89·1
	$\hat{\psi}_{PO-CV}$	-1·73	1·7	1·5	67·2	-1·76	2	1·7	69
	$\hat{\psi}_{HDBR}$	-1·13	2·6	2·3	92·3	-1·5	2·9	2·6	91·2
2,0·4	$\hat{\psi}_{OLS}$	-1·56	2·1	1·5	74·7	-1·76	2·2	1·5	68·4
	$\hat{\psi}_{PDS}$	-2·78	1·6	1·6	59·6	-2·9	1·7	1·7	57·4
	$\hat{\psi}_{PO}$	-2·78	1·6	1·7	60·9	-2·91	1·7	1·7	57·6
	$\hat{\psi}_{PDS-CV}$	-0·72	1·9	1·8	92·2	-0·77	2·1	1·9	88·2
	$\hat{\psi}_{PO-CV}$	-0·84	1·9	1·7	89·7	-0·85	2	1·8	86·7
	$\hat{\psi}_{HDBR}$	-0·66	2·1	2	92·6	-0·68	2·1	2·1	92·1
0·5,0·4	$\hat{\psi}_{OLS}$	-2·18	2·3	1·7	68·9	-2·34	2·2	1·7	65·1
	$\hat{\psi}_{PDS}$	-2·95	1·9	1·8	62·9	-3	1·8	1·8	61·1
	$\hat{\psi}_{PO}$	-2·95	1·9	1·8	63·7	-3·01	1·8	1·8	61·7
	$\hat{\psi}_{PDS-CV}$	-1·08	2·2	2	89·6	-1·2	2·1	2	87·5
	$\hat{\psi}_{PO-CV}$	-1·16	2·1	1·9	85·3	-1·31	2·1	1·9	83·8
	$\hat{\psi}_{HDBR}$	-1·02	2·3	2·2	91·4	-1·06	2·4	2·3	92·7

Table 2: Simulation results from Experiment 2 ($n = 200$). Estimators considered (Est); Monte Carlo bias multiplied by 10 (Bias); Monte Carlo standard deviation multiplied by 10 (MCSD); Mean standard error multiplied by 10 (MSE); coverage probability multiplied by 100 (Cov)

		$p = 200$				$p = 250$			
ρ, τ	Est	Bias	MCSD	MSE	Cov	Bias	MCSD	MSE	Cov
2,1	$\hat{\psi}_{OLS}$	-1·5	3·8	2·7	82·2	-1·43	4	2·6	80·5
	$\hat{\psi}_{PDS}$	-3·6	2·9	2·9	79·5	-3·6	3	2·9	78·9
	$\hat{\psi}_{PO}$	-3·6	2·9	3·1	84·3	-3·6	3	3·1	83·6
	$\hat{\psi}_{PDS-CV}$	-3·49	3·1	3·2	81·3	-3·58	3·2	3·2	81
	$\hat{\psi}_{PO-CV}$	-3·46	3	3	79·6	-3·56	3·1	3	80·1
	$\hat{\psi}_{HDBR}$	0·3	3	2·9	95	0·24	3·1	3	93·6
0·5,1	$\hat{\psi}_{OLS}$	-2·76	6·4	5·3	87·5	-1·99	6·3	5·3	89·4
	$\hat{\psi}_{PDS}$	-7·17	5·9	5·9	78·2	-6·81	5·8	5·9	81·2
	$\hat{\psi}_{PO}$	-7·17	5·9	6·4	84	-6·81	5·8	6·5	85·6
	$\hat{\psi}_{PDS-CV}$	-6·85	6·3	6·7	84·8	-6·4	6·6	6·9	85·5
	$\hat{\psi}_{PO-CV}$	-6·84	6·3	6·6	83·7	-6·31	6·6	6·7	84·7
	$\hat{\psi}_{HDBR}$	-0·17	5·9	6	95·1	0·23	6·3	6·2	95·4
2,0·4	$\hat{\psi}_{OLS}$	0·16	2·3	1·7	85	0·18	2·3	1·7	86·5
	$\hat{\psi}_{PDS}$	-1·41	2	2	89	-1·46	2·1	2	88·2
	$\hat{\psi}_{PO}$	-1·41	2	2·1	90·8	-1·47	2·1	2·1	89·9
	$\hat{\psi}_{PDS-CV}$	-1·51	2·1	2·1	88·9	-1·58	2·2	2·1	87·2
	$\hat{\psi}_{PO-CV}$	-1·55	2·1	2·1	87·1	-1·61	2·2	2·1	85·8
	$\hat{\psi}_{HDBR}$	0·06	2·4	2·3	94·8	0·1	2·5	2·3	93·3
0·5,0·4	$\hat{\psi}_{OLS}$	-0·83	3	2·5	89·5	-0·77	2·9	2·5	91·2
	$\hat{\psi}_{PDS}$	-2·73	2·8	2·9	85·4	-2·84	2·9	2·9	83·6
	$\hat{\psi}_{PO}$	-2·73	2·8	3·1	88·8	-2·84	2·9	3·1	87
	$\hat{\psi}_{PDS-CV}$	-2·72	3·1	3·2	87·8	-2·87	3·3	3·3	86·9
	$\hat{\psi}_{PO-CV}$	-2·74	3	3·1	86·7	-2·9	3·2	3·2	85
	$\hat{\psi}_{HDBR}$	0·09	3·3	3·2	94·4	0·04	3·4	3·2	94·9

Table 3: Simulation results from Experiment 3 ($n = 200$). Estimators considered (Est); Monte Carlo bias multiplied by 10 (Bias); Monte Carlo standard deviation multiplied by 10 (MCSD); Mean estimated standard error multiplied by 10 (MSE); coverage probability multiplied by 100 (Cov).

		$p = 200$				$p = 250$			
ρ, τ	Est	Bias	MCSD	MSE	Cov	Bias	MCSD	MSE	Cov
2,1	$\hat{\psi}_{OLS}$	-0·87	1·7	1·6	89·8	-0·95	1·8	1·5	84·9
	$\hat{\psi}_{PDS}$	-1·09	2·3	1·5	83·7	-1·18	2·4	1·5	81·3
	$\hat{\psi}_{PO}$	-1·36	2·1	1·6	85·6	-1·45	2·2	1·6	83·8
	$\hat{\psi}_{PDS-CV}$	-0·81	1·7	1·8	94	-0·88	1·9	1·8	90·7
	$\hat{\psi}_{PO-CV}$	-1	1·6	1·6	88·5	-1·04	1·8	1·6	85·7
	$\hat{\psi}_{HDBR}$	-0·89	1·8	1·7	89·9	-0·88	1·8	1·7	88·7
0·5,1	$\hat{\psi}_{OLS}$	-1·26	3·1	2	78·5	-1·76	3·2	2·1	72·6
	$\hat{\psi}_{PDS}$	-5·75	3·1	2·7	41·5	-6·07	3	2·7	39·8
	$\hat{\psi}_{PO}$	-5·76	2·9	2·7	43·7	-6·07	2·9	2·8	40·5
	$\hat{\psi}_{PDS-CV}$	-1·24	2·3	2·3	91·3	-1·43	2·5	2·4	88·5
	$\hat{\psi}_{PO-CV}$	-1·73	1·7	1·5	67·1	-1·88	1·9	1·6	63·5
	$\hat{\psi}_{HDBR}$	-1·21	2·5	2	91	-1·49	2·7	2·3	91·1
2,0·4	$\hat{\psi}_{OLS}$	-1·59	2·3	1·5	69·6	-1·59	2·3	1·5	70·7
	$\hat{\psi}_{PDS}$	-2·9	1·7	1·6	54·4	-2·84	1·7	1·6	57
	$\hat{\psi}_{PO}$	-2·91	1·6	1·7	57·5	-2·85	1·7	1·7	59·1
	$\hat{\psi}_{PDS-CV}$	-0·69	2	1·8	90·9	-0·79	2·1	1·8	90·1
	$\hat{\psi}_{PO-CV}$	-0·77	1·9	1·8	89·3	-0·91	2	1·8	87·6
	$\hat{\psi}_{HDBR}$	-0·6	2·1	2·1	92·9	-0·7	2·2	2·1	91·1
0·5,0·4	$\hat{\psi}_{OLS}$	-2·12	2·3	1·7	66·4	-2·31	2·2	1·7	67
	$\hat{\psi}_{PDS}$	-3·02	1·8	1·8	59·3	-3·1	1·8	1·8	58·4
	$\hat{\psi}_{PO}$	-3·03	1·8	1·8	60·5	-3·1	1·8	1·8	60·4
	$\hat{\psi}_{PDS-CV}$	-1·12	2·1	2	88·2	-1·26	2·2	2	87·3
	$\hat{\psi}_{PO-CV}$	-1·22	2	1·9	84·7	-1·31	2·1	1·9	83·3
	$\hat{\psi}_{HDBR}$	-1·02	2·2	2·1	89·6	-1·17	2·4	2·2	89·2

methods performed better in this case, 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 for post-double selection, which performed surprisingly well in Experiment 3, despite the fact that the variance estimator used is not generally robust to heteroscedastic errors. 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, we saw that 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 generally improved upon alternatives. Experiments 1-3 were repeated with $n = p = 400$, where superior coverage was seen across all settings (see Appendix B).

5 Extensions

5.1 Effect heterogeneity and categorical exposures

Suppose that interest lies in the exposure effect parameter $\psi = (\psi^{(1)}, \psi^{(2)})^T$ indexing the semiparametric model $\mathcal{M}_{int}$ defined by

$$g\{E(Y|A = a, L = l)\} - g\{E(Y|A = 0, L = l)\} = \psi_0^{(1)} a + \psi_0^{(2)} az.$$

Here, Z is a scalar component of L. We can now redefine $H(\psi_0)$, such that $H(\psi_0) = Y - \psi_0^{(1)} A - \psi_0^{(2)} AZ$ when $g(\cdot)$ is the identity link and $H(\psi_0) = Y \exp(-\psi_0^{(1)} A - \psi_0^{(2)} AZ)$ when $g(\cdot)$ is the log link.

Then because ψ_0 is now two-dimensional, we recommend the use of different nuisance parameters in the first and second doubly robust estimating functions for ψ_0. In particular, consider the following equations:

$$U(\psi,\eta) = \begin{pmatrix} \{A-\pi(L;\gamma)\}\{H(\psi)-m(L;\beta^{(1)})\} \\ Z\{A-\pi(L;\gamma)\}\{H(\psi)-m(L;\beta^{(2)})\} \end{pmatrix} = \begin{pmatrix} U^{(1)}(\psi,\eta) \\ U^{(2)}(\psi,\eta) \end{pmatrix}.$$

Following the reasoning in Section 2.3, we use the gradient of $U^{(1)}(\psi,\eta)$ with respect to γ as an estimating function for $\beta_0^{(1)}$ (and likewise for $\beta_0^{(2)}$). So we now estimate $\beta_0^{(1)}$ and $\beta_0^{(2)}$ as the solutions respectively to

$$0 = -\sum_{i=1}^{n} w(L_i;\hat{\gamma})\{H_i(\psi_0)-m(L_i;\beta^{(1)})\}L_i + \lambda_{\beta^{(1)}}\delta|\beta^{(1)}|^{\delta-1} \circ \text{sign}(\beta^{(1)}) \tag{11}$$

$$0 = -\sum_{i=1}^{n} w(L_i;\hat{\gamma})\{H_i(\psi_0)-m(L_i;\beta^{(2)})\}L_i Z_i + \lambda_{\beta^{(2)}}\delta|\beta^{(2)}|^{\delta-1} \circ \text{sign}(\beta^{(2)}). \tag{12}$$

By allowing β to take on different values in the different estimating equations for ψ, we enable the targeting of the nuisance parameter estimates towards the different parameters of interest. We now search over a two-dimensional space for ψ, estimating $\beta_0^{(1)}$ and $\beta_0^{(2)}$ at each value considered. Then $T_n\{\psi_0, \hat{\eta}(\psi_0)\}$ can be compared to a χ_2^2 distribution, and inverting the test statistic yields a confidence interval for $\psi_0^{(1)}$ and $\psi_0^{(2)}$ that is uniformly valid under model $\mathcal{M}_{int} \cap \mathcal{A}$ (following the proof of Theorem 1). Our proposal contrasts with other approaches for estimating interaction effects in high-dimensional models (e.g. Belloni et al. (2014) and Chernozhukov et al. (2018)), where separate models are postulated for $E(A|L)$ and $E(AZ|L)$. In contrast, we only require a single exposure model $\mathcal{A}$, which is also computationally more efficient, not to mention more robust if the outcome model is misspecified. For instance, if linear models are postulated for both $E(A|L)$ and $E(AZ|L)$, then they cannot generally both be correct.

Similarly, if A is an exposure with three categories (taking values 0,1 or 2), then the semiparametric model $\mathcal{M}_{cat}$ is now

$$g\{E(Y|A=a,L=l)\} - g\{E(Y|A=0,L=l)\} = \psi_0^{(1)}a_1 + \psi_0^{(2)}a_2$$

where $A_1 = 1$ if $A = 1$ and 0 otherwise, and $A_2 = 1$ if $A = 2$ and 0 otherwise. Also, $H(\psi_0)$ is redefined accordingly e.g. $H(\psi_0) = Y - \psi_0^{(1)} A_1 - \psi_0^{(2)} A_2$ when $g(\cdot)$ is the identity link. Then the estimating functions for $\psi^{(1)}$ and $\psi^{(2)}$ are

$$U(\psi, \eta) = \begin{pmatrix} \{A_1 - \pi(L; \gamma^{(1)})\}\{H(\psi) - m(L; \beta^{(1)})\} \\ \{A_2 - \pi(L; \gamma^{(2)})\}\{H(\psi) - m(L; \beta^{(2)})\} \end{pmatrix}$$

where $\pi(L; \gamma_0^{(1)})$ is a model postulated for $P(A = 1|L) \equiv P(A_1 = 1|L)$ and $\pi(L; \gamma_0^{(2)})$ is a model postulated for $P(A = 2|L) \equiv P(A_2 = 1|L)$. Postulating a multinomial logistic model $\mathcal{A}$ for A, the parameters $\gamma_0^{(1)}$ and $\gamma_0^{(2)}$ can be estimated efficiently via the Group Lasso (see Farrell (2015) for theoretical guarantees). Then $\beta_0^{(1)}$ and $\beta_0^{(2)}$ can be estimated as in (11) and (12) (except $\hat{\beta}^{(1)}$ will only depend on $\hat{\gamma}^{(1)}$, and likewise $\hat{\beta}^{(2)}$ depends only on $\hat{\gamma}^{(2)}$).

A drawback however is that even when a linear model is postulated for the conditional mean of Y, the confidence intervals are valid under model $\mathcal{M}_{int} \cap \mathcal{A}$ (or $\mathcal{M}_{cat} \cap \mathcal{A}$) but are not uniformly doubly robust. This is because, in the case of effect modification, estimating γ as in (2) will not set $\sum_{i=1}^{n} \partial U^{(2)}(\psi_0, \hat{\eta})/\partial \beta^{(2)}$ approximately to zero. This can be addressed by estimating separate parameters $\gamma^{(1)}$ and $\gamma^{(2)}$ in the same way as is done for β above. In the case of categorical exposures, one must fit separate logistic models for $P(A = 1|L)$ and $P(A = 2|L)$ rather than using a Group Lasso approach.

5.2 Controlled direct effects

Finally, we will briefly consider a study with a variable A_1 measured at baseline along with an accompanying collection of variables L_1. These may confound the association between A_1 and an end of study outcome Y. In addition, we also measure a post-baseline variable A_2, which may be influenced by L_2 (covariates measured after baseline but prior to A_2) along with L_1 and A_1. Interest is in the causal effect of A_2 on Y, but also the effect of A_1 on Y if A_2 were fixed at zero. Under extended structural assumptions (see below), the latter corresponds to *controlled direct effect* of A_1 on Y (Robins and Greenland, 1992). One may consider A_2 as a mediator of the effect of A_1, or as a subsequent measurement of a time-varying exposure A. Furthermore, we allow L_2 to depend on the baseline variable A_1, which means that standard regression-based approaches do not generally encode the controlled direct effect of A_1 (Robins, 1986).

For time $t = 1, 2$, we will postulate the model $\mathcal{M}$:

$$\begin{aligned} & g\{E(Y|A_t = a_t, \bar{A}_{t-1} = \bar{a}_{t-1}, \bar{L}_t = \bar{l}_t)\} \\ & - g\{E(Y|A_t = 0, \bar{A}_{t-1} = \bar{a}_{t-1}, \bar{L}_t = \bar{l}_t)\} = \psi_0^{(t)} a_t; \end{aligned}$$

here, $\bar{L}_t$ and $\bar{A}_{t-1}$ denote the covariate and the exposure history up to time t, with $A_0 = \emptyset$. The model imposes the restriction that the effect $\psi_0^{(t)}$ of A_t on Y does not depend on $\bar{L}_t$ and $\bar{A}_{t-1}$ at each t. In order to give either of the above contrasts a causal interpretation, a sequential 'no unmeasured confounding' assumption is required. Namely, $\bar{L}_t$ and $\bar{A}_{t-1}$ must suffice to adjust for confounding between A_t and Y. Therefore in the two time-point setting, under this condition $\psi_0^{(2)}$ encodes the (conditional) causal effect of A_2 and Y and $\psi_0^{(1)}$ the controlled direct effect of A_1 on Y (fixing A_2 at zero).

To construct tests and confidence intervals for these parameters, we will postulate the models $\mathcal{A}_t$ for $E(A_t|\bar{A}_{t-1}, \bar{L}_t)$ e.g. $E(A_t|\bar{A}_{t-1}, \bar{L}_t) = \pi_t(\bar{A}_{t-1}, \bar{L}_t; \gamma_0^{(t)})$, and $\mathcal{B}_t$ for $E(Y|A_t = 0, \bar{A}_{t-1}, \bar{L}_t)$ e.g. $E(Y|A_t = 0, \bar{A}_{t-1}, \bar{L}_t) = m_t(\bar{A}_{t-1}, \bar{L}_t; \beta_0^{(t)})$. Inference may then be obtained via the estimating functions

$$U(\psi, \eta) = \begin{pmatrix} \{A_2 - \pi_2(\bar{L}_2, A_1; \gamma^{(2)})\}\{H_2(\psi^{(2)}) - m_2(\bar{L}_2, A_1; \beta^{(2)})\} \\ \{A_1 - \pi_1(L_1; \gamma^{(1)})\}\{H_1(\psi) - m_1(L_1; \beta^{(1)})\} \end{pmatrix}$$

where $\psi = (\psi^{(1)}, \psi^{(2)})^t$ and $\eta = (\gamma^{(1)^T}, \gamma^{(2)^T}, \beta^{(1)^T}, \beta^{(2)^T})^T$. For an identity link, $H_2(\psi^{(2)}) = Y - \psi^{(2)} A_2$ and $H_1(\psi) = Y - \psi^{(2)} A_2 - \psi^{(1)} A_1$; otherwise $H_2(\psi^{(2)}) = Y \exp(-\psi^{(2)} A_2)$ and $H_1(\psi) = Y \exp(-\psi^{(2)} A_2 - \psi^{(1)} A_1)$.

6 Discussion

In this paper, we have described how to obtain uniformly valid confidence intervals for the conditional treatment effect parameters in high-dimensional linear and log-linear models. We have unified and generalised the existing doubly robust and de-biasing approaches: unified, since our proposal adapts to the sparsity conditions in each literature, depending on the modelling assumptions one is willing to make; and generalised, since we allow for misspecification and more general model choices. This allows us to extend our work to a wide selection of problems (like estimating controlled direct effects), where one does not wish to rely on an outcome model being correct. Unfortunately, it's currently unclear how our proposal could be extended to 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; the same applies to the hazard ratio.

In future work, we will look at incorporating more general machine learning methods in the construction of confidence intervals. In general, under model $\mathcal{M} \cap \mathcal{A}$, stronger rate conditions on the estimators are required than under the intersection sub-model, which are currently available for a limited selection of estimators. These include the Lasso, post-Lasso and more recently, deep neural networks (Farrell et al., 2018). In fact, if conditions equivalent to those in Appendix A are met, it follows that deep neural networks could be substituted for the Lasso for estimating the propensity score. It is an open question whether the high-dimensional bias reduction property exists for a more general class of machine learning estimators; such a development would be useful when the conditional expectation $E(Y|A = a, L)$ is difficult to estimate well.

Acknowledgments

The first author is supported by a PhD grant from the Research Foundation - Flanders (FWO).

A Appendix A

In this Appendix, we give proofs of the main results in Section 3 of the main paper. Beginning with some notation, we use $\mathbb{E}_{P_n}[]$ for taking expectation w.r.t. the local data generating process (DGP), whereas $\mathbb{E}_n[]$ refers to sample expectations. Similarly, $\mathbb{P}_{P_n}[]$ and $\text{var}_{P_n}[]$ denote probabilities and variances taken w.r.t. the local DGP respectively. In certain places, we will use the notation $\hat{\beta}(\psi_0, \hat{\gamma})$, in order to make explicit the dependence of the estimator of β on the estimated weights. The gradients $\partial U_i\{\psi_0, \hat{\eta}(\psi_0)\}/\partial\eta$ and $\partial U_i\{\psi_0, \eta_0\}/\partial\eta$ are viewed as row vectors.

In the proofs that follow, we will make the following assumptions:

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

(i) $\mathbb{E}_{P_n}[\{H(\psi_0) - m(L;\beta_0)\}^4|A, L] < C$ *with probability approaching 1.*

(ii) $\mathbb{E}_{P_n}[|H(\psi_0) - m(L;\beta_0)|^r] < C$.

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

Assumption 2. *(Concentration bound). Let* $d(L;\beta_0) = \partial m(L;\beta_0)/\partial\beta$*; then*

$$\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{A_i - \pi(L_i;\gamma_0)\} d(L_i;\beta_0) \right\|_\infty = O_{P_n}(\sqrt{\log(p \vee n)}).$$

Assumption 3. *(Rates of convergence of the parameter estimates). For a given sequence P_n, the estimators $\hat{\gamma}$ and $\hat{\beta}(\psi_0)$ satisfy:*

(i) $\|\hat{\gamma} - \gamma_0\|_1 = O_{P_n}(s_\gamma \sqrt{\log(p \vee n)/n})$.

(ii) $\|\hat{\gamma} - \gamma_0\|_2 = O_{P_n}(\sqrt{s_\gamma \log(p \vee n)/n})$.

(iii) $\left\|\hat{\beta}(\psi_0) - \beta_0\right\|_1 = O_{P_n}(s^* \sqrt{\log(p \vee n)/n})$.

(iv) $\left\|\hat{\beta}(\psi_0) - \beta_0\right\|_2 = O_{P_n}(\sqrt{s^* \log(p \vee n)/n})$.

where $s^* = s_\gamma \vee s_\beta$.

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

(i) $\mathbb{E}_n[\{\pi(L_i; \gamma_0) - \pi(L_i; \hat{\gamma})\}^2] = O_{P_n}(s_\gamma \log(p \vee n)/n)$.

(ii) $\mathbb{E}_n\left([\{m(L_i; \beta_0) - m\{L_i; \hat{\beta}(\psi_0)\}]^2\right) = O_{P_n}(s^* \log(p \vee n)/n)$.

Assumption 5. *(Dependency on estimated weights). Assuming that $m(L; \beta_0)$ is linear in β_0, then for a given sequence P_n we have that*

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

Assumption 6. *(Regularity conditions on the errors).*
$\max_{i \leq n} |H_i(\psi_0) - m(L_i; \beta_0)| \sqrt{s_\gamma s^*} \log(p \vee n) = o(\sqrt{n})$ *with probability approaching 1.*

Assumption 1 places mild moment conditions on the residuals. Given that we are working under model $\mathcal{M} \cap \mathcal{A}$, Assumption 2 can be shown to hold using the moderate deviations theory of self-normalised sums (De la Peña et al., 2009; Belloni et al., 2012), assuming that $\log(p) = o(n^{1/3})$ and (for example) the regressors L are Gaussian or have bounded support. Alternatively, one can place sub-exponential type conditions on $\mathbb{E}_n[\partial U_i(\psi_0, \eta_0)/\partial\beta]$ as in Ning and Liu (2017). The rates in Assumption 3 and 4 are known to hold for several sparse estimators, including Lasso, post-Lasso and weighted Lasso (Belloni et al., 2014, 2016). That rate 4(i) holds for Lasso logistic regression follows from Farrell (2015); rate 4(ii) holds for weighted lasso and post-lasso estimators (Belloni et al., 2016). Note that obtaining these results implies certain restrictions on the penalties, namely that

$$\lambda_\gamma = O(\sqrt{\log(p \vee n)/n}) \tag{A.1}$$

$$\lambda_\beta = O(\sqrt{\log(p \vee n)/n}). \tag{A.2}$$

The result in Assumption 5 is shown to hold for the proposed estimators of β_0 in Dukes et al. (2019) (so long as the model is linear); we refer to that paper for a list of primitive conditions required for it to hold. Assumption 6 allows us to trade off restrictions on the distribution of the errors with stronger sparsity conditions.

Proof of Theorem 1

Proof. In order to obtain result (9), we will show each of the following in turn:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_i\{\psi_0, \hat{\eta}(\psi_0)\} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_i(\psi_0, \eta_0) + o_{P_n}(1) \tag{A.3}$$

$$V^{-1/2}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_i(\psi_0, \eta_0) \xrightarrow{d} \mathcal{N}(0, 1) \tag{A.4}$$

$$\hat{V}^{-1/2} = V^{-1/2} + o_{P_n}(1). \tag{A.5}$$

Step 1 (Asymptotic linearity). We have

$$
\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_i\{\psi_0, \hat{\eta}(\psi_0)\} - \frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_i(\psi_0, \eta_0) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{A_i - \pi(L_i;\gamma_0)\}[m(L_i;\beta_0) - m\{L_i;\hat{\beta}(\psi_0)\}] \\
&\quad + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{H_i(\psi_0) - m(L_i;\beta_0)\}\{\pi(L_i;\gamma_0) - \pi(L_i;\hat{\gamma})\} \\
&\quad + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[m\{L_i;\hat{\beta}(\psi_0)\} - m(L_i;\beta_0)]\{\pi(L_i;\hat{\gamma}) - \pi(L_i;\gamma_0)\} \\
&= \mathcal{I}_1 + \mathcal{I}_2 + \mathcal{I}_3.
\end{aligned}
$$

For $\mathcal{I}_1$, using a Taylor expansion,

$$
\mathcal{I}_1 = -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{\partial U_i(\psi_0,\eta_0)}{\partial \beta}\right\}\{\beta_0 - \hat{\beta}(\psi_0)\} + O_{P_n}(\sqrt{n}||\beta_0 - \hat{\beta}(\psi_0)||_2^2).
$$

By Assumption 3(iv) and the sparsity condition (i) stated in Theorem 1, $O_{P_n}(\sqrt{n}||\beta_0 - \hat{\beta}(\psi_0)||_2^2) = o_{P_n}(1)$. Then using Hölder's inequality,

$$
\begin{aligned}
&\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{\partial U_i(\psi_0,\eta_0)}{\partial \beta}\right\}\{\beta_0 - \hat{\beta}(\psi_0)\}\right| \\
&\leq \left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{A_i - \pi(L_i;\gamma_0)\}d(L_i;\beta_0)\right\|_\infty \left\|\beta_0 - \hat{\beta}(\psi_0)\right\|_1 \\
&= O_{P_n}(\sqrt{\log(p \vee n)}) \left\|\beta_0 - \hat{\beta}(\psi_0)\right\|_1
\end{aligned}
$$

following Assumption 2. Then by Assumption 3(iii) and condition (i), $|\mathcal{I}_1| = o_{P_n}(1)$.

Considering now $\mathcal{I}_2$,

$$
\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{H_i(\psi_0) - m(L_i;\beta_0)\}\{\pi(L_i;\gamma_0) - \pi(L_i;\hat{\gamma})\} \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[H_i(\psi_0) - m\{L_i;\hat{\beta}(\psi_0)\}]\{\pi(L_i;\gamma_0) - \pi(L_i;\hat{\gamma})\} \\
&\quad + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[m\{L_i;\hat{\beta}(\psi_0)\} - m(L_i;\beta_0)]\{\pi(L_i;\gamma_0) - \pi(L_i;\hat{\gamma})\} \\
&= \mathcal{I}_{2a} + \mathcal{I}_{2b}
\end{aligned}
$$

For $\mathcal{I}_{2a}$, by a Taylor expansion,

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[H_i(\psi_0) - m\{L_i; \hat{\beta}(\psi_0)\}]\{\pi(L_i;\gamma_0) - \pi(L_i;\hat{\gamma})\} \\
&= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{\partial U_i\{\psi_0, \hat{\eta}(\psi_0)\}}{\partial \gamma}\right](\gamma_0 - \hat{\gamma}) + O_{P_n}(\sqrt{n}||\gamma_0 - \hat{\gamma}||_2^2)
\end{aligned}$$

and

$$\begin{aligned}
&\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{\partial U_i\{\psi_0, \hat{\eta}(\psi_0)\}}{\partial \gamma}\right](\gamma_0 - \hat{\gamma})\right| \\
&\leq \sqrt{n}\left\|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial U_i\{\psi_0, \hat{\eta}(\psi_0)\}}{\partial \gamma}\right\|_\infty ||\gamma_0 - \hat{\gamma}||_1 \\
&\leq \sqrt{n}\lambda_\beta \delta ||\gamma_0 - \hat{\gamma}||_1
\end{aligned}$$

since $\left\|\delta|\hat{\gamma}|^{\delta-1} \circ \text{sign}(\hat{\gamma})\right\|_\infty \leq 1$ for $\delta \to 1+$. Then given Assumption 3(i), 3(ii), (A.1) and condition (i), it follows that $|\mathcal{I}_{2a}| = o_{P_n}(1)$. Moving onto $\mathcal{I}_{2b}$, by Hölder's inequality

$$\begin{aligned}
&\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[m\{L_i; \hat{\beta}(\psi_0)\} - m(L_i;\beta_0)]\{\pi(L_i;\gamma_0) - \pi(L_i;\hat{\gamma})\}\right| \\
&\leq \sqrt{n}\mathbb{E}_n\left([m\{L_i; \hat{\beta}(\psi_0)\} - m(L_i;\beta_0)]^2\right)^{1/2}\mathbb{E}_n[\{\pi(L_i;\gamma_0) - \pi(L_i;\hat{\gamma})\}^2]^{1/2}
\end{aligned}$$

Then given the joint sparsity condition (i) on s_γ and s_β, it follows that

$$\begin{aligned}
&\sqrt{n}\mathbb{E}_n\left([m\{L_i; \hat{\beta}(\psi_0)\} - m(L_i;\beta_0)]^2\right)^{1/2}\mathbb{E}_n[\{\pi(L_i;\gamma_0) - \pi(L_i;\hat{\gamma})\}^2]^{1/2} \\
&= o_{P_n}(1)
\end{aligned}$$

Repeating the above reasoning, we have $|\mathcal{I}_3| = o_{P_n}(1)$, and the result (A.3) follows.

Step 2 (Asymptotic normality). It follows from Assumptions 1(ii) and 1(iii) that $\mathbb{E}_{P_n}\{U_i(\psi_0, \eta_0)\}$ is bounded away from zero and above uniformly in n. Furthermore, $\mathbb{E}_{P_n}\{|U_i(\psi_0, \eta_0)|^{2+\epsilon}\} \leq C$ by Assumption 1(ii). Hence the Lyapunov condition is verified, and one can invoke the Lyapunov central limit theorem for triangular arrays to arrive at (A.4).

Step 3 (Consistency of the variance estimator). Further details on the following arguments can be found in the appendix of Dukes et al. (2019). Given that $\mathbb{E}_{P_n}\{U_i(\psi_0, \eta_0)^2\}$ can be bounded above and below uniformly in n by Assumption 1, it will suffice to prove that $\mathbb{E}_n[U_i\{\psi_0, \hat{\eta}(\psi_0)]^2\} = \mathbb{E}_{P_n}\{U_i(\psi_0, \eta_0)^2\} + o_{P_n}(1)$. One can show that

$$\mathbb{E}_n\{U_i(\psi_0, \eta_0)^2\} = \mathbb{E}_{P_n}\{U_i(\psi_0, \eta_0)^2\} + o_{P_n}(1)$$

using the Von-Bahr Esseen Inequality (von Bahr and Esseen, 1965) in combination with Assumptions 1(i) and 1(ii). Then it remains to show that $\mathbb{E}_n[U_i\{\psi_0, \hat{\eta}(\psi_0)\}^2] = \mathbb{E}_n\{U_i(\psi_0, \eta_0)^2\} = o_{P_n}(1)$.

Applying the triangle inequality,

$$\begin{aligned}
&|\mathbb{E}_n[U_i\{\psi_0, \hat{\eta}(\psi_0)\}^2] - \mathbb{E}_n\{U_i(\psi_0, \eta_0)^2\}| \\
&\leq \mathbb{E}_n[\{\pi(L_i; \hat{\gamma}) - \pi(L_i; \gamma_0)\}^2\{H_i(\psi_0) - m(L_i; \beta_0)\}^2] \\
&\quad + |2\mathbb{E}_n[\{A_i - \pi(L_i; \gamma_0)\}\{\pi(L_i; \hat{\gamma}) - \pi(L_i; \gamma_0)\}\{H_i(\psi_0) - m(L_i; \beta_0)\}^2]| \\
&\quad + \mathbb{E}_n\left([m\{L_i; \hat{\beta}(\psi_0)\} - m(L_i; \beta_0)]^2\{A_i - \pi(L_i; \hat{\gamma})\}^2\right) \\
&\quad + |2\mathbb{E}_n[\{H_i(\psi_0) - m(L_i; \beta_0)\}[m\{L_i; \hat{\beta}(\psi_0)\} - m(L_i; \beta_0)]\{A_i - \pi(L_i; \hat{\gamma})\}^2]| \\
&= \mathcal{I}_4 + \mathcal{I}_5 + \mathcal{I}_6 + \mathcal{I}_7
\end{aligned}$$

Then applying the von Bahr-Esseen inequality and given Assumptions 1(ii), 4(i), 4(ii) and condition (i), we have:

$$
\begin{aligned}
\mathcal{I}_4 \leq & \max_{i\leq n}|\pi(L_i;\hat{\gamma})-\pi(L_i;\gamma_0)|\mathbb{E}_n[\{\pi(L_i;\hat{\gamma})-\pi(L_i;\gamma_0)\}^2]^{1/2} \\
& \times \mathbb{E}_n\{|H_i(\psi_0)-m(L_i;\beta_0)|^4\}^{1/2} = o_{P_n}(1) \\
\mathcal{I}_5 \leq & 2\max_{i\leq n}|A_i-\pi(L_i;\gamma_0)|\mathbb{E}_n[\{\pi(L_i;\hat{\gamma})-\pi(L_i;\gamma_0)\}^2]^{1/2} \\
& \times \mathbb{E}_n\{|H_i(\psi_0)-m(L_i;\beta_0)|^4\}^{1/2} = o_{P_n}(1) \\
\mathcal{I}_6 \leq & \max_{i\leq n}\{A_i-\pi(L_i;\hat{\gamma})\}^2\mathbb{E}_n\left([m\{L_i;\hat{\beta}(\psi_0)\}-m(L_i;\beta_0)]^2\right) = o_{P_n}(1) \\
\mathcal{I}_7 \leq & 2\max_{i\leq n}\{A_i-\pi(L_i;\hat{\gamma})\}^2\mathbb{E}_n[\{H_i(\psi_0)-m(L_i;\beta_0)\}^2]^{1/2} \\
& \times \mathbb{E}_n\left([m\{L_i;\hat{\beta}(\psi_0)\}-m(L_i;\beta_0)]^2\right)^{1/2} = o_{P_n}(1).
\end{aligned}
$$

Step 4 (Uniform validity). Note that it is immediate from steps 1-3 that

$$T_n\{\psi_0,\hat{\eta}(\psi_0)\} \xrightarrow{d} \mathcal{N}(0,1).$$

under any sequence P_n. Then along the lines in of the proof of Proposition 1 in Chernozhukov et al. (2015), for any sequence $\delta_n \to 0$, let us consider an arbitrary sequence P_n^* where

$$\sup_{P_n\in\mathcal{P}'}|\mathbb{P}_{P_n}(\psi_0\in[l_s,u_s])-(1-\alpha)| \leq |\mathbb{P}_{P_n^*}(\psi_0\in[l_s,u_s])-(1-\alpha)|+\delta_n$$

However, by (A.3), (A.4) and (A.5) we have that

$$\mathbb{P}_{P_n^*}(\psi_0\in[l_s,u_s]) = \mathbb{P}_{P_n^*}[|T_n\{\psi_0,\hat{\eta}(\psi_0)\}|\leq\Phi(1-\alpha/2)] \to 1-\alpha$$

and the main result follows.

□

Proof of Corollary 1

Proof. We begin by noting that one can motivate the proposed estimator of γ as the solution to the penalised estimating equations

$$\begin{aligned} 0 &= \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \beta} U_i\{\psi_0, \hat{\eta}(\psi_0)\} + \lambda_\gamma \delta |\hat{\gamma}|^{\delta-1} \circ \operatorname{sign}(\hat{\gamma}) \\ &= \frac{1}{n}\sum_{i=1}^{n} -\{A_i - \pi(L_i;\hat{\gamma})\}L_i + \lambda_\gamma \delta |\hat{\gamma}|^{\delta-1} \circ \operatorname{sign}(\hat{\gamma}), \end{aligned}$$

letting $\delta \to 1+$.

Then repeating the arguments in Step 1 of Theorem 1, for $\mathcal{I}_1$ we now have

$$\begin{aligned} \mathcal{I}_1 = & \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{A_i - \pi(L_i;\hat{\gamma})\}[m(L_i;\beta_0) - m\{L_i;\hat{\beta}(\psi_0)\}] \\ & + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{\pi(L_i;\hat{\gamma}) - \pi(L_i;\gamma_0)\}[m(L_i;\beta_0) - m\{L_i;\hat{\beta}(\psi_0)\}] \end{aligned} \tag{A.6}$$

The second term in the right hand side is $o_{P_n}(1)$ under Assumption 4 and sparsity condition (i). Then

$$\begin{aligned} & \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{A_i - \pi(L_i;\hat{\gamma})\}[m(L_i;\beta_0) - m\{L_i;\hat{\beta}(\psi_0)\}] \\ & = -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{\partial U_i\{\psi_0, \hat{\eta}(\psi_0)\}}{\partial \beta}\right\}\{\beta_0 - \hat{\beta}(\psi_0)\} + O_{P_n}(\sqrt{n}||\beta_0 - \hat{\beta}(\psi_0)||_2^2). \end{aligned}$$

and

$$\begin{aligned} & \left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{\partial U_i\{\psi_0, \hat{\eta}(\psi_0)\}}{\partial \beta}\right\}\{\beta_0 - \hat{\beta}(\psi_0)\}\right| \\ & \leq \sqrt{n}\left\|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial U_i\{\psi_0, \hat{\eta}(\psi_0)\}}{\partial \beta}\right\|_\infty \left\|\beta_0 - \hat{\beta}(\psi_0)\right\|_1 \\ & \leq \sqrt{n}\lambda_\gamma \delta ||\beta_0 - \hat{\beta}(\psi_0)||_1 \end{aligned}$$

since $\left\|\delta|\hat{\beta}(\psi_0)|^{\delta-1} \circ \operatorname{sign}\{\hat{\beta}(\psi_0)\}\right\|_\infty \leq 1$ for $\delta \to 1+$. Then by Assumptions 3(iii), 3(iv), (A.2) and condition (i), $|\mathcal{I}_{1b}| = o_{P_n}(1)$. The result follows by repeating the remaining steps in the above proof. □

Proof of Theorem 2

Proof for linear models

Proof. We first consider the case where $m(L;\beta_0)$ is linear in β_0 (so that we can invoke Assumption 5). Repeating Step 1 of the proof of Theorem 1, for $\mathcal{I}_1$ we now have

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{A_i-\pi(L_i;\gamma_0)\}\{m(L_i;\beta_0)-m(L_i;\hat{\beta}(\psi_0,\hat{\gamma}))\}\\
&=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{A_i-\pi(L_i;\gamma_0)\}\{m(L_i;\beta_0)-m\{L_i;\hat{\beta}(\psi_0,\gamma_0)\}\}\\
&\quad+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{A_i-\pi(L_i;\gamma_0)\}\{m\{L_i;\hat{\beta}(\psi_0,\gamma_0)\}-m(L_i;\hat{\beta}(\psi_0,\hat{\gamma}))\}\\
&=\mathcal{I}_{1a}+\mathcal{I}_{1b}
\end{aligned}$$

Considering first $\mathcal{I}_{1a}$, by the location-shift condition (iv) in Theorem 2,

$$\begin{aligned}
&\mathbb{E}_{P_n}[\mathcal{I}_{1a}|\{H_i(\psi_0),L_i\}_{i=1}^n]\\
&=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\mathbb{E}_{P_n}[A_i|\{H_i(\psi_0),L_i\}_{i=1}^n]-\pi(L_i;\gamma_0)\right)\left[m(L_i;\beta_0)-m\{L_i;\hat{\beta}(\psi_0,\gamma_0)\right]\\
&=0
\end{aligned}$$

and

$$\mathbb{E}_{P_n}[\mathcal{I}_{1a}^2|\{H_i(\psi_0),L_i\}_{i=1}^n]\le C\mathbb{E}_n\left([m(L_i;\beta_0)-m\{L_i;\hat{\beta}(\psi_0,\gamma_0)\}]^2\right).$$

By Assumption 4(ii) and sparsity condition (ii), $\mathbb{E}_{P_n}[R_{1a}^2]=o(1)$, and therefore$|R_{1a}|=o_{P_n}(1)$.

For $\mathcal{I}_{1b}$,

$$\begin{aligned}
&\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{A_i-\pi(L_i;\gamma_0)\}[m\{L_i;\hat{\beta}(\psi_0,\gamma_0)\}-m\{L_i;\hat{\beta}(\psi_0,\hat{\gamma})\}]\right|\\
&\le\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{A_i-\pi(L_i;\gamma_0)\}L_i\right\|_\infty\left\|\hat{\beta}(\psi_0,\gamma_0)-\hat{\beta}(\psi_0,\hat{\gamma})\right\|_1\\
&=O_{P_n}(\sqrt{\log(p\vee n)})\left\|\hat{\beta}(\psi_0,\gamma_0)-\hat{\beta}(\psi_0,\hat{\gamma})\right\|_1=o_{P_n}(1)
\end{aligned}$$

by Assumptions 5, 6 and sparsity conditions (ii) and (iii). Note that if β_0 is estimated without weights, this step is not required.

Then for $\mathcal{I}_2$,

$$\begin{aligned}
&\mathbb{E}_{P_n}\{\mathcal{I}_2|(A_i, L_i)_{i=1}^n\}\\
&= \mathbb{E}_n\left(\mathbb{E}_{P_n}[\{H_i(\psi_0) - m(L;\beta_0)\}^2|(A_i, L_i)_{i=1}^n]\{\pi(L_i;\gamma_0) - \pi(L_i;\hat{\gamma})\}^2\right)\\
&\leq C\mathbb{E}_n[\{\pi(L_i;\gamma_0) - \pi(L_i;\hat{\gamma})\}^2],
\end{aligned}$$

by Assumption 1(i); then $|\mathcal{I}_2| = o_{P_n}(1)$ given condition (ii). One can show that $|\mathcal{I}_3| = o_{P_n}(1)$ under Assumption 4 and sparsity condition (iii), without requiring the condition (i) that was invoked in the proof of Theorem 1. Indeed, one can then repeat steps 2-4 of the proof of Theorem 1, invoking conditions (ii) and (iii) instead of (i), to obtain the main result. □

Proof using sample splitting

Proof. For non-linear models, for simplicity we will consider a simple scheme whereby the data is split into approximately equal subsamples k and k^c, where k has sample size $n_k = n/2$ (so $n_{k^c} = n - n_k$). This sketch proof also extends to more complex schemes as in Chernozhukov et al. (2018). We estimate β_0 using sample k^c only, such that the resulting estimates will be denoted by $\hat{\beta}^{k^c}(\psi_0)$. Let $\hat{\gamma}^k$ denote an estimate of γ_0 obtained using sample k (sample k^c could be used instead without changing the final result). Then

$$\begin{aligned}
&\frac{1}{\sqrt{n_k}}\sum_{i=1}^{n_k} U_i^k(\psi_0, \hat{\eta}^{k^c}) - \frac{1}{\sqrt{n_k}}\sum_{i=1}^{n_k} U_i(\psi_0, \eta_0) \\
&= \frac{1}{\sqrt{n_k}}\sum_{i=1}^{n_k}\{A_i^k - \pi(L_i^k;\gamma_0)\}[m(L_i^k;\beta_0) - m\{L_i^k;\hat{\beta}^{k^c}(\psi_0)\}] \\
&\quad + \frac{1}{\sqrt{n_k}}\sum_{i=1}^{n_k}\{H_i^k(\psi_0) - m(L_i^k;\beta_0)\}\{\pi(L_i^k;\gamma_0) - \pi(L_i^k;\hat{\gamma}^k)\} \\
&\quad + \frac{1}{\sqrt{n_k}}\sum_{i=1}^{n_k}[m\{L_i^k;\hat{\beta}^{k^c}(\psi_0)\} - m(L_i^k;\beta_0)]\{\pi(L_i^k;\hat{\gamma}^k) - \pi(L_i^k;\gamma_0)\}. \\
&= \check{\mathcal{I}}_1 + \check{\mathcal{I}}_2 + \check{\mathcal{I}}_3.
\end{aligned}$$

By Assumptions 1(i),

$$\begin{aligned}
\mathbb{E}_{P_n}\{\check{\mathcal{I}}_1^2|(L_i^k)_{i=1}^{n_k}, k^c\} &\leq C\mathbb{E}_n\left([m(L_i^k;\beta_0) - m\{L_i^k;\hat{\beta}^{k^c}(\psi_0)\}]^2\right) \\
\mathbb{E}_{P_n}\{\check{\mathcal{I}}_2^2|(A_i^k, L_i^k)_{i=1}^{n_k}\} &\leq C\mathbb{E}_n[\{\pi(L_i^k;\gamma_0) - \pi(L_i^k;\hat{\gamma}^k)\}^2]
\end{aligned}$$

and

$$|\check{\mathcal{I}}_3| \leq \sqrt{n}\mathbb{E}_n\left([m\{L_i^k;\hat{\beta}^{k^c}(\psi_0)\} - m(L_i^k;\beta_0)]^2\right)^{1/2}\mathbb{E}_n[\{\pi(L_i^k;\hat{\gamma}^k) - \pi(L_i^k;\gamma_0)\}^2]^{1/2}$$

Hence invoking Assumptions 4 and conditions (ii) and (iii), it follows that $\check{\mathcal{I}}_1$, $\check{\mathcal{I}}_2$ and $\check{\mathcal{I}}_3$ are all $o_{P_1}(1)$. Validity of the confidence intervals follows from repeating steps 2-4 of the proof of Theorem 1. □

B Appendix B

B.1 Additional simulation results

Table 4: Simulation results from repeating Experiments 1-3 at $n = p = 400$. Estimators considered (Est); Monte Carlo bias multiplied by 10 (Bias); Monte Carlo standard deviation multiplied by 10 (MCSD); Mean standard error multiplied by 10 (MSE); coverage probability multiplied by 100 (Cov).

			Experiment 1				Experiment 2				Experiment 3		
ρ, τ	Est	Bias	MCSD	MSE	Cov	Bias	MCSD	MSE	Cov	Bias	MCSD	MSE	Cov
2,1	$\hat{\psi}_{OLS}$	-0·54	1·4	1·1	86·2	-2·94	2·7	2	63·6	-0·59	1·2	1·1	88·3
	$\hat{\psi}_{PDS}$	-0·39	1·2	1·2	93	-3·74	2·1	2·1	57·6	-0·42	1·1	1·1	92·5
	$\hat{\psi}_{PO}$	-0·62	1·1	1·1	91·7	-3·74	2·1	2·2	62·1	-0·65	1	1·1	93·2
	$\hat{\psi}_{PDS-CV}$	-0·56	1·3	1·2	91·4	-3·63	2·2	2·2	61·7	-0·64	1·2	1·2	91·9
	$\hat{\psi}_{PO-CV}$	-0·66	1·3	1·2	89·4	-3·63	2·2	2·1	59·8	-0·77	1·1	1·2	89·2
	$\hat{\psi}_{HDBR}$	-0·34	1·4	1·4	94·6	0·14	2	2	95·6	-0·36	1·2	1·2	93·6
0·5,0·4	$\hat{\psi}_{OLS}$	-1·42	1·8	1·2	68·3	-0·95	2·2	1·8	87·2	-1·4	1·8	1·2	67·6
	$\hat{\psi}_{PDS}$	-2·2	1·5	1·3	59·3	-2·75	2	2·1	74·8	-2·3	1·5	1·3	54·4
	$\hat{\psi}_{PO}$	-2·22	1·5	1·3	58·8	-2·75	2	2·2	78·3	-2·32	1·5	1·3	55·6
	$\hat{\psi}_{PDS-CV}$	-0·78	1·4	1·4	90	-2·79	2·1	2·2	77·3	-0·72	1·4	1·4	91·2
	$\hat{\psi}_{PO-CV}$	-0·92	1·3	1·3	87·1	-2·8	2	2·2	75·9	-0·85	1·4	1·3	86·1
	$\hat{\psi}_{HDBR}$	-0·56	1·5	1·6	94·4	0·07	2·2	2·2	95·5	-0·55	1·5	1·6	93·8

References

Athey, S., Imbens, G. W., and Wager, S. (2018). Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

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(6):2369–2429.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on Treatment Effects after Selection among High-Dimensional Controls. *The Review of Economic Studies*, 81(2):608–650.

Belloni, A., Chernozhukov, V., and Wei, Y. (2016). Post-Selection Inference for Generalized Linear Models With Many Controls. *Journal of Business & Economic Statistics*, 34(4):606–619.

Benkeser, D., Carone, M., Laan, M. J. V. D., and Gilbert, P. B. (2017). Doubly robust nonparametric inference on the average treatment effect. *Biometrika*, 104(4):863–880.

Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Chernozhukov, V., Hansen, C., and Spindler, M. (2015). Valid Post-Selection and Post-Regularization Inference: An Elementary, General Approach. *Annual Review of Economics*, 7(1):649–688.

Chernozhukov, V., Hansen, C., and Spindler, M. (2016). hdm: High-Dimensional Metrics. *The R Journal*, 8(2):185–199.

De la Peña, V., Lai, T. L., and Shao, Q.-M. (2009). *Self-normalized processes: limit theory and statistical applications*. Probability and its applications. Springer, Berlin. OCLC: ocn244765605.

Dukes, O., Avagyan, V., and Vansteelandt, S. (2019). High-dimensional doubly robust tests for regression parameters. *arXiv:1805.06714 [stat]*. arXiv: 1805.06714.

Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189(1):1–23.

Farrell, M. H., Liang, T., and Misra, S. (2018). Deep Neural Networks for Estimation and Inference: Application to Causal Effects and Other Semiparametric Estimands. *arXiv:1809.09953 [cs, econ, math, stat]*. arXiv: 1809.09953.

Fu, W. J. (2003). Penalized estimating equations. *Biometrics*, 59(1):126–132.

Javanmard, A. and Montanari, A. (2014). Confidence Intervals and Hypothesis Testing for High-dimensional Regression. *J. Mach. Learn. Res.*, 15(1):2869–2909.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics*, 28(5):1356–1378.

Leeb, H. and Pötscher, B. M. (2005). Model Selection and Inference: Facts and Fiction. *Econometric Theory*, 21(1):21–59.

Ning, Y. and Liu, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics*, 45(1):158–195.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9):1393–1512.

Robins, J. M. (1997). Causal Inference from Complex Longitudinal Data. In Berkane, M., editor, *Latent Variable Modeling and Applications to Causality*, Lecture Notes in Statistics, pages 69–117. Springer New York.

Robins, J. M. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology (Cambridge, Mass.)*, 3(2):143–155.

Robins, J. M., Mark, S. D., and Newey, W. K. (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, 48(2):479–495.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of Regression Coefficients When Some Regressors are not Always Observed. *Journal of the American Statistical Association*, 89(427):846–866.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202.

van der Laan, M. J. and Rubin, D. (2006). Targeted Maximum Likelihood Learning. *The International Journal of Biostatistics*, 2(1).

Vermeulen, K. and Vansteelandt, S. (2015). Bias-Reduced Doubly Robust Estimation. *Journal of the American Statistical Association*, 110(511):1024–1036.

von Bahr, B. and Esseen, C.-G. (1965). Inequalities for the $r$th Absolute Moment of a Sum of Random Variables, $1 \leqq r \leqq 2$. *The Annals of Mathematical Statistics*, 36(1):299–303.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242.

Zhu, Y. and Bradic, J. (2016). Significance testing in non-sparse high-dimensional linear models. *arXiv:1610.02122 [math, stat]*. arXiv: 1610.02122.