Statistical Science
2020, Vol. 35, No. 3, 518–539
https://doi.org/10.1214/20-STS786

On Nearly Assumption-Free Tests of Nominal Confidence Interval Coverage for Causal Parameters Estimated by Machine Learning[1]

Lin Liu, Rajarshi Mukherjee and James M. Robins

Abstract. For many causal effect parameters of interest, doubly robust machine learning (DRML) estimators $\hat{\psi}_1$ are the state-of-the-art, incorporating the good prediction performance of machine learning; the decreased bias of doubly robust estimators; and the analytic tractability and bias reduction of sample splitting with cross-fitting. Nonetheless, even in the absence of confounding by unmeasured factors, the nominal $(1-\alpha)$ Wald confidence interval $\hat{\psi}_1 \pm z_{\alpha/2}\widehat{\mathsf{s.e.}}[\hat{\psi}_1]$ may still undercover even in large samples, because the bias of $\hat{\psi}_1$ may be of the same or even larger order than its standard error of order $n^{-1/2}$.

In this paper, we introduce essentially assumption-free tests that (i) can falsify the null hypothesis that the bias of $\hat{\psi}_1$ is of smaller order than its standard error, (ii) can provide a upper confidence bound on the true coverage of the Wald interval, and (iii) are valid under the null under no smoothness/sparsity assumptions on the nuisance parameters. The tests, which we refer to as Assumption Free Empirical Coverage Tests (AFECTs), are based on a U-statistic that estimates part of the bias of $\hat{\psi}_1$.

Our claims need to be tempered in several important ways. First no test, including ours, of the null hypothesis that the ratio of the bias to its standard error is smaller than some threshold δ can be consistent [without additional assumptions (e.g., smoothness or sparsity) that may be incorrect]. Second, the above claims only apply to certain parameters in a particular class. For most of the others, our results are unavoidably less sharp. In particular, for these parameters, we cannot directly test whether the nominal Wald interval $\hat{\psi}_1 \pm z_{\alpha/2}\widehat{\mathsf{s.e.}}[\hat{\psi}_1]$ undercovers. However, we can often test the validity of the smoothness and/or sparsity assumptions used by an analyst to justify a claim that the reported Wald interval's actual coverage is no less than nominal. Third, in the main text, with the exception of the simulation study in Section 1, we assume we are in the semisupervised data setting (wherein there is a much larger dataset with information only on the covariates), allowing us to regard the covariance matrix of the covariates as known. In the simulation in Section 1, we consider the setting in which estimation of the covariance matrix is required. In the simulation, we used a data adaptive estimator which performs very well in our simulations, but the estimator's theoretical sampling behavior remains unknown.

Key words and phrases: Causal inference, assumption-free, valid inference, U-statistics, higher-order influence functions.

Lin Liu is Assistant Professor, Institute of Natural Sciences, School of Mathematical Sciences and SJTU-Yale Joint Center for Biostatistics, Shanghai Jiao Tong University, Shanghai 200420, China (e-mail: linliu@alumni.tongji.edu.cn). Rajarshi Mukherjee is Assistant Professor, Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, Massachusetts 02115, USA (e-mail: ram521@mail.harvard.edu). James M. Robins is Mitchell L.

1. INTRODUCTION AND MOTIVATION

Valid inference (i.e., valid confidence intervals) for causal effects is of importance in many subject matter areas. For example, in medicine it is critical to evaluate whether a nonnull treatment effect estimate could differ from zero simply because of sampling variability and, conversely, whether a null treatment effect estimate is compatible with a clinically important effect.

In observational studies, control of confounding is necessary for valid inference. Historically, and assuming no confounding by unmeasured covariates, two statistical approaches have been used to control confounding by potential measured confounders, both of which require the building of noncausal purely predictive algorithms:

- One approach builds an algorithm to predict the conditional mean $b(x)$ of the outcome of interest given data on potential confounders and (sometimes) treatment (referred to as the outcome regression);
- The other approach builds an algorithm to predict the conditional probability $p(x)$ of treatment given data on potential confounders (referred to as the propensity score).

The validity of a nominal $(1-\alpha)$ Wald confidence interval (CI) $\hat{\psi}_1 \pm z_{\alpha/2}\widehat{\mathsf{s.e.}}(\hat{\psi}_1)$[2] for a parameter ψ of interest centered at a particular estimator $\hat{\psi}_1$ quite generally requires that the bias of $\hat{\psi}_1$ is much less than than its estimated standard error $\widehat{\mathsf{s.e.}}(\hat{\psi}_1)$. A nominal $(1-\alpha)$ confidence interval is said to be valid if the actual coverage rate under repeated sampling is no smaller than $(1-\alpha)$. Under either of the above approaches, obtaining estimators with small bias generally depends on good performance of the corresponding prediction algorithm. This has motivated the application of modern machine learning (ML) methods to these prediction problems for the following reason. When the vector of potential confounding factors is high-dimensional, as is now standard owing to the "big data revolution," it has become noted that, so-called machine learning algorithms (e.g., neural nets (Krizhevsky, Sutskever and Hinton, 2012), support vector machines (Cortes and Vapnik, 1995), boosting (Freund and Schapire, 1997), regression trees and random forests (Breiman, 2001), etc., especially when combined with cross-validation) can often do a much better job of prediction than traditional parametric or nonparametric approaches (e.g., kernel or series regression). However, even the best machine learning methods may fail to give predictions that are sufficiently accurate to provide nearly unbiased causal effect estimates, and thus, may fail to control bias due to confounding.

To partially guard against this possibility, so-called doubly robust machine learning (DRML) (Chernozhukov et al., 2018) estimators have been developed that can be nearly unbiased for the causal effect ψ, even when both of the above approaches fail. DRML estimators employ ML estimators of both the outcome regression $b(x)$ and the propensity score $p(x)$. DRML estimators are the state-of-the-art for estimation of causal effects, combining the benefits of sample splitting, machine learning and double robustness (Scharfstein, Rotnitzky and Robins, 1999a, 1999b, Robins and Rotnitzky, 2001, Bang and Robins, 2005). By sample splitting, we mean that the data is randomly divided into two (or more) samples— the estimation sample and the training sample. The ML estimators $\hat{b}(x)$ and $\hat{p}(x)$ of $b(x)$ and $p(x)$ are fit using the training sample data. The estimator $\hat{\psi}_1$ of our causal parameter ψ is computed from the estimation sample treating the ML estimators as fixed functions. This approach is required because the ML estimates of the regression functions generally have unknown statistical properties and, in particular, may not lie in a so-called Donsker class— a condition often needed for valid inference when sample splitting is not employed. Under conditions given in Theorem 1.4, the efficiency lost due to sample splitting can be recovered by cross-fitting. The cross-fitting estimator $\hat{\psi}_{\mathsf{cf},1}$ averages $\hat{\psi}_1$ with its "twin" obtained by exchanging the roles of the estimation and training sample. In the semiparametric statistics literature, the possibility of using sample-splitting with cross-fitting to avoid imposing Donsker conditions has a long history (Schick, 1986, van der Vaart, 1998, p. 391), although the idea of explicitly combining cross-fitting with ML was not emphasized until recently. Ayyagari (2010) Ph.D. thesis (subsequently published as Robins et al. (2013)) and Zheng and van der Laan (2011) are early examples that emphasized the theoretical and finite sample advantages of DRML estimators.

However, even the use of DRML estimators is not guaranteed to provide valid inferences owing to the possibility that the two ML prediction algorithms are not sufficiently accurate for the bias to be small compared to the standard error. In particular, if the bias of the DRML estimator is of the same (or greater) order than its standard error, the actual coverage of nominal $(1-\alpha)$ CIs for the causal effect will be smaller (and often much smaller) than the nominal level, thereby producing misleading inferences.

Suppose an author publishes a paper with a nominal $(1-\alpha)$ Wald CI $\hat{\psi}_{\mathsf{cf},1} \pm z_{\alpha/2}\widehat{\mathsf{s.e.}}(\hat{\psi}_{\mathsf{cf},1})$ for a parameter ψ. The previous discussion leads to the following question. Can $\alpha^{\dagger}$-level tests be developed that have the ability to falsify whether the bias of the DRML estimator $\hat{\psi}_1$ or $\hat{\psi}_{\mathsf{cf},1}$ is of the same or greater order than its standard error? In particular, can we provide an upper confidence

and Robin LaFoley Dong Professor, Department of Epidemiology and Biostatistics, Harvard T. H. Chan School of Public Health, Boston, Massachusetts 02115, USA (e-mail: robins@hsph.harvard.edu).

[1]Discussed in 10.1214/20-STS796; rejoinder at 10.1214/20-STS804.

[2]In this paper, we use the standard notation z_α to denote the $1-\alpha$ standard normal quantile and $\Phi(x)$ to denote the standard normal CDF.

bound on the actual coverage of a nominal $(1-\alpha)$ CI $\hat{\psi}_{\mathsf{cf},1} \pm z_{\alpha/2}\widehat{\mathsf{s.e.}}(\hat{\psi}_{\mathsf{cf},1})$? If so, when such excess bias is detected, can we construct new estimators $\hat{\psi}_2$ that are less biased? Furthermore, is it possible to construct such tests and estimators without: (i) refitting, modifying or even having knowledge of the ML algorithms that have been employed and (ii) without making any assumptions about the smoothness or sparsity of the true outcome regression $b(x)$ or propensity score function $p(x)$?

Throughout, we assume that we have been given access to the data set used to obtain both the estimate $\hat{\psi}_1$ and the estimated regression functions outputted by some ML prediction algorithms. We do not require any knowledge of or access to the ML algorithms used, other than the functions $\hat{b}(x)$ and $\hat{p}(x)$ that they outputted.

In this paper, we show that, perhaps surprisingly, for parameters in a certain class, the *monotone bias class* defined in Definition 2.2 of Section 2, the answer to these questions is "yes" by using higher-order influence function tests and estimators (Robins et al., 2008, 2017, Mukherjee, Newey and Robins, 2017). We refer to such tests as Assumption-Free Empirical Coverage Tests (AFECTs). For parameters not in the *monotone bias class*, we cannot test whether the bias of $\hat{\psi}_1$ is small compared to its standard error. The best we can do is to empirically test the author's *justification* for the claim that his intervals are valid. In general, a data analyst who reports the interval $\hat{\psi}_{\mathsf{cf},1} \pm z_{\alpha/2}\widehat{\mathsf{s.e.}}(\hat{\psi}_{\mathsf{cf},1})$ justifies its validity by (i) imposing restrictive assumptions on the complexities of b and p (in terms of smoothness or sparsity) and then (ii) appealing to theorems that guarantee the asymptotic validity of the Wald CI under these assumptions. However, these assumptions may be incorrect. We show that we can often construct AFECTs that can falsify the complexity reducing assumptions on b and p.

To make the above more concrete, we describe our approach at a high level. Throughout, we let A denote the treatment indicator, Y a bounded outcome of interest, and X the vector of potential confounders with compact support. Let $\hat{\psi}_1$ and $\hat{\psi}_1 \pm z_{\alpha/2}\widehat{\mathsf{s.e.}}(\hat{\psi}_1)$ denote a DRML estimator of and associated $(1-\alpha)$ Wald CI for a particular parameter ψ. In this paper, for didactic purposes only, we will choose ψ to be (components) of the so-called variance-weighted average treatment effect (ATE) of a binary treatment A on Y given a vector X of confounding variables. Specifically, these components are the expected conditional variance $\mathbb{E}[\mathsf{var}(A|X)]$ of A given X and the expected conditional covariance $\mathbb{E}[\mathsf{cov}(A,Y|X)]$ of A and Y given X, with the variance weighted ATE being $\mathbb{E}[\mathsf{cov}(A,Y|X)]/\mathbb{E}[\mathsf{var}(A|X)]$. We chose the variance weighted ATE precisely because $\mathbb{E}[\mathsf{var}(A|X)]$ is in the *monotone bias class* but $\mathbb{E}[\mathsf{cov}(A,Y|X)]$ is not, thereby allowing us to highlight the critical difference between these classes. The methods developed herein can be applied essentially unchanged to many other causal effect parameters (e.g., the average treatment effect and the effect of treatment on the treated) regardless of the state spaces of A and Y, as well as to many noncausal parameters.

Even for the parameter $\mathbb{E}[\mathsf{var}(A|X)]$, as explained in Remark 1.2, there is an unavoidable limitation to what can be achieved with our or any other method: No test, including ours, of the null hypothesis that the bias of a DRML estimator is negligible compared to its standard error can be consistent [without making additional, possibly incorrect, complexity reducing assumptions on $b(x)$ and $p(x)$]. Thus, when our $\alpha^\dagger$-level test rejects the null for $\alpha^\dagger$ small, we can have strong evidence that the estimators $\hat{\psi}_1$ and $\hat{\psi}_{\mathsf{cf},1}$ have bias at least the order of its standard error; nonetheless, when the test does not reject, we cannot conclude that there is good evidence that the bias is less than the standard error, no matter how large the sample size. In fact, in the absence of complexity reducing assumptions, no consistent estimator of $\mathbb{E}[\mathsf{var}(A|X)]$ exists; hence we can never empirically rule out that the bias of $\hat{\psi}_1$ and $\hat{\psi}_{\mathsf{cf},1}$ is as large as order 1, and thus $n^{1/2}$ times greater than $\widehat{\mathsf{s.e.}}(\hat{\psi}_1)$! Put another way, because we make essentially no assumptions, no methodology can (nontrivially) upper bound the bias of any estimator or lower bound the coverage of any confidence interval.

In this paper, we are adopting a skeptic's stance, which is illuminated by comparing two social norms. The first is the social norm most of our parents taught us and the second is the skeptic's social norm.

- Parental Social Norm: If You Don't Have Anything Positive to Contribute, Don't Go Criticizing Others.
- Skeptic's Social Norm: Not Having Anything Positive to Contribute Does Not Relieve You of Your Duty to Criticize What Others Say.

As we saw above, because we do not impose complexity reducing assumptions on b and p, we have nothing to contribute if we follow parental social norms. However, in this paper, we adopt the *skeptic's social norms* and criticize, where possible, an author who reports a state of the art $(1-\alpha)$ Wald CI $\hat{\psi}_{\mathsf{cf},1} \pm z_{\alpha/2}\widehat{\mathsf{s.e.}}(\hat{\psi}_{\mathsf{cf},1})$ as valid. However, our critique will have to be stronger than simply informing the author that one can prove (when possible) that his interval will not be valid if his complexity-reducing assumptions are incorrect, as he will likely respond that he believes his assumptions to be reasonable and likely true under the law actually generating the data. Instead, for parameters in the *monotone bias class*, we will employ AFECTs to prove to the author that his Wald CI is invalid.

For other parameters such as the $\mathbb{E}[\mathsf{cov}(A,Y|X)]$, we can only falsify the validity of the author's Wald interval under the so-called faithfulness assumption given in Section 4.1. Heuristically, faithfulness is the assumption that near perfect cancelling of the nonnegligible bias of two

separate components of the bias of $\hat{\psi}_{\mathsf{cf},1}$ (one estimable and the other not) to give near zero total bias will essentially never occur.

If we do not assume faithfulness, we must consider the less ambitious goal of demonstrating to the author, when possible, that his complexity reducing assumptions are incorrect [without being able to ever empirically prove that the bias of his estimator is of the order of its standard error or greater]. If successfully achieved, the author would then have to admit that he can no longer justify his earlier claim of validity for his state-of-the-art confidence interval. The approach described here is one of being "in dialogue with current practices and practitioners." This is not surprising, as it is the justifications of the practitioners that the skeptic is critiquing.

To be concrete, suppose, as is often the case, an author justifies the validity of $\hat{\psi}_1 \pm z_{\alpha/2}\widehat{\mathsf{s.e.}}(\hat{\psi}_1)$, and thus its cross-fit version $\hat{\psi}_{\mathsf{cf},1} \pm z_{\alpha/2}\widehat{\mathsf{s.e.}}(\hat{\psi}_{\mathsf{cf},1})$ by (i) first proving that, under his complexity reducing assumptions, the Cauchy–Schwarz (CS) bias functional

$$\mathsf{CSBias}(\hat{\psi}_1) = \{\mathbb{E}[(\hat{b}(X) - b(X))^2]\}^{1/2} \times \{\mathbb{E}[(\hat{p}(X) - p(X))^2]\}^{1/2}$$

is $o(n^{-1/2})$,[3] conditional on the training sample[4] (and thus also on the functions $\hat{b}$, $\hat{p}$ computed from the training sample) and (ii) then noting the CS bias upper bounds the absolute conditional bias

$$|\mathbb{E}[(\hat{b}(X) - b(X))(\hat{p}(X) - p(X))]|$$

of $\hat{\psi}_1$ for $\psi(\theta) = \mathbb{E}_\theta[\mathsf{cov}_\theta(A, Y|X)]$. It then follows if we can empirically show that $\mathsf{CSBias}(\hat{\psi}_1)$ exceeds some given multiple $\delta > 0$, for example, $\delta = 0.75$, times $\hat{\psi}_1$'s conditional standard error of order $n^{-1/2}$, then we have falsified the analysts' justification of the claim that his nominal $(1-\alpha)$ Wald CIs are valid.

To this end, we shall construct $\alpha^\dagger$-level AFECTs of the null hypothesis $\mathsf{CSBias}(\hat{\psi}_1) < \mathsf{s.e.}(\hat{\psi}_1)\delta$, which can be done because, as we shall see, the $\mathsf{CSBias}(\hat{\psi}_1)$ parameter is in the *monotone bias class*.

We now describe our AFECT tests and related estimators at a high level. DRML estimators are based on the first-order influence function of the parameter ψ (van der Vaart, 1998). Our proposed approach begins by computing a *second-order influence function estimator* $\widehat{\mathbb{IF}}_{22,k}$ of the estimable part of the conditional bias $\mathbb{E}[\hat{\psi}_1 - \psi]$ of $\hat{\psi}_1$ *given the training sample data*. The bias corrected estimator is $\hat{\psi}_{2,k} \equiv \hat{\psi}_1 - \widehat{\mathbb{IF}}_{22,k}$, where $\widehat{\mathbb{IF}}_{22,k}$ is a second-order U-statistic that depends on a choice of k (with $k = o(n^2)$ for reasons explained in Remark 2.9), a vector of basis functions $\bar{\mathsf{Z}}_k \equiv \bar{\mathsf{z}}_k(X) \equiv (\mathsf{z}_1(X), \ldots, \mathsf{z}_k(X))^\top$ of X and an estimator $\widehat{\Omega}_k^{-1}$ of the inverse expected outer product $\Omega_k^{-1} := \{\mathbb{E}[\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top]\}^{-1}$. Both $\hat{\psi}_{2,k}$ and $\widehat{\mathbb{IF}}_{22,k}$ will be asymptotically normal when, as in our asymptotic setup, $k \to \infty$ and $k = o(n^2)$ as $n \to \infty$. (If k did not increase with n, the asymptotic distribution of $\widehat{\mathbb{IF}}_{22,k}$ would be the so-called Gaussian chaos distribution (Rubin and Vitale, 1980).)

The degree of the bias corrected by $\widehat{\mathbb{IF}}_{22,k}$ depends critically on (i) the choice of k, (ii) the accuracy of the estimator $\widehat{\Omega}_k^{-1}$ of Ω_k^{-1} when Ω_k^{-1} is unknown (see Section S3), and (iii) the particular k-vector of (basis) functions $\bar{\mathsf{Z}}_k \equiv \bar{\mathsf{z}}_k(X)$ selected from a much larger, possibly countably infinite, dictionary of candidate functions.

One sometimes has X-semisupervised data available; that is, a data set in which the number N of subjects with complete data on (A, Y, X) is many fold less than the number of subjects on whom only data on the covariates X are available. In that case, assuming the subjects with complete data are effectively a random sample of all subjects, we can estimate Ω_k by the empirical covariance matrix from subjects with incomplete data; and then treat Ω_k^{-1} as known in an analysis based on the N subjects with complete data (Chapelle, Schölkopf and Zien, 2010, Chakrabortty and Cai, 2018). Since, for the most of the paper we assume access to semisupervised data, we will omit the notational dependence on Ω_k^{-1} and denote $\widehat{\mathbb{IF}}_{22,k}(\Omega_k^{-1})$ and $\hat{\psi}_{2,k}(\Omega_k^{-1})$ by $\widehat{\mathbb{IF}}_{22,k}$ and $\hat{\psi}_{2,k}$. However, we write $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Omega}_k^{-1})$ and $\hat{\psi}_{2,k}(\widehat{\Omega}_k^{-1})$ when an estimator $\widehat{\Omega}_k^{-1}$ is substituted for Ω_k^{-1}. In the simulations below, we use a particular data-adaptive estimator $\widehat{\Omega}_k^{-1}$, described in the Appendix. Both $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Omega}_k^{-1})$ and $\hat{\psi}_{2,k}(\widehat{\Omega}_k^{-1})$ performed very well in our simulations; nonetheless, in contrast to $\widehat{\mathbb{IF}}_{22,k}$ and $\hat{\psi}_{2,k}$, we, as yet, lack a theoretical understanding of their statistical behavior. Consequently, we have relegated the definition and discussion of the estimators $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Omega}_k^{-1})$ and $\hat{\psi}_{2,k}(\widehat{\Omega}_k^{-1})$ to the Appendix and the Supplementary Material (Liu, Mukherjee and Robins, 2020), as requested by a referee.

For further motivation, we now summarize the results from one of the simulation studies that are described in detail in Section S9. We simulated 100 estimation samples each with sample size $n = 5000$. The same training sample, also of size 5000, and thus the same estimates of the nuisance regression functions were used in each simulation. Thus the results are conditional on that training sample. The dimension d of X is chosen to be 2 in order to allow estimation of the nuisance functions by kernel regression (with bandwidth selected by cross validation) in a timely fashion. We let $\psi = \mathbb{E}[\mathsf{var}(A|X)]$. We took k to be less than n for the following three reasons: $k < n$ is necessary (i) for CIs centered at $\hat{\psi}_{2,k} \equiv \hat{\psi}_1 - \widehat{\mathbb{IF}}_{22,k}$ to

[3]Here, the asymptotic statement would be in probability had we not treated the training sample as fixed.

[4]In this paper, essentially all expectations and probabilities are to be understood as conditioning on the training sample. Hence we can and do omit this conditioning event in our notation.

have length approximately equal to CIs centered at $\hat{\psi}_1$, (ii) for $\mathsf{s.e.}(\widehat{\mathbb{IF}}_{22,k})$ to be of order smaller than or equal to the order $n^{-1/2}$ of the standard error of $\hat{\psi}_1$, thereby creating the possibility of detecting that the ratio of the bias of $\hat{\psi}_1$ to its standard error exceeds a constant δ, if n is sufficiently large and (iii) to be able to estimate Ω_k^{-1} accurately without imposing the additional (possibly incorrect) smoothness or sparsity assumptions on the marginal density f_X. Thus we were able to use quite nonsmooth densities f_X in simulations; see Section S9.

In simulation studies, we chose a data generating process for which the minimax rates of estimation were known, in order to be able to better evaluate the properties of our proposed procedures. Specifically, both the true propensity score and outcome regression functions in our simulation studies were chosen to lie in particular Hölder smoothness classes chosen to ensure that $\hat{\psi}_1$ had significant asymptotic bias. We estimated these regression functions using nonparametric kernel regression estimators that are known to obtain the minimax optimal rate of convergence for these smoothness classes (Tsybakov, 2009), thereby guaranteeing that $\hat{\psi}_1$ performed close to as well as any other DRML estimator. [Out of interest, in Section S9, we also report simulation results when the regression functions are estimated by neural networks.] The basis functions $\bar{\mathsf{z}}_k(x)$ were chosen to be particular Cohen–Vial–Daubechies wavelets that Robins et al. (2009, 2017) showed to be minimax optimal for estimation of ψ by $\hat{\psi}_{2,k}$ for the chosen smoothness classes. In summary, we used optimal versions of $\hat{\psi}_1$ and $\hat{\psi}_{2,k}$ to ensure a fair comparison.

Table 1 reports results from this simulation study. We examined the empirical behavior of our data adaptive estimator as k varies by comparing the estimators $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Omega}_k^{-1})$ and $\hat{\psi}_{2,k}(\widehat{\Omega}_k^{-1})$ that use $\widehat{\Omega}_k^{-1}$ to the oracle estimators $\widehat{\mathbb{IF}}_{22,k}$ and $\hat{\psi}_{2,k}$ that use the true inverse covariance matrix Ω_k^{-1} (see the Appendix and Section S3). The target parameter ψ of this simulation study is the expected conditional variance of A given X. Simulation results for the expected conditional covariance were similar and are reported in Section S9.

Note the unmodified estimator $\hat{\psi}_1$ is included as the first row of Table 1 as, by definition, it equals $\hat{\psi}_{2,k}$ for $k = 0$. Also by definition, $\widehat{\mathbb{IF}}_{22,k=0}$ and $\widehat{\mathbb{IF}}_{22,k=0}(\widehat{\Omega}_k^{-1})$ are zero. As seen in row 1, column 2 of Table 1, nominal 90% Wald CIs centered at $\hat{\psi}_1 = \hat{\psi}_{2,k=0}$ had empirical coverage of 0% in 100 simulations. However, as seen in column 2 of both the upper and lower panels of the last row, 90% Wald

TABLE 1
A simulation result

k	$\widehat{\mathbb{IF}}_{22,k}$	MC coverage ($\hat{\psi}_{2,k}$ 90% Wald CI)	$\mathsf{Bias}(\hat{\psi}_{2,k})$	$\widehat{\chi}_k(\Omega_k^{-1}; z_{0.10}, \delta = 0.75(1.5))$
0	0 (0)	0%	0.229 (0.0161)	0% (0%)
64	0.0457 (0.00782)	0%	0.183 (0.0144)	99% (44%)
128	0.0484 (0.00831)	0%	0.180 (0.0145)	100% (54%)
256	0.125 (0.0144)	0%	0.103 (0.0114)	100% (100%)
512	0.127 (0.0147)	0%	0.101 (0.0122)	100% (100%)
1024	0.129 (0.0172)	0%	0.100 (0.0147)	100% (100%)
2048	0.161 (0.0238)	4%	0.0672 (0.0191)	100% (100%)
4096	0.180 (0.0271)	46%	0.0483 (0.0259)	100% (100%)
k	$\widehat{\mathbb{IF}}_{22,k}(\widehat{\Omega}_k^{-1})$	MC coverage ($\hat{\psi}_{2,k}(\widehat{\Omega}_k^{-1})$ 90% Wald CI)	$\mathsf{Bias}(\hat{\psi}_{2,k}(\widehat{\Omega}_k^{-1}))$	$\widehat{\chi}_k(\widehat{\Omega}_k^{-1}; z_{0.10}, \delta = 0.75(1.5))$
0	0 (0)	0%	0.229 (0.0252)	0% (0%)
64	0.0465 (0.00785)	0%	0.182 (0.0143)	100% (47%)
128	0.0498 (0.00831)	0%	0.180 (0.0143)	100% (64%)
256	0.131 (0.0142)	0%	0.0972 (0.0116)	100% (100%)
512	0.136 (0.0150)	0%	0.0922 (0.0125)	100% (100%)
1024	0.142 (0.0173)	0%	0.0868 (0.0143)	100% (100%)
2048	0.165 (0.0222)	4%	0.0636 (0.0185)	100% (100%)
4096	0.225 (0.0374)	90%	0.00314 (0.0296)	100% (100%)

Here, the parameter of interest is $\psi(\theta) = \mathbb{E}_\theta[\mathsf{var}_\theta(A|X)]$. We reported the Monte Carlo averages (MCavs) of point estimates and Monte Carlo standard deviations (MCsds) in the parenthesis (first column in each panel) of $\widehat{\mathbb{IF}}_{22,k}$ and $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Omega}_k^{-1})$, together with the coverage probability of 90% CIs (second column in each panel) of $\hat{\psi}_{2,k}$ and $\hat{\psi}_{2,k}(\widehat{\Omega}_k^{-1})$, the MCavs of the bias and MCsds in the parenthesis (third column in each panel) of $\hat{\psi}_{2,k}$ and $\hat{\psi}_{2,k}(\widehat{\Omega}_k^{-1})$ and the empirical rejection rate based on the test statistic $\widehat{\chi}_k(\zeta_k, \delta = 0.75 \text{ or } 1.5)$ and $\widehat{\chi}_k(\widehat{\Omega}_k^{-1}; \zeta_k, \delta = 0.75 \text{ or } 1.5)$ (see Section 2) with $\zeta_k = z_{0.10} = 1.28$ (fourth column in each panel). In the simulation, we choose $A \sim p(X) + N(0, 1)$. For more details on the simulation setup, see Section S9.

CIs centered at $\hat{\psi}_{2,k}$ at $k = 4096$ had empirical coverage around 46%.[5] The standard error of $\hat{\psi}_{2,k}$ did not greatly exceed that of $\hat{\psi}_1$.

In more detail, the left panel of Table 1 displays the Monte Carlo averages (MCavs) of the point estimates and Monte Carlo standard deviations (MCsds) (in parentheses) of $\widehat{\mathbb{IF}}_{22,k}$ in the first column; the empirical probability that a nominal 90% Wald CI centered at $\hat{\psi}_{2,k}$ covered the true parameter value in the second column; the MC bias (i.e., MCav of $\hat{\psi}_{2,k} - \psi$) and MCsd of $\hat{\psi}_{2,k}$ in the third column; and, in the fourth column, the empirical rejection rate of a one-sided $\alpha^\dagger = 0.10$ level test $\widehat{\chi}_k^{(1)}(z_{\alpha^\dagger=0.10}, \delta = 0.75 \text{ or } 1.5)$ (defined in equation (3.2) of Section 2) of the null hypothesis that the bias of $\hat{\psi}_1$ is smaller than $\delta = 0.75$ or 1.5 of its standard error. The test rejects when the ratio $\widehat{\mathbb{IF}}_{22,k}/\widehat{\mathsf{s.e.}}(\hat{\psi}_1)$ is large. Similarly, the bottom panel displays these same summary statistics but with the data adaptive estimator $\widehat{\Omega}_k^{-1}$ in place of Ω_k^{-1}. The difference between the MC bias of $\hat{\psi}_{2,k}(\widehat{\Omega}_k^{-1})$ and $\hat{\psi}_{2,k}$ is an estimate of the additional bias due to the estimation of Ω_k^{-1} by $\widehat{\Omega}_k^{-1}$. (The uncertainty in the estimate of the bias itself is not given in the table but it is negligible as it approximately equals $(1/100)^{1/2}$ times the standard error given in the table.)

Reading from the first row of Table 1, we see that the MC bias of $\hat{\psi}_1$ was 0.229. The MC bias of $\hat{\psi}_{2,k}$ and $\hat{\psi}_{2,k}(\widehat{\Omega}_k^{-1})$ decreased with increasing k, becoming nearly zero at $k = 4096$. The observation that the bias decreases as k increases is predicted by the theory developed in Section 2 and reflects the fact that $\psi = \mathbb{E}[\mathsf{var}(A|X)]$ is in the *monotone bias class*. The decrease in bias reflects the increase in the absolute value of $\widehat{\mathbb{IF}}_{22,k}$ with k. Note further that both the MCavs of $\widehat{\mathbb{IF}}_{22,k}$ and $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Omega}_k^{-1})$ are relatively close, as are their MCsds, implying that our estimator $\widehat{\Omega}_k^{-1}$ performs similar to the true Ω_k^{-1}. The actual coverages of 90% Wald CIs centered at $\hat{\psi}_{2,k}$ and $\hat{\psi}_{2,k}(\widehat{\Omega}_k^{-1})$ both increase from 0% at $k = 0$ to more than 40% at $k = 4096$. Also, reading from the third column, we see that the MCsd (0.0259) of $\hat{\psi}_{2,k=4096}$ is only 1.6 times the standard error (0.0161) of $\hat{\psi}_1$, confirming that the dramatic difference in coverage rates of their associated CIs is due to the bias of $\hat{\psi}_1$. Reading from the 4th column of each panel, we see that the rejection rates of both $\widehat{\chi}_k^{(1)}(z_{\alpha^\dagger=0.10}, \delta)$ and $\widehat{\chi}_k^{(1)}(\widehat{\Omega}_k^{-1}; z_{\alpha^\dagger=0.10}, \delta)$ for $\delta = 0.75$ (for $\delta = 1.5$) are already 100% when k is 64 (256), indicating that the bias of $\hat{\psi}_1$ is much greater than 0.75 (1.5) of its standard error. Indeed, reading from row 1 of column 3, we see that the ratio of the MC bias of $\hat{\psi}_1 = \hat{\psi}_{2,k=0}$ (0.229) to its MCsd (0.0161) is nearly 14. In Remark 2.4, we show that this ratio is close to that predicted by theory.

Figure S1 in Section S10.1 provides a histogram over the 100 estimation samples of $(1 - \alpha^\dagger)$ upper confidence bounds $\mathsf{UCB}^{(1)}(\Omega_{k=2048}^{-1}; \alpha, \alpha^\dagger)$ (defined in equation (3.4) of Section 2) and $\mathsf{UCB}^{(1)}(\widehat{\Omega}_{k=2048}^{-1}; \alpha, \alpha^\dagger)$ (defined in equation (A.2) of the Appendix) for the actual conditional asymptotic coverage of the nominal $(1 - \alpha)$ CI $\hat{\psi}_1 \pm z_{\alpha/2}\widehat{\mathsf{s.e.}}(\hat{\psi}_1)$. To clarify the meaning of $\mathsf{UCB}^{(1)}(\Omega_{k=2048}^{-1}; \alpha, \alpha^\dagger)$, let $\mathsf{coverage}(\alpha) = P(\psi \in \{\hat{\psi}_1 \pm z_{\alpha/2}\widehat{\mathsf{s.e.}}(\hat{\psi}_1)\})$ be the conditional actual coverage of ψ, given the training sample. Then, by definition, a $(1 - \alpha^\dagger)$ conditional upper confidence bound $\mathsf{UCB}^{(1)}(\Omega_{k=2048}^{-1}; \alpha, \alpha^\dagger)$ is a random variable satisfying[6]

$$P\{\mathsf{coverage}(\alpha) \le \mathsf{UCB}^{(1)}(\Omega_{k=2048}^{-1}; \alpha, \alpha^\dagger)\} \ge 1 - \alpha^\dagger. \tag{1.1}$$

Recall from row 1, column 2 of the right panel of Table 1, that the actual Monte Carlo coverage of the nominal 90% interval $\hat{\psi}_1 \pm 1.64\widehat{\mathsf{s.e.}}(\hat{\psi}_1)$ was 0%. As expected, our nominal 90% upper confidence bounds $\mathsf{UCB}^{(1)}(\Omega_{k=2048}^{-1}; \alpha, \alpha^\dagger)$ and $\mathsf{UCB}^{(1)}(\widehat{\Omega}_{k=2048}^{-1}; \alpha, \alpha^\dagger)$ were nearly 0% in all the 100 simulated estimation samples.

Organization of the paper. The remainder of the paper is organized as follows. In Section 1.1 to Section 1.3, we describe our data structure, our parameters of interest ψ, the state of the art DRML estimators and the statistical properties of these estimators.

In Section 2, we present a second-order U-statistic $\widehat{\mathbb{IF}}_{22,k}$ that is an unbiased estimator of the "estimable" part of the bias of $\hat{\psi}_1$.

In Section 3 and Section 4, we develop $\alpha^\dagger$ level tests that have the ability to detect whether the bias of $\hat{\psi}_1$ is of the same or greater order than its standard error, for the expected conditional variance; in the case of the expected conditional covariance we test whether the Cauchy–Schwarz (CS) bias is the same or greater than the standard error of $\hat{\psi}_1$.

In the Appendix and the Supplementary Material (Liu, Mukherjee and Robins, 2020), we propose an estimator $\widehat{\Omega}_k^{-1}$ of Ω_k^{-1} which performs well in simulations but lacks theoretical guarantees.

In Section 5, we consider a semisupervised setting with $k > n$, based on the following motivation. The estimator $\hat{\psi}_{2,k} = \hat{\psi}_1 - \widehat{\mathbb{IF}}_{22,k}$ of $\psi = \mathbb{E}[\mathsf{var}(A|X)]$ with k less

[5] Our data generating process implied that $\hat{\psi}_{2,k}$ was $\sqrt{n}$-consistent but asymptotically biased, so the expected coverage of the Wald CI centered at $\hat{\psi}_{2,k}$ was less than 90%.

[6] For example, if $\mathsf{UCB}^{(1)}(\Omega_{k=2048}^{-1}; \alpha = 0.10, \alpha^\dagger = 0.10) = 0.14$, then the actual coverage of the nominal 90% interval $\hat{\psi}_1 \pm 1.64\widehat{\mathsf{s.e.}}(\hat{\psi}_1)$ is no more than 14% with confidence at least $1 - \alpha^\dagger = 0.90$. More precisely, the random interval $[0, \mathsf{UCB}^{(1)}(\Omega_{k=2048}^{-1}; \alpha = 0.10, \alpha^\dagger = 0.10)]$ is guaranteed to include the actual coverage of $\hat{\psi}_1 \pm 1.64\widehat{\mathsf{s.e.}}(\hat{\psi}_1)$ at least 90% of the time in repeated sampling of the estimation sample with the training sample fixed.

than but near n has standard error not much larger than the standard error of $\hat{\psi}_1$, but has smaller bias. This suggests foregoing the estimation of an upper bound on the actual coverage of a nominal $(1-\alpha)$ Wald CI centered at $\widehat{\psi}_1$; rather always report, with Ω_k^{-1} known, the nominal $(1-\alpha)$ Wald CI $\hat{\psi}_{2,k} \pm z_{\alpha/2}\widehat{\mathsf{s.e.}}(\hat{\psi}_{2,k})$ with k just less than n. However, doing so naturally raises the question of whether the interval $\hat{\psi}_{2,k} \pm z_{\alpha/2}\widehat{\mathsf{s.e.}}(\hat{\psi}_{2,k})$ itself covers ψ at its nominal $1-\alpha$ rate. In Section 5, we develop a test of the null hypothesis that the ratio of the conditional bias of $\hat{\psi}_{2,k}$ to its standard error is smaller than a fraction δ using an AFECT statistic based on $\widehat{\mathbb{IF}}_{22,k'}$ with $k' > n$.

In Section 6, we conclude by discussing several open problems.

The following common asymptotic notation are used throughout the paper: $x \lesssim y$ (equivalently $x = O(y)$) denotes that there exists some constant $C > 0$ such that $x \leq Cy$, $x \asymp y$ means there exist some constants $c_1 > c_2 > 0$ such that $c_2|y| \leq |x| \leq c_1|y|$. $x = o(y)$ or $y \gg x$ is equivalent to $\lim_{x,y\to\infty} \frac{x}{y} = 0$. For a random variable X_n with law P possibly depending on the sample size n, $X_n = O_P(a_n)$ denotes that X_n/a_n is bounded in P-probability, and $X_n = o_P(a_n)$ means that $\lim_{n\to\infty} P(|X_n/a_n| \geq \epsilon) = 0$ for every positive real number ϵ.

1.1 Parameter of Interest

In this part, we begin to make precise the issues discussed above. For didactic purposes, we will restrict our discussion to the variance-weighted average treatment effect (variance weighted ATE, defined below) for a binary treatment A and binary outcome Y given a vector X of d-dimensional baseline covariates compactly supported in $[0,1]^d$. We suppose we observe N i.i.d. copies from the joint distribution of (Y, A, X).

We parametrize the joint distribution P_θ of (Y, A, X) by the variation independent parameters $\theta \equiv (b, p, f_X, OR_{YA|X=x})$, where

$$b(X) \equiv \mathbb{E}_\theta[Y|X],$$
$$p(X) \equiv \mathbb{E}_\theta[A|X]$$

are respectively the regression of Y on X and the regression of A on X, f_X is the marginal density of X, and $OR_{YA|X=x}$ is the conditional odds ratio. We let $\hat{\theta} = (\hat{b}, \hat{p}, \theta \setminus \{b, p\})$. Throughout the paper, we use $\mathbb{E}_\theta$, var_θ and cov_θ with subscript θ to indicate the conditional expectation, variance and covariance, given the training sample, under the probability measure P_θ indexed by θ. We assume a nonparametric infinite dimensional model $\mathcal{M}(\Theta) := \{P_\theta, \theta \in \Theta\}$ where Θ indexes all possible θ subject to weak regularity conditions given later in Condition W.

Under the assumption that the vector X of measured covariates suffices to control confounding, the variance weighted ATE $\tau(\theta)$ is identified as $\tau(\theta) := \frac{\mathbb{E}_\theta[\gamma_\theta(X)\mathsf{var}_\theta(A|X)]}{\mathbb{E}_\theta[\mathsf{var}_\theta(A|X)]}$ where $\gamma_\theta(X) \equiv \mathbb{E}_\theta[Y|A=1,X] - \mathbb{E}_\theta[Y|A=0,X]$ is the conditional treatment effect given X and $\mathsf{var}_\theta(A|X) = p(X)(1-p(X))$. In applications, the variance weighted ATE arises when we want to downweight the subjects whose propensity scores are extreme. Moreover, the parameter $\tau(\theta)$ can also be identified as the regression coefficient of A[7] in the classical semiparametric partially linear model $Y = \tau A + b(X) + \mathsf{noise}$.

Some algebra shows that

$$\tau(\theta) = \frac{\mathbb{E}_\theta[\mathsf{cov}_\theta(Y, A|X)]}{\mathbb{E}_\theta[\mathsf{var}_\theta(A|X)]}.$$

Henceforth, we shall restrict attention to the estimation of the expected conditional covariance

$$\psi(\theta) \equiv \mathbb{E}_\theta\big[\mathsf{cov}_\theta(Y, A|X)\big] = \mathbb{E}_\theta\big[\{Y - b(X)\}\{A - p(X)\}\big].$$

and the expected conditional variance $\mathbb{E}_\theta[\mathsf{var}_\theta(A|X)]$, which is simply the special case of $\mathbb{E}_\theta[\mathsf{cov}_\theta(Y, A|X)]$ in which $A = Y$ w.p.1. If we can construct asymptotically unbiased and normal estimators of $\mathbb{E}_\theta[\mathsf{cov}_\theta(Y, A|X)]$ and $\mathbb{E}_\theta[\mathsf{var}_\theta(A|X)]$, we also can construct the same for $\tau(\theta)$ by the functional delta method.

REMARK 1.1. We shall see that the statistical guarantees of our bias correction methodology differ depending on whether the parameter of interest is $\mathbb{E}_\theta[\mathsf{cov}_\theta(Y, A|X)]$ versus $\mathbb{E}_\theta[\mathsf{var}_\theta(A|X)]$. In fact, the insight into our methodology offered by this difference is the reason we chose the variance weighted average treatment effect rather than the average treatment effect as the causal effect of interest in this paper.

In the next section, we describe the current state-of-the-art DRML estimators $\hat{\psi}_1$ and $\hat{\psi}_{\mathsf{cf},1}$. They will depend on estimators $\hat{b}(x)$ and $\hat{p}(x)$ of $b(x)$ and $p(x)$, which may have been outputted by machine learning algorithms for estimating conditional means, with completely unknown statistical properties.

REMARK 1.2. The methods in Robins et al. (2009) and Ritov et al. (2014) can be straightforwardly combined to show that, without further unverifiable assumptions (such as smoothness or sparsity assumptions that may be incorrect), for some $\sigma > 0$, no consistent α-level test of the null hypothesis $\mathbb{E}_\theta[\mathsf{cov}_\theta(A, Y|X)] = \sigma$ for $\sigma > 0$ versus the alternative $\mathbb{E}_\theta[\mathsf{cov}_\theta(A, Y|X)] = \sigma + c$ for some fixed constant $c > 0$ exists, whenever some components of X have a continuous distribution. Furthermore, there is no consistent estimator of the expected conditional covariance without further unverifiable assumptions. The above negative result also applies to the expected conditional variance $\mathbb{E}_\theta[\mathsf{var}_\theta(A|X)]$.

[7] A does not need to be binary.

1.2 State-of-the-Art Estimators $\hat{\psi}_1$ and $\hat{\psi}_{\mathsf{cf},1}$ and Their Asymptotic Properties

The state-of-the-art DRML estimator $\hat{\psi}_1$ uses sample splitting, because $\hat{b}(x)$ and $\hat{p}(x)$ have unknown statistical properties and, in particular, may not lie in a so-called Donsker class (see, e.g., van der Vaart and Wellner, 1996, Chapter 2)—a condition often needed for valid inference when we do not split the sample. The cross-fitting estimator $\hat{\psi}_{\mathsf{cf},1}$ is a DRML estimator that can recover the information lost by $\hat{\psi}_1$ due to sample splitting, provided that $\hat{\psi}_1$ is asymptotically unbiased given the training sample.

The following algorithm defines $\hat{\psi}_1$ and $\hat{\psi}_{\mathsf{cf},1}$ for $\psi(\theta) = \mathbb{E}_\theta[\mathsf{cov}_\theta(A, Y|X)]$ and can be easily modified for $\psi(\theta) = \mathbb{E}_\theta[\mathsf{var}_\theta(A|X)]$:

(i) The N study subjects are randomly split into 2 parts: an estimation sample of size n and a training (nuisance) sample of size $n_{\mathsf{tr}} = N - n$ with $n/N \approx 1/2$. Without loss of generality we shall assume that $i = 1, \ldots, n$ corresponds to the estimation sample.

(ii) Estimators $\hat{b}(x)$, $\hat{p}(x)$ are constructed from the training sample data using ML methods.

(iii) Compute

$$\hat{\psi}_1 = \frac{1}{n}\sum_{i=1}^{n}[\{Y_i - \hat{b}(X_i)\}\{A_i - \hat{p}(X_i)\}]$$

from n subjects in the estimation sample and

$$\hat{\psi}_{\mathsf{cf},1} = (\hat{\psi}_1 + \overline{\hat{\psi}}_1)/2,$$

where $\overline{\hat{\psi}}_1$ is $\hat{\psi}_1$ but with the training and estimation samples reversed.

1.3 Asymptotic Properties of $\hat{\psi}_1$ and $\hat{\psi}_{\mathsf{cf},1}$

The following theorems (Theorem 1.3 and Theorem 1.4) give the asymptotic properties of the estimator $\hat{\psi}_1$ of the expected conditional covariance, conditional on the training sample.

THEOREM 1.3. *Conditional on the training sample, $\hat{\psi}_1$ is asymptotically normal with conditional bias*

$$\begin{aligned}\mathsf{Bias}_\theta(\hat{\psi}_1) &:= \mathbb{E}_\theta\big[\hat{\psi}_1 - \psi(\theta)\big] \\ &= \mathbb{E}_\theta\big[\{b(X) - \hat{b}(X)\}\{p(X) - \hat{p}(X)\}\big].\end{aligned} \tag{1.2}$$

PROOF. Since conditionally $\hat{b}(x)$ and $\hat{p}(x)$ are fixed functions, $\hat{\psi}_1$ is the sum of i.i.d. bounded random variables and thus is asymptotically normal. A straightforward calculation shows $\mathsf{Bias}_\theta(\hat{\psi}_1)$ is the conditional bias. □

We note that $\hat{\psi}_1$ is, by definition, doubly robust because $\mathsf{Bias}_\theta(\hat{\psi}_1) = 0$ if either $b(X) = \hat{b}(X)$ or $p(X) = \hat{p}(X)$ with P_θ-probability 1. Finally, before proceeding, we summarize the statistical properties of the DRML estimator in the following theorem, the proof of which is standard and can be found in Chernozhukov et al. (2018). Recall that $\mathsf{Bias}_\theta(\hat{\psi}_1)$ is random only through its dependence on the training sample data via $\hat{b}$ and $\hat{p}$.

THEOREM 1.4. *If* (a) $\mathsf{Bias}_\theta(\hat{\psi}_1)$ *is* $o(n^{-1/2})$ *and* (b) $\hat{b}(x)$ *and* $\hat{p}(x)$ *converge to* $b(x)$ *and* $p(x)$ *in* $L_2(P_\theta)$, *then*:

1.

$$\hat{\psi}_1 - \psi(\theta) = n^{-1}\sum_{i=1}^{n}\mathsf{IF}_{1,i}(\theta) + o\big(n^{-1/2}\big),$$

$$\hat{\psi}_{\mathsf{cf},1} - \psi(\theta) = N^{-1}\sum_{i=1}^{N}\mathsf{IF}_{1,i}(\theta) + o\big(N^{-1/2}\big),$$

where $\mathsf{IF}_1(\theta) = \{Y - b(X)\}\{A - p(X)\} - \psi(\theta)$ *is the first-order influence function of* $\psi(\theta)$ *under* P_θ. *Further,* $n^{1/2}(\hat{\psi}_1 - \psi(\theta))$ *converges conditionally and unconditionally to a normal distribution with mean zero*; $\hat{\psi}_{\mathsf{cf},1}$ *is a regular, asymptotically linear estimator*; *that is,* $N^{1/2}(\hat{\psi}_{\mathsf{cf},1} - \psi(\theta))$ *converges unconditionally to a normal distribution with mean zero and variance equal to the semiparametric variance bound* $\mathsf{var}_\theta[\mathsf{IF}_1(\theta)]$.

2. *The* $(1-\alpha)$ *nominal Wald CIs* (*CIs*)

$$\hat{\psi}_1 \pm z_{\alpha/2}\widehat{\mathsf{s.e.}}[\hat{\psi}_1],$$

$$\hat{\psi}_{\mathsf{cf},1} \pm z_{\alpha/2}\widehat{\mathsf{s.e.}}[\hat{\psi}_{\mathsf{cf},1}]$$

are $(1-\alpha)$ *asymptotic CI for* $\psi(\theta)$. *Here,* $\widehat{\mathsf{s.e.}}[\hat{\psi}_1] = (\widehat{\mathsf{var}}[\hat{\psi}_1])^{1/2}$ *with*

$$\widehat{\mathsf{var}}[\hat{\psi}_1] = \frac{1}{n^2}\sum_{i=1}^{n}[\{Y_i - \hat{b}(X_i)\}\{A_i - \hat{p}(X_i)\}]^2,$$

$$\widehat{\mathsf{var}}[\hat{\psi}_{\mathsf{cf},1}] = \frac{1}{4}\{\widehat{\mathsf{var}}[\hat{\psi}_1] + \widehat{\mathsf{var}}[\overline{\hat{\psi}}_1]\}.$$

REMARK 1.5. Had we chosen $\psi(\theta) = \mathbb{E}_\theta[\mathbb{E}_\theta[Y|A = 1, X]]$, the mean response of Y under missing at random rather than the variance weighted ATE as our parameter of interest, the outcome regression function appearing in the first-order influence function would be $\mathbb{E}_\theta[Y|A = 1, X]$ rather than $\mathbb{E}_\theta[Y|X]$ and $\hat{\psi}_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{A_i}{\hat{p}(X_i)} \times (Y - \hat{b}(X_i)) + \hat{b}(X_i)$.

REMARK 1.6 (Training sample squared error loss cross-validation). How can we choose among the many (say, J) available ML algorithms if our goal is to minimize the conditional mean squared error $\mathbb{E}_\theta[(b(X) - \hat{b}(X))^2]$? One approach is to let the data decide by applying cross-validation restricted to the training sample. Specifically, we randomly split the training sample into S subsamples of size n_{tr}/S. For each subsample s, we fit the jth ML algorithm to the other $S-1$ subsamples to obtain outputs $\hat{b}_s^{(j)}(\cdot)$, for $j = 1, \ldots, J$. Next, we compute, for each j, the squared error loss $CV^{(j)} = \sum_{s=1}^{S} CV_s^{(j)}$ with

$CV_s^{(j)} = \sum_{i\in s}\{Y_i - \hat{b}_s^{(j)}(X_i)\}^2$, and finally select the ML algorithm $j_* = \arg\min_j CV^{(j)}$. Analogous results apply to the estimation of $p(X)$.

REMARK 1.7. Although a standard result, Theorem 1.4 is of minor interest to us in this paper for several reasons. First, because of their asymptotic nature, there is no finite sample size n at which any test could empirically falsify $\mathsf{Bias}_\theta(\hat{\psi}_1) = o(n^{-1/2})$. Rather, as discussed in Section 1, our interest, instead, lies in testing and rejecting hypotheses such as, at the actual estimation sample size n, the actual coverage of the interval $\hat{\psi}_1 \pm z_{\alpha/2}\widehat{\mathsf{s.e.}}[\hat{\psi}_1]$, conditional on the training sample, is less than a fraction $\varrho < 1$ of its nominal coverage.

Second, we make no assumptions concerning either the complexity of the unknown functions b and p or the statistical behavior of their ML estimators $\hat{b}$ and $\hat{p}$, our inferential statements will regard the training sample as fixed rather than random. In particular, the only randomness referred to in any theorem is that of the estimation sample. Our inferences rely on being in "asymptopia" to be able to posit that, at our estimation sample size of n, (1) the quantiles of the finite sample distribution of a conditionally asymptotically normal statistic (e.g., $\widehat{\mathbb{IF}}_{22,k}$ defined later in equation (2.8)) are close to the quantiles of a normal and (2) the standard error estimators of $\hat{\psi}_1$ and $\widehat{\mathbb{IF}}_{22,k}$ are close to their true standard errors. (It will often be useful to consider the power functions of our proposed tests as a function of the sample size, which we do by taking $n \to \infty$.)

REMARK 1.8. Indeed, when the constants in the nonasymptotic concentration inequalities (Boucheron, Lugosi and Massart, 2013, Vershynin, 2018) are explicit and can be estimated from data, then our reliance on asymptotics could be eliminated at the expense of decreased power and increased CI width. However, such finite sample bounds are beyond the scope of this paper.

Before starting to explain our methodology in detail, we collect some frequently used notation.

Notation. For a (random) vector V, $\|V\|_\theta \equiv \mathbb{E}_\theta[V^\top V]^{1/2}$ denotes its $L_2(P_\theta)$ norm conditioning on the training sample, $\|V\| \equiv (V^\top V)^{1/2}$ denotes its ℓ_2 norm and $\|V\|_\infty$ denotes its L_∞ norm. For any matrix A, $\|A\|$ will be used for its operator norm. Given a k, the random vector $\bar{\mathsf{Z}}_k = \bar{\mathsf{z}}_k(X) = (\mathsf{z}_1(X), \ldots, \mathsf{z}_k(X))^\top$, $\Pi[\cdot|\bar{\mathsf{Z}}_k]$ denotes the population linear projection operator onto the space spanned by $\bar{\mathsf{Z}}_k$ conditioning on the training sample: with $\Omega_k := \mathbb{E}_\theta[\bar{\mathsf{Z}}_k\bar{\mathsf{Z}}_k^\top]$, $\Pi[\cdot|\bar{\mathsf{Z}}_k^\perp] = I - \Pi[\cdot|\bar{\mathsf{Z}}_k]$ is the projection onto the orthogonal complement of $\bar{\mathsf{Z}}_k$ in the Hilbert space $L_2(f_X)$. Hence, for a random variable W,

$$\Pi[W|\bar{\mathsf{Z}}_k] = \bar{\mathsf{Z}}_k^\top \beta_{k,W}, \Pi\big[W|\bar{\mathsf{Z}}_k^\perp\big] = W - \Pi[W|\bar{\mathsf{Z}}_k], \tag{1.3}$$

where $\beta_{k,W} = \Omega_k^{-1}\mathbb{E}_\theta[\bar{\mathsf{Z}}_k W]$ is the vector of population regression coefficients. It should be noted that we allow selection of the vector $\bar{\mathsf{Z}}_k$ to depend on the training sample data (for further discussions, see Section 6). $\widehat{\Omega}_k^{-1}$ denotes a generic estimator of Ω_k^{-1}. When referring to a particular estimator of Ω_k^{-1} (mostly in the Appendix), an identifying superscript will often be attached.

We also denote the following commonly used residuals as

$$\widehat{\varepsilon}_{b,i} := Y_i - \hat{b}(X_i), \qquad \widehat{\varepsilon}_{p,i} := A_i - \hat{p}(X_i),$$
$$\widehat{\xi}_{b,i} := b(X_i) - \hat{b}(X_i), \qquad \widehat{\xi}_{p,i} := p(X_i) - \hat{p}(X_i)$$

for $i = 1, 2, \ldots, n$, where $\hat{b}$ and $\hat{p}$ are estimated from the training sample.

If $\bar{\mathsf{Z}}_{k_1}$ and $\bar{\mathsf{Z}}_{k_2}$ are vectors depending on different values of k, we impose the following restriction.

CONDITION B. For any $k_1 < k_2 = o(n^2)$, the space spanned by $\bar{\mathsf{Z}}_{k_1}$ is a subspace of the space spanned by $\bar{\mathsf{Z}}_{k_2}$.

REMARK 1.9. For example, when choosing the basis functions $\bar{\mathsf{Z}}_k$ from a dictionary $\mathcal{V}$ of (candidate) functions greedily, Condition B holds.

2. THE PROJECTED CONDITIONAL BIAS AND TWO DIFFERENCES BETWEEN $\mathbb{E}_\theta[\mathsf{var}_\theta(A|X)]$ AND $\mathbb{E}_\theta[\mathsf{cov}_\theta(A, Y|X)]$

In the main text, following the recommendation by a referee, we only discuss an "oracle" procedure that assumes Ω_k^{-1} to be known, as with semisupervised data.

Let $\mathcal{V}$ be a set (i.e., dictionary) of (basis) functions of X that is either countable or finite with cardinality $p > n$. Given the vector $X = (X_l; l = 1, \ldots, d)$ of d covariates, many choices for $\mathcal{V}$ are possible. For example, $\mathcal{V}$ could be tensor products of spline, wavelet or local polynomial partition series (or the union of all three types) in defining $\mathcal{V}$.

We decompose $b(X) - \hat{b}(X) = \Pi[b(X) - \hat{b}(X)|\bar{\mathsf{Z}}_k] + \Pi[b(X) - \hat{b}(X)|\bar{\mathsf{Z}}_k^\perp]$ (and similarly for $p(X) - \hat{p}(X)$), where the first term is the $L_2(P_\theta)$-orthogonal (population least squares) projection of $b(X) - \hat{b}(X)$ on the linear span of the vector $\bar{\mathsf{Z}}_k$ and the second term is the projection onto the orthocomplement $\bar{\mathsf{Z}}_k^\perp$. Specifically, following equation (1.3), we have

$$\begin{aligned}\Pi\big[b(X) - \hat{b}(X)|\bar{\mathsf{Z}}_k\big] &= \bar{\mathsf{Z}}_k^\top \beta_{k,b-\hat{b}} \\ &= \bar{\mathsf{Z}}_k^\top \Omega_k^{-1}\mathbb{E}_\theta\big[\bar{\mathsf{Z}}_k\big(b(X) - \hat{b}(X)\big)\big] \\ &= \bar{\mathsf{Z}}_k^\top \Omega_k^{-1}\mathbb{E}_\theta\big[\bar{\mathsf{Z}}_k\big(Y - \hat{b}(X)\big)\big] \\ &= \bar{\mathsf{Z}}_k^\top \Omega_k^{-1}\mathbb{E}_\theta[\bar{\mathsf{Z}}_k\hat{\varepsilon}_b],\end{aligned} \tag{2.1}$$

$$\begin{aligned}\Pi\big[p(X) - \hat{p}(X)|\bar{\mathsf{Z}}_k\big] &= \bar{\mathsf{Z}}_k^\top \beta_{k,p-\hat{p}} \\ &= \bar{\mathsf{Z}}_k^\top \Omega_k^{-1}\mathbb{E}_\theta\big[\bar{\mathsf{Z}}_k\big(p(X) - \hat{p}(X)\big)\big] \\ &= \bar{\mathsf{Z}}_k^\top \Omega_k^{-1}\mathbb{E}_\theta\big[\bar{\mathsf{Z}}_k\big(A - \hat{p}(X)\big)\big] \\ &= \bar{\mathsf{Z}}_k^\top \Omega_k^{-1}\mathbb{E}_\theta[\bar{\mathsf{Z}}_k\hat{\varepsilon}_p],\end{aligned} \tag{2.2}$$

where in the second lines of the above two equations we use the definitions of $b(X)$, $p(X)$, $\hat{\varepsilon}_b$ and $\hat{\varepsilon}_p$.

Then we have the following lemma that decomposes $\mathsf{Bias}_\theta(\hat{\psi}_1)$ (see the LHS of equation (1.2)).

LEMMA 2.1. *$\mathsf{Bias}_\theta(\hat{\psi}_1)$ can be decomposed into the sum of two terms $\mathsf{Bias}_{\theta,k}(\hat{\psi}_1)$ and $\mathsf{TB}_{\theta,k}(\hat{\psi}_1)$:*[8]

$$\mathsf{Bias}_\theta(\hat{\psi}_1) \equiv \mathsf{Bias}_{\theta,k}(\hat{\psi}_1) + \mathsf{TB}_{\theta,k}(\hat{\psi}_1), \tag{2.3}$$

where we define

$$\begin{aligned}\mathsf{Bias}_{\theta,k}(\hat{\psi}_1) &:= \mathbb{E}_\theta\big[\{\Pi\big[b(X) - \hat{b}(X)|\bar{\mathsf{Z}}_k\big]\} \\ &\quad \times \{\Pi\big[p(X) - \hat{p}(X)|\bar{\mathsf{Z}}_k\big]\}\big].\end{aligned}$$

Then

$$\begin{aligned}\mathsf{Bias}_{\theta,k}(\hat{\psi}_1) &= \beta_{k,b-\hat{b}}^\top \Omega_k \beta_{k,p-\hat{p}} \\ &\equiv \mathbb{E}_\theta[\hat{\varepsilon}_b \bar{\mathsf{Z}}_k]^\top \Omega_k^{-1} \mathbb{E}_\theta[\bar{\mathsf{Z}}_k \hat{\varepsilon}_p], \\ \mathsf{TB}_{\theta,k}(\hat{\psi}_1) &= \mathbb{E}_\theta\big[\{\Pi\big[b(X) - \hat{b}(X)|\bar{\mathsf{Z}}_k^\perp\big]\} \\ &\quad \times \{\Pi\big[p(X) - \hat{p}(X)|\bar{\mathsf{Z}}_k^\perp\big]\}\big].\end{aligned} \tag{2.4}$$

PROOF. By definition,

$$\begin{aligned}\mathsf{Bias}_{\theta,k}(\hat{\psi}_1) &:= \mathbb{E}_\theta\big[\{\Pi\big[b(X) - \hat{b}(X)|\bar{\mathsf{Z}}_k\big]\} \\ &\quad \times \{\Pi\big[p(X) - \hat{p}(X)|\bar{\mathsf{Z}}_k\big]\}\big] \\ &= \mathbb{E}_\theta\big[\beta_{k,b-\hat{b}}^\top \bar{\mathsf{Z}}_k \bar{\mathsf{Z}}_k^\top \beta_{k,p-\hat{p}}\big] \\ &= \beta_{k,b-\hat{b}}^\top \Omega_k \beta_{k,p-\hat{p}} \\ &= \mathbb{E}_\theta\big[\big(Y - \hat{b}(X)\big)\bar{\mathsf{Z}}_k\big]^\top \Omega_k^{-1} \\ &\quad \times \mathbb{E}_\theta\big[\bar{\mathsf{Z}}_k\big(A - \hat{p}(X)\big)\big],\end{aligned} \tag{2.5}$$

where the last equality follows from equation (2.1). The second part of equation (2.4) directly follows from the Pythagorean theorem. □

We now define the *monotone bias class* of parameters that we mentioned in Section 1.

DEFINITION 2.2 (Monotone bias class of parameters). For the parameter $\psi(\theta)$, given any DRML estimator $\hat{\psi}_1$, under Condition B, if $|\mathsf{TB}_{\theta,k}(\hat{\psi}_1)|$ is nonincreasing with k, or equivalently if $|\mathsf{Bias}_{\theta,k}(\hat{\psi}_1)|$ is nondecreasing with k, $\psi(\theta)$ is said to be in the *monotone bias class*.

2.1 Orderings Between $\mathsf{Bias}_\theta(\hat{\psi}_1)$ and $\mathsf{Bias}_{\theta,k}(\hat{\psi}_1)$: Difference Between $\mathbb{E}_\theta[\mathsf{cov}_\theta(Y, A|X)]$ and $\mathbb{E}_\theta[\mathsf{var}_\theta[A|X]]$

We first compare certain properties of the parameters $\mathbb{E}_\theta[\mathsf{cov}_\theta(Y, A|X)] = \mathbb{E}_\theta[(Y - b(X))(A - p(X))]$ and $\mathbb{E}_\theta[\mathsf{var}_\theta(A|X)] = \mathbb{E}_\theta[(A - p(X))^2]$, where we note that all the earlier results and definitions concerning $\mathbb{E}_\theta[\mathsf{cov}_\theta(Y, A|X)]$ also apply to $\psi(\theta) = \mathbb{E}_\theta[\mathsf{var}_\theta(A|X)]$ when we everywhere substitute A, p, $\hat{p}$ for Y, b, $\hat{b}$. However, we observe a first key difference between these two parameters, which are collected in the following lemma, whose proof is trivial once we note that for $\mathbb{E}_\theta[\mathsf{var}_\theta(A|X)]$, unlike $\mathbb{E}_\theta[\mathsf{cov}_\theta(Y, A|X)]$, $\mathsf{Bias}_\theta(\hat{\psi}_1) = \mathbb{E}_\theta[(p(X) - \hat{p}(X))^2]$, $\mathsf{Bias}_{\theta,k}(\hat{\psi}_1) = \mathbb{E}_\theta[\{\Pi[p(X) - \hat{p}(X)|\bar{\mathsf{Z}}_k]\}^2]$ and $\mathsf{TB}_{\theta,k}(\hat{\psi}_1) = \mathbb{E}_\theta[\{\Pi[p(X) - \hat{p}(X)|\bar{\mathsf{Z}}_k^\perp]\}^2]$ are all nonnegative. We thus have the following.

LEMMA 2.3. *The following statements are true for $\psi(\theta) = \mathbb{E}_\theta[\mathsf{var}_\theta[A|X]]$ but not always true for $\psi(\theta) = \mathbb{E}_\theta[\mathsf{cov}_\theta[Y, A|X]]$:*

(i) *$\mathsf{Bias}_{\theta,k}(\hat{\psi}_1)$ is nondecreasing in k (since, by Condition B, the space spanned by $\bar{\mathsf{Z}}_k$ increases with k), and thus, $\mathsf{TB}_{\theta,k}(\hat{\psi}_1)$ is nonincreasing in k. That is, for $k_2 > k_1$,*

$$0 \le \mathsf{Bias}_{\theta,k_1}(\hat{\psi}_1) \le \mathsf{Bias}_{\theta,k_2}(\hat{\psi}_1) \le \mathsf{Bias}_\theta(\hat{\psi}_1),$$

$$\mathsf{TB}_{\theta,k_1}(\hat{\psi}_1) \ge \mathsf{TB}_{\theta,k_2}(\hat{\psi}_1) \ge 0.$$

(ii) $\mathsf{Bias}_\theta(\hat{\psi}_{2,k}) \le \mathsf{Bias}_\theta(\hat{\psi}_1)$.

(iii) *For any $\delta > 0$, consider the null hypotheses*

$$\mathsf{H}_0(\delta) : \frac{|\mathsf{Bias}_\theta(\hat{\psi}_1)|}{\mathsf{s.e.}_\theta[\hat{\psi}_1]} \equiv \frac{|\mathsf{Bias}_{\theta,k}(\hat{\psi}_1) + \mathsf{TB}_{\theta,k}(\hat{\psi}_1)|}{\mathsf{s.e.}_\theta[\hat{\psi}_1]} < \delta \tag{2.6}$$

and its surrogate hypothesis

$$\mathsf{H}_{0,k}(\delta) : \frac{|\mathsf{Bias}_{\theta,k}(\hat{\psi}_1)|}{\mathsf{s.e.}_\theta[\hat{\psi}_1]} < \delta. \tag{2.7}$$

If $\mathsf{H}_0(\delta)$ (2.6) is true, then the surrogate null $\mathsf{H}_{0,k}(\delta)$ (2.7) is true. Hence rejection of the surrogate $\mathsf{H}_{0,k}(\delta)$ (2.7) implies rejection of $\mathsf{H}_0(\delta)$ (2.6).

Thus $\psi(\theta) = \mathbb{E}_\theta[\mathsf{var}_\theta(A|X)]$, unlike $\psi(\theta) = \mathbb{E}_\theta[\mathsf{cov}_\theta(A, Y|X)]$, belongs to the *monotone bias class*. The null hypothesis $\mathsf{H}_0(\delta)$ (2.6) states that $\mathsf{Bias}_\theta(\hat{\psi}_1)$ is less than a fraction δ of its standard error. In Theorem 3.2 and Theorem 4.2 below, we construct valid $\alpha^\dagger$-level tests for the null hypothesis $\mathsf{H}_{0,k}(\delta)$ (2.7). In Section 3.1 and Section 4.1, we consider the role of these null hypotheses when our goal is to either falsify (i) an analyst's claim that the Wald confidence interval centered at $\hat{\psi}_1$ has at least nominal coverage or (ii), less ambitiously, the analyst's justification for the claim.

REMARK 2.4. The simulation study reported in Table 1 was for the parameter $\psi(\theta) = \mathbb{E}_\theta[\mathsf{var}_\theta(A|X)]$. Were it not, our claim that the observation that the bias of $\hat{\psi}_{2,k}$ decreases as k increases as predicted by the theory developed in Section 2 would have been false. Similarly, our claim that the test $\widehat{\chi}_k^{(1)}(z_{\alpha^\dagger}, \delta)$ is an $\alpha^\dagger$-level test of $\mathsf{H}_0(\delta)$ (2.6) would also have been false.

[8] The notation $\mathsf{TB}_{\theta,k}(\hat{\psi}_1)$ was adopted because it is the so-called *truncation bias* in Robins et al. (2008).

In our simulation studies for the parameter $\mathbb{E}_\theta[\mathsf{cov}_\theta(Y, A|X)]$ reported in Table S8 and Table S11 in Section S9, the results were qualitatively similar to those in Table 1 (e.g., the MCav of $\hat{\psi}_{2,k}$ increased with k). However, this was due to the particular data generating process used and is not always true for $\psi(\theta) = \mathbb{E}_\theta[\mathsf{cov}_\theta(Y, A|X)]$.

An additional point in regard to the study reported in Table 1, the ratio of the MC bias 0.229 of $\hat{\psi}_1$ for $\psi(\theta) = \mathbb{E}_\theta[\mathsf{var}_\theta(A|X)]$ to the MCav 0.0161 of its estimated standard error was approximately 14. The theoretical prediction based on rates of convergence, ignoring constants, was reasonably close (given that we ignore unknown constants), being equal to 4.1, calculated as follows. In the simulation, $p(x)$ had a Hölder exponent s_p of 0.25 and, therefore, the conditional bias $\mathbb{E}_\theta[\{\hat{p}(X) - p(X)\}^2]$ was of order $n^{-2s_p/(2s_p+1)} = n^{-1/3}$, because we used a rate minimax estimator $\hat{p}(x)$ (see Section S9). Hence the order of the bias over the standard error is $n^{-1/3}/n^{-1/2} = n^{1/6}$, which evaluated at the sample size $n = 5000$ gives $4.1 = 5000^{1/6}$.

It follows from Remark 1.2 above that in the absence of further assumptions, $\mathsf{TB}_{\theta,k}(\hat{\psi}_1)$ could be of order 1 and cannot be consistently estimated without further assumptions on $(b, p, \hat{b}, \hat{p})$. However, it is immediate from equation (2.4) that the oracle second-order U-statistic estimator $\widehat{\mathbb{IF}}_{22,k}$[9] is an unbiased estimator of $\mathsf{Bias}_{\theta,k}(\hat{\psi}_1)$ conditional on the training sample,[10] where

$$\begin{aligned}\widehat{\mathbb{IF}}_{22,k} &\equiv \widehat{\mathbb{IF}}_{22,k}\big(\Omega_k^{-1}\big)\\ &:= \frac{1}{n(n-1)} \sum_{1\le i_1\ne i_2\le n} \widehat{\mathsf{IF}}_{22,k,i_1,i_2}\big(\Omega_k^{-1}\big),\end{aligned} \tag{2.8}$$

$$\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\big(\Omega_k^{-1}\big) = \big[\hat{\varepsilon}_b \bar{\mathsf{z}}_k(X)\big]_{i_1}^\top \Omega_k^{-1} \big[\bar{\mathsf{z}}_k(X)\hat{\varepsilon}_p\big]_{i_2}.$$

Thus the conditional bias of the bias corrected estimator[11] $\hat{\psi}_{2,k} \equiv \hat{\psi}_{2,k}(\Omega_k^{-1}) := \hat{\psi}_1 - \widehat{\mathbb{IF}}_{22,k}$ for $\psi(\theta)$ and conditional mean of $\hat{\psi}_{2,k}$ are

$$\begin{aligned}&\mathsf{Bias}_\theta(\hat{\psi}_{2,k}) \equiv \mathbb{E}_\theta\big[\hat{\psi}_{2,k} - \psi(\theta)\big] = \mathsf{TB}_{\theta,k}(\hat{\psi}_1),\\ &\mathbb{E}_\theta[\hat{\psi}_{2,k}] = \psi(\theta) + \mathsf{TB}_{\theta,k}(\hat{\psi}_1)\end{aligned} \tag{2.9}$$

since

$$\begin{aligned}\mathbb{E}_\theta\big[\hat{\psi}_{2,k} - \psi(\theta)\big] &= \mathbb{E}_\theta\big[\hat{\psi}_1 - \psi(\theta)\big] - \mathbb{E}_\theta[\widehat{\mathbb{IF}}_{22,k}]\\ &= \mathsf{Bias}_\theta(\hat{\psi}_1) - \mathsf{Bias}_{\theta,k}(\hat{\psi}_1)\\ &= \mathsf{TB}_{\theta,k}(\hat{\psi}_1).\end{aligned}$$

Thus for $\psi(\theta) = \mathbb{E}_\theta[\mathsf{var}_\theta(A|X)]$, we are certain that $\hat{\psi}_{2,k}$ has smaller bias than $\hat{\psi}_1$ and the bias of $\hat{\psi}_{2,k}$ decreases as we increase k, following Lemma 2.3(i).

2.2 Statistical Properties of $\widehat{\mathbb{IF}}_{22,k}$: Another Difference Between $\mathbb{E}_\theta[\mathsf{cov}_\theta(Y, A|X)]$ and $\mathbb{E}_\theta[\mathsf{var}_\theta(A|X)]$

Throughout the rest of this paper, our results require the following weak regularity conditions (Condition W) to hold:

CONDITION W.

1. All the eigenvalues of Ω_k are bounded away from 0 and ∞;
2. $A, Y, b(X), \hat{b}(X), p(X)$ and $\hat{p}(X)$ are bounded with probability 1;
3. $\|\bar{\mathsf{Z}}_k^\top \bar{\mathsf{Z}}_k\|_\infty \le Bk$ for some constant $B > 0$, $\|\Pi[b - \hat{b}|\bar{\mathsf{Z}}_k]\|_\infty \le C$ (where $\|\Pi[b - \hat{b}|\bar{\mathsf{Z}}_k] \equiv \bar{\mathsf{Z}}_k^\top \beta_{k,b-\hat{b}}$) and $\|\Pi[p - \hat{p}|\bar{\mathsf{Z}}_k]\|_\infty \le C$ (where $\|\Pi[p - \hat{p}|\bar{\mathsf{Z}}_k] \equiv \bar{\mathsf{Z}}_k^\top \beta_{k,p-\hat{p}}$) for some constant $C > 0$.

REMARK 2.5. Condition W(2) was assumed to allow us to focus on important issues. We believe we should be able replace the boundedness assumption with an assumption of light tails (Vershynin, 2018, Kuchibhotla and Chakrabortty, 2018). However, most of the existing results on U-statistics that we use, require the U-statistic kernel to be bounded.

Condition W(3) will only be needed in Section S3 when Ω_k^{-1} is unknown. Even though the main text only concerns the case with known Ω_k^{-1}, we still keep this assumption to emphasize its importance in the setting where Ω_k^{-1} must be estimated. Condition W(3) holds for Cohen–Daubechies–Vial wavelet series, B-spline series, and local polynomial partition series following from Belloni et al. (2015), Examples 3.8–3.10.

We have the following result regarding the statistical properties of the oracle estimator $\widehat{\mathbb{IF}}_{22,k}$ of the projected bias $\mathsf{Bias}_{\theta,k}(\hat{\psi}_1)$. For notational convenience, we define the following $L_2(P_\theta)$ norms:

$$\mathbb{L}_{2,b,k} := \big\{\mathbb{E}_\theta\big[\Pi\big[b(X) - \hat{b}(X)|\bar{\mathsf{Z}}_k\big]^2\big]\big\}^{1/2},$$

$$\mathbb{L}_{2,p,k} := \big\{\mathbb{E}_\theta\big[\Pi\big[p(X) - \hat{p}(X)|\bar{\mathsf{Z}}_k\big]^2\big]\big\}^{1/2}.$$

Note that $\mathbb{L}_{2,p,k}$ is equal to $\mathsf{Bias}_{\theta,k}(\hat{\psi}_1)$ when $\psi(\theta) = \mathbb{E}_\theta[\mathsf{var}_\theta(A|X)]$.

[9] Following the definitions in Robins et al. (2008), $\widehat{\mathbb{IF}}_{22,k}$ is the unique second-order influence function of $\mathsf{Bias}_{\theta,k}(\hat{\psi}_1)$ under the law $P_{\hat{\theta}}$. But the definition of $\widehat{\mathbb{IF}}_{22,k}$ in Robins et al. (2008) differs from that in the current paper in the sign; thus $\hat{\psi}_{2,k} \equiv \hat{\psi}_1 - \widehat{\mathbb{IF}}_{22,k}$ would be $\hat{\psi}_1 + \widehat{\mathbb{IF}}_{22,k}$ in Robins et al. (2008). We reversed the sign because it seems didactically useful to have $\widehat{\mathbb{IF}}_{22,k}$ be an unbiased estimator of $\mathsf{Bias}_{\theta,k}(\hat{\psi}_1)$. Robins et al. (2008) refer to $\psi(\theta) + \mathsf{TB}_{\theta,k}(\hat{\psi}_1)$ as the truncated parameter.

[10] Because it simply replaces the expectations of equation (2.5) by U-statistics.

[11] We discuss in Section S1.1 the connection between $\hat{\psi}_{2,k}$ and a triple sample splitting estimator proposed in Newey and Robins (2018) for $\psi(\theta) = \mathbb{E}_\theta[\mathsf{var}_\theta(A|X)]$.

THEOREM 2.6. *Under Condition* W, *with* $k, n \to \infty$, *and* $k = o(n^2)$, *conditional on the training sample, we have*:

(i) $\widehat{\mathbb{IIF}}_{22,k}$ *is unbiased for* $\mathsf{Bias}_{\theta,k}(\hat{\psi}_1)$ *with variance of order*

$$\frac{1}{n}\max\left\{\frac{k}{n}, \mathbb{L}_{2,b,k}^2, \mathbb{L}_{2,p,k}^2\right\},$$

where $\mathbb{L}_{2,b,k}$ *and* $\mathbb{L}_{2,p,k}$ *are defined above.*

(ii) $\frac{\widehat{\mathbb{IIF}}_{22,k} - \mathsf{Bias}_{\theta,k}(\hat{\psi}_1)}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IIF}}_{22,k}]}$ *converges in law to a standard normal* $N(0,1)$. *Further,* $\mathsf{s.e.}_\theta[\widehat{\mathbb{IIF}}_{22,k}] := \mathsf{var}_\theta^{1/2}[\widehat{\mathbb{IIF}}_{22,k}]$ *can be estimated by* $\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IIF}}_{22,k}] := \widehat{\mathsf{var}}^{1/2}[\widehat{\mathbb{IIF}}_{22,k}]$ *defined in Section* S5 *satisfying* $\frac{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IIF}}_{22,k}]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IIF}}_{22,k}]} = 1 + o_{P_\theta}(1)$.

(iii) $\widehat{\mathbb{IIF}}_{22,k} \pm z_{\alpha^\dagger/2}\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IIF}}_{22,k}]$ (*resp.*, $[\widehat{\mathbb{IIF}}_{22,k} - z_{\alpha^\dagger} \times \widehat{\mathsf{s.e.}}[\widehat{\mathbb{IIF}}_{22,k}], \infty)$) *is a* $(1-\alpha^\dagger)$ *asymptotic two-sided* (*resp., one-sided*) *Wald CI for* $\mathsf{Bias}_{\theta,k}(\hat{\psi}_1)$ *with length of order*

$$\frac{1}{\sqrt{n}}\max\left\{\sqrt{\frac{k}{n}}, \mathbb{L}_{2,b,k}, \mathbb{L}_{2,p,k}\right\}.$$

PROOF. The variance order of $\widehat{\mathbb{IIF}}_{22,k}$ is proved in Section S5. When $k = o(n^2)$ and $k \to \infty$ as $n \to \infty$, the conditional asymptotic normality of $\frac{\widehat{\mathbb{IIF}}_{22,k} - \mathsf{Bias}_{\theta,k}(\hat{\psi}_1)}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IIF}}_{22,k}]}$ follows directly from Hoeffding decomposition, with the conditional asymptotic normality of the degenerate second-order U-statistic part implied by Bhattacharya and Ghosh (1992), Corollary 1.2. □

REMARK 2.7. Now we consider a second key difference between the parameters $\mathbb{E}_\theta[\mathsf{var}_\theta(A|X)]$ and $\mathbb{E}_\theta[\mathsf{cov}_\theta(A, Y|X)]$. It follows from Theorem 2.6(i) that for $\mathbb{E}_\theta[\mathsf{cov}(A, Y|X)]$,

$$\mathsf{var}_\theta[\widehat{\mathbb{IIF}}_{22,k}] = O\left(\frac{1}{n}\left\{\frac{k}{n} + \mathbb{L}_{2,b,k}^2 + \mathbb{L}_{2,p,k}^2\right\}\right),$$

whereas for $\mathbb{E}_\theta[\mathsf{var}(A|X)]$,

$$\mathsf{var}_\theta[\widehat{\mathbb{IIF}}_{22,k}] = O\left(\frac{1}{n}\left\{\frac{k}{n} + \mathbb{L}_{2,p,k}^2\right\}\right).$$

For $\mathbb{E}_\theta[\mathsf{var}_\theta(A|X)]$, when $\mathbb{L}_{2,p,k}^2 = O(n^{-1/2})$, with $k = o(n)$, we always have $\mathsf{var}_\theta[\widehat{\mathbb{IIF}}_{22,k}] \ll n^{-1}$. However, for $\mathbb{E}_\theta[\mathsf{cov}_\theta(A, Y|X)]$, when $\mathsf{Bias}_{\theta,k}(\hat{\psi}_1) = O(n^{-1/2})$, $\mathbb{L}_{2,b,k}^2$ and $\mathbb{L}_{2,p,k}^2$ can still be $O(1)$, with $k = o(n)$, and then we have $\mathsf{var}_\theta[\widehat{\mathbb{IIF}}_{22,k}] \asymp n^{-1}$. We shall see below that the above implies the statistical behavior of tests of the hypothesis $\mathsf{H}_{0,k}(\delta)$ differ for $\mathbb{E}_\theta[\mathsf{var}_\theta(A|X)]$ and $\mathbb{E}_\theta[\mathsf{cov}_\theta(A, Y|X)]$.

REMARK 2.8. The qqplots in the left panel of Figure S2 (see Section S10.2) provide empirical evidence that, in our simulation experiments, in Section S9, the quantiles of $\widehat{\mathbb{IIF}}_{22,k}/\mathsf{s.e.}_\theta[\widehat{\mathbb{IIF}}_{22,k}]$ are close to normal quantiles.

REMARK 2.9. When k is of order greater than or equal to n^2, the conditional asymptotic normality of $\frac{\widehat{\mathbb{IIF}}_{22,k} - \mathsf{Bias}_{\theta,k}(\hat{\psi}_1)}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IIF}}_{22,k}]}$ does not hold. Moreover, when $k \gg n^2$, $\mathsf{var}_\theta[\widehat{\mathbb{IIF}}_{22,k}] \asymp \frac{k}{n^2}$ is of order greater than 1 and, therefore, $\widehat{\mathbb{IIF}}_{22,k}$ is not consistent for $\mathsf{Bias}_{\theta,k}(\hat{\psi}_1)$ even if $\mathsf{Bias}_{\theta,k}(\hat{\psi}_1)$ is of order 1. As mentioned in Section 1, when k is bounded (not growing with n), after standardization $\widehat{\mathbb{IIF}}_{22,k}$ converges to a Gaussian chaos distribution instead of a normal distribution, conditional on the training sample.

3. THE NULL HYPOTHESIS AND AN ORACLE TEST FOR $\mathbb{E}_\theta[\mathsf{var}_\theta(A|X)]$

3.1 The Null Hypothesis

We next consider the implications of rejection of the null hypothesis $\mathsf{H}_{0,k}(\delta)$ in the case of $\psi(\theta) = \mathbb{E}_\theta[\mathsf{var}_\theta(A|X)]$. In Section 4.1, we extend this discussion to $\psi(\theta) = \mathbb{E}_\theta[\mathsf{cov}_\theta(A, Y|X)]$. We shall require the following elementary lemma, which follows from the conditional asymptotic normality of $\hat{\psi}_1$ in Theorem 1.4.

LEMMA 3.1. *If* $\frac{\mathsf{Bias}_\theta(\hat{\psi}_1)}{\mathsf{s.e.}_\theta(\hat{\psi}_1)} = \delta$, *the actual asymptotic coverage of a two-sided* $(1-\alpha)$ *Wald CI* $\hat{\psi}_1 \pm z_{\alpha/2}\widehat{\mathsf{s.e.}}[\hat{\psi}_1]$ *for* $\psi(\theta)$ *is*

$$\mathsf{TC}_\alpha(\delta) := \Phi(z_{\alpha/2} - \delta) - \Phi(-z_{\alpha/2} - \delta). \tag{3.1}$$

The dependence of $\mathsf{TC}_\alpha(\delta)$ on δ for several α is shown in Figure 1. It follows that if $\mathsf{H}_0(\delta)$ is false, the true coverage rate is no more than $\mathsf{TC}_\alpha(\delta)$. It follows that $\mathsf{H}_0(\delta)$ is equivalent to the null hypothesis that the actual asymptotic coverage (given the training sample) of $\hat{\psi}_1 \pm z_{\alpha/2}\widehat{\mathsf{s.e.}}[\hat{\psi}_1]$ for $\psi(\theta)$ is greater than or equal to $\mathsf{TC}_\alpha(\delta)$. This result holds for both $\psi(\theta) = \mathbb{E}_\theta[\mathsf{cov}_\theta(A, Y|X)]$ and $\psi(\theta) = \mathbb{E}_\theta[\mathsf{var}_\theta(A|X)]$. For $\psi(\theta) = \mathbb{E}_\theta[\mathsf{var}_\theta(A|X)]$, but not for $\psi(\theta) = \mathbb{E}_\theta[\mathsf{cov}_\theta(A, Y|X)]$, if $\mathsf{H}_{0,k}(\delta)$ is false and, therefore, $\mathsf{H}_0(\delta)$ is false, the true coverage rate is no more than $\mathsf{TC}_\alpha(\delta)$.

In Theorem 3.2 below, we construct an asymptotically level $\alpha^\dagger$ test for the surrogate null hypothesis $\mathsf{H}_{0,k}(\delta)$, which by Lemma 2.3(iii) is also an asymptotically level $\alpha^\dagger$ test of $\mathsf{H}_0(\delta)$ for $\psi(\theta) = \mathbb{E}_\theta[\mathsf{var}_\theta(A|X)]$ but not for $\psi(\theta) = \mathbb{E}_\theta[\mathsf{cov}_\theta(Y, A|X)]$. Thus, one might reasonably ask whether our methods are useful for inference concerning the parameter $\mathbb{E}_\theta[\mathsf{cov}_\theta(Y, A|X)]$, a question to which we return in Section 4.

3.2 An Oracle Test

Based on the statistical properties of $\hat{\psi}_1$ and $\widehat{\mathbb{IIF}}_{22,k}$ summarized in Theorem 1.4 and Theorem 2.6, for $\psi(\theta) := \mathbb{E}_\theta[\mathsf{var}_\theta(A|X)]$, we now consider the properties

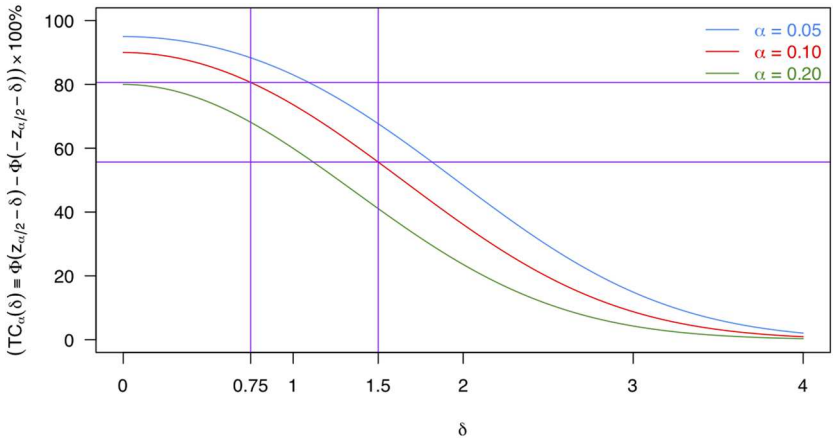

FIG. 1. $\mathsf{TC}_\alpha(\delta) \equiv \Phi(z_{\alpha/2}-\delta)-\Phi(-z_{\alpha/2}-\delta)$ *as a function of* δ *over several different* α*'s.*

of the following one-sided test $\widehat{\chi}_k^{(1)}(\zeta_k,\delta)$ of the surrogate null $\mathsf{H}_{0,k}(\delta)$:

$$\begin{aligned}\widehat{\chi}_k^{(1)}(\zeta_k,\delta) &\equiv \widehat{\chi}_k^{(1)}(\Omega_k^{-1};\zeta_k,\delta)\\ &:= \mathbb{1}\left\{\frac{\widehat{\mathbb{IF}}_{22,k}}{\widehat{\mathsf{s.e.}}[\hat{\psi}_1]}-\zeta_k\frac{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}]}{\widehat{\mathsf{s.e.}}[\hat{\psi}_1]}>\delta\right\}\end{aligned} \tag{3.2}$$

for user-specified $\zeta_k,\delta>0$. We use a one-sided test because the sign of $\mathsf{Bias}_{\theta,k}(\hat{\psi}_1)\geq 0$ is known *a priori*.

The following theorem characterizes the asymptotic level and power of the oracle one-sided test $\widehat{\chi}_k^{(1)}(\zeta_k,\delta)$ of the surrogate null $\mathsf{H}_{0,k}(\delta)$ when $\psi(\theta)=\mathbb{E}_\theta[\mathsf{var}_\theta(A|X)]$.

THEOREM 3.2. *For* $\psi(\theta)=\mathbb{E}_\theta[\mathsf{var}_\theta(A|X)]$, *under Condition* W, *when* $k\to\infty$ *but* $k=o(n)$, *for any given* $\delta,\zeta_k>0$, *suppose that* $\frac{\mathbb{L}_{2,p,k}^2}{\mathsf{s.e.}_\theta[\hat{\psi}_1]}=\frac{\mathsf{Bias}_{\theta,k}(\hat{\psi}_1)}{\mathsf{s.e.}_\theta[\hat{\psi}_1]}=\gamma$ *for some* (*sequence*) $\gamma=\gamma(n)$ (*where* $\gamma(n)$ *can diverge with* n), *then the rejection probability of* $\widehat{\chi}_k^{(1)}(\zeta_k,\delta)$ *converges to*

$$1-\Phi\left(\zeta_k-\lim_{n\to\infty}(\gamma-\delta)\frac{\mathsf{s.e.}_\theta[\hat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}]}\right) \tag{3.3}$$

as $n\to\infty$. *In particular,*

(1) *under* $\mathsf{H}_{0,k}(\delta):\gamma\leq\delta$, $\widehat{\chi}_k^{(1)}(\zeta_k,\delta)$ *rejects the null with probability less than or equal to* $1-\Phi(\zeta_k)$, *as* $n\to\infty$;

(2) *under the following alternative to* $\mathsf{H}_{0,k}(\delta)$: $\gamma=\delta+c$, *for any fixed* $c>0$ *or any diverging sequence* $c=c(n)\to\infty$, $\widehat{\chi}_k^{(1)}(\zeta_k,\delta)$ *rejects the null with probability converging to* 1, *as* $n\to\infty$.

REMARK 3.3. In Section S2, we prove equation (3.3). We now prove that equation (3.3) implies Theorem 3.2(1)–(2).

- Regarding (1), under $\mathsf{H}_{0,k}(\delta)$,

$$-(\gamma-\delta)\frac{\mathsf{s.e.}_\theta[\hat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}]}\geq 0,$$

which implies that the rejection probability is less than $1-\Phi(\zeta_k)$, as $n\to\infty$. Choose $\zeta_k=z_{\alpha^\dagger}$, $1-\Phi(\zeta_k)=1-\Phi(z_{\alpha^\dagger})=\alpha^\dagger$ and conclude that the test is a valid level $\alpha^\dagger$ test of the null.

- Regarding (2), under the alternative $\gamma=\delta+c$ for some $c>0$, it follows from Remark 2.7 and equation (3.3) that the rejection probability of $\widehat{\chi}_k^{(1)}(\zeta_k,\delta)$, as $n\to\infty$, is no smaller than

$$1-\Phi\left(\zeta_k-c\Theta(p,\hat{p},f_X,\bar{\mathsf{Z}}_k)\left\{\frac{k}{n}+\mathbb{L}_{2,p,k}\right\}^{-1}\right),$$

where $\Theta(p,\hat{p},f_X,\bar{\mathsf{Z}}_k)$ is some positive constant depending on the true regression function p, the estimated function $\hat{p}$ from the training sample, the density f_X of X and the chosen basis functions $\bar{\mathsf{Z}}_k$. For fixed $c>0$, $\mathsf{Bias}_{\theta,k}(\hat{\psi}_1)\equiv\mathbb{L}_{2,p,k}^2=(\delta+c)\mathsf{s.e.}_\theta(\hat{\psi}_1)=O(n^{-1/2})=o(1)$, which implies that the power converges to $1-\Phi(-\infty)=1$.

Theorem 3.2 implies that $\widehat{\chi}_k^{(1)}(z_{\alpha^\dagger},\delta)$ is an asymptotically valid level $\alpha^\dagger$ one-sided test of the surrogate null $\mathsf{H}_{0,k}(\delta)$. This allows us to define the following upper confidence bound that we briefly described in Section 1:

$$\begin{aligned}&\mathsf{UCB}^{(1)}(\Omega_k^{-1};\alpha,\alpha^\dagger)\\ &:=\mathsf{TC}_\alpha\left(\left[\frac{\widehat{\mathbb{IF}}_{22,k}-z_{\alpha^\dagger}\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}]}{\widehat{\mathsf{s.e.}}[\hat{\psi}_1]}\right]\right).\end{aligned} \tag{3.4}$$

Given the mapping $\mathsf{TC}_\alpha(\delta)$ between δ and the minimal asymptotic coverage of a nominal $(1-\alpha)$ two-sided Wald CI centered at $\hat{\psi}_1$ under $\mathsf{H}_{0,k}(\delta)$, the following corollary is an immediate consequence of Theorem 3.2:

COROLLARY 3.4. *Under the conditions in Theorem* 3.2, $\mathsf{UCB}^{(1)}(\Omega_k^{-1};\alpha,\alpha^\dagger)$ *is an asymptotically valid*[12]

[12]Recall that the validity of a nominal $(1-\alpha^\dagger)$ upper confidence bound is defined in equation (1.1) with P replaced by P_θ. That is,

nominal $(1-\alpha^\dagger)$ *upper confidence bound for the true coverage of a nominal* $(1-\alpha)$ *two-sided Wald CI centered at* $\hat{\psi}_1$ *for the parameter* $\mathbb{E}_\theta[\hat{\psi}_{2,k}] \equiv \psi(\theta) + \mathsf{TB}_{\theta,k}(\hat{\psi}_1)$ *when* $\psi(\theta) = \mathbb{E}_\theta[\mathsf{var}_\theta(A|X)]$.

Finally, the following corollary, implied by Theorem 3.2, Corollary 3.4 and Lemma 2.3, summarizes (1) the implication of $\widehat{\chi}_k^{(1)}(\zeta_k, \delta)$ on the actual null hypothesis of interest $\mathsf{H}_0(\delta)$ and (2) the implication of a nominal $(1-\alpha^\dagger)$ upper confidence bound $\mathsf{UCB}^{(1)}(\Omega_k^{-1}; \alpha, \alpha^\dagger)$ on the true coverage of a nominal $(1-\alpha)$ two-sided Wald CI centered at $\hat{\psi}_1$ for $\psi(\theta)$.

COROLLARY 3.5. *Under the conditions in Theorem 3.2*:

- $\widehat{\chi}_k^{(1)}(\zeta_k, \delta)$ *is an asymptotically level* $1 - \Phi(\zeta_k)$ *one-sided test of* $\mathsf{H}_0(\delta)$, *as* $n \to \infty$.
- $\mathsf{UCB}^{(1)}(\Omega_k^{-1}; \alpha, \alpha^\dagger)$ *is an asymptotically valid nominal* $(1-\alpha^\dagger)$ *upper confidence bound for the true coverage of a nominal* $(1-\alpha)$ *two-sided Wald CI centered at* $\hat{\psi}_1$ *for* $\psi(\theta) = \mathbb{E}_\theta[\mathsf{var}_\theta(A|X)]$. *That is, actual asymptotic coverage of a nominal* $(1-\alpha)$ *two-sided Wald CI centered at* $\hat{\psi}_1$ *is no greater than the random variable* $\mathsf{UCB}^{(1)}(\Omega_k^{-1}; \alpha, \alpha^\dagger)$ *with probability at least* $1-\alpha^\dagger$.

For $\psi(\theta) = \mathbb{E}_\theta[\mathsf{var}_\theta(A|X)]$, when $\widehat{\chi}_k^{(1)}(\zeta_k, \delta)$ rejects $\mathsf{H}_{0,k}(\delta)$, we should also reject $\mathsf{H}_0(\delta)$. Nevertheless, $\widehat{\chi}_k^{(1)}(\zeta_k, \delta)$ can be a powerless test under the alternative to $\mathsf{H}_0(\delta)$ for which $\mathsf{H}_{0,k}(\delta)$ holds. In fact, as discussed earlier, $\mathsf{Bias}_{\theta,k}(\hat{\psi}_1)$ may be zero and yet $\mathsf{Bias}_\theta(\hat{\psi}_1) = \mathsf{TB}_{\theta,k}(\hat{\psi}_1)$ may be order 1, owing to the fact we are not controlling the magnitude of $\mathsf{TB}_{\theta,k}(\hat{\psi}_1)$ by imposing sparsity or smoothness assumptions.

4. THE NULL HYPOTHESIS AND AN ORACLE TEST FOR $\mathbb{E}_\theta[\mathsf{cov}_\theta(A, Y|X)]$

4.1 The Null Hypothesis

In this section, we turn our attention to the parameter $\psi(\theta) = \mathbb{E}_\theta[\mathsf{cov}_\theta(A, Y|X)]$. In fact, the discussion in this section actually applies to any parameter $\psi(\theta)$ with a unique first order influence function depending on unknown regression functions or densities for which the absolute value $|\mathsf{TB}_{\theta,k}(\hat{\psi}_1)|$ of the truncation bias need not be a nonincreasing function of k, that is, outside the *monotone bias class*. In particular, it applies to the class of doubly robust functionals in Robins et al. (2008). Such parameters cover many causal parameters, including the average treatment effect and the effect of treatment on the treated, as well as many noncausal parameters. It is the class of parameters mentioned in the Section 1 for which our results are unavoidably less sharp. For the *monotone bias class*, we obtain much sharper results, as for $\psi(\theta) = \mathbb{E}_\theta[\mathsf{var}_\theta(A|X)]$ in Section 3.

In fact, for $\mathbb{E}_\theta[\mathsf{cov}_\theta(A, Y|X)]$ we shall have to settle for statements that are "in dialogue" with current practices and literature. To do so, we must return to the setting of Theorem 1.4 as, in current literature, authors often report a nominal $(1-\alpha)$ Wald CI $\hat{\psi}_1 \pm z_{\alpha/2}\widehat{\mathsf{s.e.}}[\hat{\psi}_1]$, or more commonly $\hat{\psi}_{\mathsf{cf},1} \pm z_{\alpha/2}\widehat{\mathsf{s.e.}}[\hat{\psi}_{\mathsf{cf},1}]$, and then appeal to Theorem 1.4 to support a claim that the true unconditional coverage is not less than nominal. Specifically, Theorem 1.4 implies validity under the null hypothesis $\mathsf{Bias}_\theta(\hat{\psi}_1) = o(n^{-1/2})$. The authors' justification for the claim that $\mathsf{Bias}_\theta(\hat{\psi}_1) = o(n^{-1/2})$ quite generally follows from making untestable complexity reducing assumptions (e.g., sparsity or smoothness) about the unknown nuisance regression functions appearing in the first-order influence function. Even given such complexity reducing assumptions, their appeal to the asymptotic $o(n^{-1/2})$ is implicitly justified by the tacit assumption that, at their sample size of $N = 2n = 2n_{\mathsf{tr}}$, they are nearly in asymptopia both in regards to the estimation sample n and in regards to the ratio $\mathsf{Bias}_\theta(\hat{\psi}_1)/\mathsf{s.e.}_\theta[\hat{\psi}_1]$ being close to its asymptotic limit of 0 (implied by their complexity reducing assumptions.)

However, most authors fail to quantify or operationalize their claims. In line with the approach of this paper, whenever a null hypothesis is defined in terms of an asymptotic rate of convergence such as $o(n^{-1/2})$ in the training sample data, we will (1) ask the authors to specify a positive number $\delta = \delta(N)$ possibly depending on the actual sample size N of their study and (2) then operationalize the asymptotic null hypothesis $\mathsf{Bias}_\theta(\hat{\psi}_1) = o(n^{-1/2})$ as the null hypothesis $\mathsf{H}_0(\delta)$. That is, we have the operationalized pair

$$\mathsf{NH}_0 : \mathsf{Bias}_\theta(\hat{\psi}_1) = o\big(n^{-1/2}\big),$$

$$\mathsf{H}_0(\delta) : \frac{|\mathsf{Bias}_\theta(\hat{\psi}_1)|}{\mathsf{s.e.}_\theta[\hat{\psi}_1]} < \delta$$

by which we mean that if $\mathsf{H}_0(\delta)$ is (not) rejected, we, by convention, will declare NH_0 (not) rejected. The authors' choice of δ depends on the degree of under coverage they are willing to tolerate. For example, if one allows the coverage of a 90% two-sided Wald CI centered at $\hat{\psi}_1$ to be at least 80.6% (or 55.6%), then the authors choose $\delta = 0.75$ as $\mathsf{TC}_{\alpha=0.1}(0.75) = 0.806$ (or choose $\delta = 1.5$ as $\mathsf{TC}_{\alpha=0.1}(1.5) = 0.556$).

Similarly, we have the surrogate operationalized pair

$$\mathsf{NH}_{0,k} : \mathsf{Bias}_{\theta,k}(\hat{\psi}_1) = o\big(n^{-1/2}\big),$$

$$\mathsf{H}_{0,k}(\delta) : \frac{|\mathsf{Bias}_{\theta,k}(\hat{\psi}_1)|}{\mathsf{s.e.}_\theta[\hat{\psi}_1]} < \delta.$$

$\mathsf{UCB}^{(1)}(\Omega_k^{-1}; \alpha, \alpha^\dagger)$ must be greater than the true asymptotic coverage probability of a $(1-\alpha)$ two-sided Wald CI covering $\mathbb{E}_\theta[\hat{\psi}_{2,k}]$ more than $(1-\alpha^\dagger) \times 100\%$ of the time over repeated sampling from the true data generating law P_θ.

Suppose now the authors of a research paper agree that in reporting $\hat{\psi}_1 \pm z_{\alpha/2}\widehat{\mathsf{s.e.}}[\hat{\psi}_1]$ as a $(1-\alpha)$ Wald CI for $\psi(\theta) = \mathbb{E}_\theta[\mathsf{cov}_\theta[Y, A|X]]$, their implicit or explicit null hypothesis is that $\mathsf{Bias}_\theta(\hat{\psi}_1)$ is $o(n^{-1/2})$. Further, suppose the test $\widehat{\chi}^{(2)}_k(z_{\alpha^\dagger}, \delta)$ developed in Section 4.2 rejects the surrogate $\mathsf{H}_{0,k}(\delta)$, equivalently $\mathsf{NH}_{0,k}$. However, unlike for $\mathbb{E}_\theta[\mathsf{var}_\theta[A|X]]$, rejecting the surrogate $\mathsf{H}_{0,k}(\delta)$ does not logically imply rejecting $\mathsf{H}_0(\delta)$, equivalently NH_0.

What, if anything, can be done? One approach is to adopt an additional "faithfulness" assumption under which rejection of the surrogate $\mathsf{NH}_{0,k}$ logically implies rejection of NH_0.

CONDITION FAITHFULNESS. Given a fixed k,

$$\frac{\mathsf{Bias}_\theta(\hat{\psi}_1)}{\mathsf{Bias}_{\theta,k}(\hat{\psi}_1)} = 1 + \frac{\mathsf{TB}_{\theta,k}(\hat{\psi}_1)}{\mathsf{Bias}_{\theta,k}(\hat{\psi}_1)}$$

is not $o(1)$.

One might find this assumption rather natural because it holds unless $\mathsf{TB}_{\theta,k}(\hat{\psi}_1)$ and $\mathsf{Bias}_{\theta,k}(\hat{\psi}_1)$ are of the same order and their leading constants sum to zero, which seems highly unlikely to be the case. In finite samples, we can also operationalize the above asymptotic faithfulness condition by choosing some $\delta' > 0$ and imposing the following.

CONDITION FAITHFULNESS(δ'). For a given k,

$$\left|\frac{\mathsf{Bias}_\theta(\hat{\psi}_1)}{\mathsf{Bias}_{\theta,k}(\hat{\psi}_1)}\right| = \left|1 + \frac{\mathsf{TB}_{\theta,k}(\hat{\psi}_1)}{\mathsf{Bias}_{\theta,k}(\hat{\psi}_1)}\right| \geq \delta'.$$

Under Condition Faithfulness(δ'), rejection of $\mathsf{H}_{0,k}(\delta)$ implies rejection of $\mathsf{H}_0(\delta\delta')$. If we choose $\delta' = 0.15$, Condition Faithfulness(δ') holds unless $-1.15 \leq \frac{\mathsf{TB}_{\theta,k}(\hat{\psi}_1)}{\mathsf{Bias}_{\theta,k}(\hat{\psi}_1)} \leq -0.85$. When we reject $\mathsf{H}_{0,k}(\delta)$ for some large δ, say $\delta = 10$, we will reject $\mathsf{H}_0(\delta\delta' = 1.5)$, suggesting that the true asymptotic coverage of a 90% two-sided Wald CI should be lower than 55.6%. To some extent, imposing Condition Faithfulness or Condition Faithfulness(δ') may seem inconsistent with the goal of falsifying the validity of reported Wald CIs without unverifiable assumptions.

Cauchy–Schwarz bias. What else can be done if we are not willing to impose Condition Faithfulness or Condition Faithfulness(δ')?

In what follows, we shall assume that the implicit or explicit goal in using a machine learning algorithm to learn the regression functions $b(x)$ and $p(x)$ is to construct $\hat{b}(x)$ and $\hat{p}(x)$ that (nearly) minimize the conditional mean square errors $\mathbb{E}_\theta[\{b(X) - \hat{b}(X)\}^2]$ and $\mathbb{E}_\theta[\{p(X) - \hat{p}(X)\}^2]$ over the set of functions computable by the algorithm. In fact, researchers who use the "training sample squared-error loss cross-validation" algorithm described in Remark 1.6 are explicitly acknowledging this as their goal.

It follows that researchers who report a nominal $(1-\alpha)$ Wald CI $\hat{\psi}_1 \pm z_{\alpha/2}\widehat{\mathsf{s.e.}}[\hat{\psi}_1]$ or $\hat{\psi}_{\mathsf{cf},1} \pm z_{\alpha/2}\widehat{\mathsf{s.e.}}(\hat{\psi}_{\mathsf{cf},1})$, based on a DRML estimator $\hat{\psi}_1$ for $\psi(\theta) = \mathbb{E}_\theta[\mathsf{cov}_\theta(A, Y|X)]$ should naturally appeal to the following Cauchy–Schwarz (CS) null hypothesis $\mathsf{NH}_{0,\mathrm{CS}}$ and its operationalization $\mathsf{H}_{0,\mathrm{CS}}(\delta)$

$$\begin{aligned}\mathsf{NH}_{0,\mathrm{CS}} : \mathsf{CSBias}_\theta(\hat{\psi}_1) &:= \{\mathbb{E}_\theta[\{b(X) - \hat{b}(X)\}^2] \\ &\quad \times \mathbb{E}_\theta[\{p(X) - \hat{p}(X)\}^2]\}^{1/2} \\ &= o(n^{-1/2}),\end{aligned} \tag{4.1}$$

$$\mathsf{H}_{0,\mathrm{CS}}(\delta) : \frac{\mathsf{CSBias}_\theta(\hat{\psi}_1)}{\mathsf{s.e.}_\theta[\hat{\psi}_1]} < \delta$$

as the *justification* of a validity claim that the Wald CI's true coverage of $\psi(\theta)$ is (within the tolerance level set by δ) nominal. The CS null hypothesis $\mathsf{NH}_{0,\mathrm{CS}}$ is the hypothesis that the Cauchy–Schwarz (CS) bias, $\mathsf{CSBias}_\theta(\hat{\psi}_1)$, is $o(n^{-1/2})$. We have the following logical orderings between the null hypotheses defined above.

LEMMA 4.1.

1. $\mathsf{NH}_{0,\mathrm{CS}} \Rightarrow \mathsf{NH}_0$, *and similarly* $\mathsf{H}_{0,\mathrm{CS}}(\delta) \Rightarrow \mathsf{H}_0(\delta)$;
2. $\mathsf{NH}_{0,\mathrm{CS}} \Rightarrow \mathsf{NH}_{0,k}$ *for all* k, *and similarly* $\mathsf{H}_{0,\mathrm{CS}}(\delta) \Rightarrow \mathsf{H}_{0,k}(\delta)$ *for all* k.

PROOF. The first part simply follows from CS inequality. The second part follows from the derivation below:

$$\begin{aligned}|\mathsf{Bias}_{\theta,k}(\hat{\psi}_1)| &= |\mathbb{E}_\theta[\Pi[b(X) - \hat{b}(X)|\bar{\mathsf{Z}}_k] \\ &\quad \times \Pi[p(X) - \hat{p}(X)|\bar{\mathsf{Z}}_k]]| \\ &\leq \mathbb{L}_{2,b,k}\mathbb{L}_{2,p,k} \\ &\leq \{\mathbb{E}_\theta[(b(X) - \hat{b}(X))^2]\}^{1/2} \\ &\quad \times \{\mathbb{E}_\theta[(p(X) - \hat{p}(X))^2]\}^{1/2} \\ &\equiv \mathsf{CSBias}_\theta(\hat{\psi}_1),\end{aligned} \tag{4.2}$$

where the first inequality follows from CS inequality and the second inequality is a consequence of the fact that a projection contracts $L_2(P_\theta)$ norms. □

However, the converse statements of Lemma 4.1 are not always true: for example, NH_0 may be true (and thus, by Theorem 1.4 the above the Wald CI centered at $\hat{\psi}_1$ is valid) even when the CS null hypothesis is false. Suppose we empirically falsify the *justification* $\mathsf{NH}_{0,\mathrm{CS}}$ ($\mathsf{H}_{0,\mathrm{CS}}(\delta)$) for the null hypothesis of actual interest NH_0 ($\mathsf{H}_0(\delta)$). Then, although logically NH_0 may be true, there seems to us, neither a substantive nor a philosophical reason to assume NH_0 is true in the absence of $\mathsf{NH}_{0,\mathrm{CS}}$. In Bayesian language, our (subjective) posterior probability that NH_0 is true conditional on $\mathsf{NH}_{0,\mathrm{CS}}$ being false is small; equivalently the rejection of $\mathsf{NH}_{0,\mathrm{CS}}$ undermines our belief in NH_0. Thus we will make the following.

CONDITION CS. If the CS null hypothesis $\mathsf{NH}_{0,\mathrm{CS}}$ and $\mathsf{H}_{0,\mathrm{CS}}(\delta)$ being true is used as the justification for the validity of the Wald interval $\hat{\psi}_1 \pm z_{\alpha/2}\widehat{\mathsf{s.e.}}(\hat{\psi}_1)$, but in fact are false, one should refuse to support claims whose validity rests on the truth of NH_0 or $\mathsf{H}_0(\delta)$; in particular, the claims that the Wald CIs centered at $\hat{\psi}_1$ have true coverage greater than or equal to their nominal.

Clearly, Condition CS will allow meaningful inferences regarding $\psi(\theta) = \mathbb{E}_\theta[\mathsf{cov}_\theta(A, Y|X)]$ only if it is possible to empirically reject the CS null hypothesis $\mathsf{H}_{0,\mathrm{CS}}(\delta)$. Indeed, it follows from Lemma 4.1(2), that the rejection of the surrogate $\mathsf{H}_{0,k}(\delta)$ implies rejection of $\mathsf{H}_{0,\mathrm{CS}}(\delta)$. In the next section, we will construct a test $\widehat{\chi}^{(2)}_k(\zeta_k, \delta)$ that can empirically reject $\mathsf{H}_{0,k}(\delta)$, and hence reject $\mathsf{H}_{0,\mathrm{CS}}(\delta)$ (and also reject $\mathsf{H}_0(\delta\delta')$ under Condition Faithfulness(δ')).

4.2 An Oracle Test

Based on the statistical properties of $\hat{\psi}_1$ and $\widehat{\mathbb{IF}}_{22,k}$ summarized in Theorem 1.4 and Theorem 2.6, for $\psi(\theta) := \mathbb{E}_\theta[\mathsf{cov}_\theta(A, Y|X)]$, we now consider the properties of the following two-sided test $\widehat{\chi}^{(2)}_k(\zeta_k, \delta)$ for $\mathsf{H}_{0,k}(\delta)$ (2.7):

$$\begin{aligned}\widehat{\chi}^{(2)}_k(\zeta_k, \delta) &\equiv \widehat{\chi}^{(2)}_k\big(\Omega_k^{-1}; \zeta_k, \delta\big) \\ &:= \mathbb{1}\left\{\frac{|\widehat{\mathbb{IF}}_{22,k}|}{\widehat{\mathsf{s.e.}}[\hat{\psi}_1]} - \zeta_k \frac{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}]}{\widehat{\mathsf{s.e.}}[\hat{\psi}_1]} > \delta\right\}\end{aligned} \tag{4.3}$$

for user-specified $\zeta_k, \delta > 0$. We use a two-sided test rather than a one-sided test because the sign of $\mathsf{Bias}_{\theta,k}(\hat{\psi}_1)$ is unknown *a priori*.

The following theorem characterizes the asymptotic level and power of the oracle two-sided test $\widehat{\chi}^{(2)}_k(\zeta_k, \delta)$ for $\mathsf{H}_{0,k}(\delta)$ (2.7) when $\psi(\theta) = \mathbb{E}_\theta[\mathsf{cov}_\theta(A, Y|X)]$.

THEOREM 4.2. *For $\psi(\theta) = \mathbb{E}_\theta[\mathsf{cov}_\theta(A, Y|X)]$, under Condition W, when $k \to \infty$ but $k = o(n)$, for any given $\delta, \zeta_k > 0$, suppose that $\frac{|\mathsf{Bias}_{\theta,k}(\hat{\psi}_1)|}{\mathsf{s.e.}_\theta[\hat{\psi}_1]} = \gamma$ for some (sequence) $\gamma = \gamma(n)$ (where $\gamma(n)$ can diverge with n), then the rejection probability of $\widehat{\chi}^{(2)}_k(\zeta_k, \delta)$ converges to*

$$\begin{aligned}&2 - \Phi\left(\zeta_k - \lim_{n\to\infty}(\gamma - \delta)\frac{\mathsf{s.e.}_\theta[\hat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}]}\right) \\ &\quad - \Phi\left(\zeta_k + \lim_{n\to\infty}(\gamma + \delta)\frac{\mathsf{s.e.}_\theta[\hat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}]}\right)\end{aligned} \tag{4.4}$$

as $n \to \infty$. In particular,

(1) *under $\mathsf{H}_{0,k}(\delta) : \gamma \le \delta$, $\widehat{\chi}^{(2)}_k(\zeta_k, \delta)$ rejects the null with probability less than or equal to $2(1 - \Phi(\zeta_k))$, as $n \to \infty$;*

(2) *under the following alternative to $\mathsf{H}_{0,k}(\delta)$: $\gamma = \delta + c$, for any diverging sequence $c = c(n) \to \infty$, $\widehat{\chi}^{(2)}_k(\zeta_k, \delta)$ rejects the null with probability converging to 1, as $n \to \infty$.*

(2') *If $\hat{b}$ and $\hat{p}$ converge to b and p in $L_2(P_\theta)$ norm, under the following alternative to $\mathsf{H}_{0,k}(\delta)$: $\gamma = \delta + c$, for any fixed $c > 0$ or any diverging sequence $c = c(n) \to \infty$, $\widehat{\chi}^{(2)}_k(\zeta_k, \delta)$ has rejection probability converging to 1, as $n \to \infty$.*

REMARK 4.3. In Section S2, we prove equation (4.4). We now prove that equation (4.4) implies Theorem 4.2(1)–(2) and (2').

- Regarding (1), under $\mathsf{H}_{0,k}(\delta) : \gamma \le \delta$,
$$-(\gamma - \delta)\frac{\mathsf{s.e.}_\theta[\hat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}]} \ge 0 \quad \text{and}$$
$$(\gamma + \delta)\frac{\mathsf{s.e.}_\theta[\hat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}]} \ge 0,$$
which implies that the rejection probability is less than or equal to $2 - 2\Phi(\zeta_k)$. Choose $\zeta_k = z_{\alpha^\dagger/2}$, $2(1 - \Phi(\zeta_k)) = 2\alpha^\dagger/2 = \alpha^\dagger$ and conclude that the test is a valid level $\alpha^\dagger$ test of the null.
- Theorem 4.2(2) and (2') are less sharp than Theorem 3.2(2) when $\psi(\theta) = \mathbb{E}_\theta[\mathsf{var}_\theta(A|X)]$. Under the alternative to $\mathsf{H}_{0,k}(\delta)$ with $\gamma = \delta + c$ for some $c > 0$, it follows from Theorem 2.6 and equation (4.4) that the rejection probability of $\widehat{\chi}^{(2)}_k(\zeta_k, \delta)$, as $n \to \infty$, is no smaller than
$$\begin{aligned}&2 - \Phi\Bigg(\zeta_k - c\Theta(b, p, \hat{b}, \hat{p}, f_X, \bar{\mathsf{Z}}_k) \\ &\quad \times \left\{\frac{k}{n} + \mathbb{L}_{2,p,k} + \mathbb{L}_{2,b,k}\right\}^{-1}\Bigg) \\ &\quad - \Phi(\infty),\end{aligned}$$
where $\Theta(b, p, \hat{b}, \hat{p}, f_X, \bar{\mathsf{Z}}_k)$ is some positive constant depending on the true regression functions b and p, the estimated functions $\hat{b}$, $\hat{p}$ from the training sample, the density f_X of X and the chosen basis functions $\bar{\mathsf{Z}}_k$. To have power approaching 1 to reject $\mathsf{H}_{0,k}(\delta)$, we need one of the following:
  - If one of $\mathbb{L}_{2,p,k}$ and $\mathbb{L}_{2,b,k}$ is $O(1)$, we need $c \to \infty$ to guarantee the rejection probability of $\widehat{\chi}^{(2)}_k(\zeta_k, \delta)$ to converge to $1 - \Phi(-\infty) = 1$. Hence we have Theorem 4.2(2).
  - If c is fixed, we need both $\mathbb{L}_{2,p,k}$ and $\mathbb{L}_{2,b,k}$ to be $o(1)$ to guarantee the rejection probability of $\widehat{\chi}^{(2)}_k(\zeta_k, \delta)$ to converge to $1 - \Phi(-\infty) = 1$. Note if $\hat{b}$ and $\hat{p}$ converge to b and p in $L_2(P_\theta)$-norm, then both $\mathbb{L}_{2,p,k}$ and $\mathbb{L}_{2,b,k}$ are $o(1)$. Hence we have Theorem 4.2(2').

Theorem 4.2 implies that $\widehat{\chi}^{(2)}_k(z_{\alpha^\dagger/2}, \delta)$ is an asymptotically valid level $\alpha^\dagger$ two-sided test of the surrogate null $\mathsf{H}_{0,k}(\delta)$, and hence by Lemma 4.1(2) it is also an asymptotically $\alpha^\dagger$ level test of $\mathsf{H}_{0,\mathrm{CS}}(\delta)$. Thus when

$\widehat{\chi}_k^{(2)}(z_{\alpha^\dagger/2}, \delta)$ rejects $\mathsf{H}_{0,k}(\delta)$, we also reject $\mathsf{H}_{0,\mathrm{CS}}(\delta)$ and by Condition CS, we conclude that we have no justification for assuming the validity of the Wald CI centered at $\hat{\psi}_1$ (even though $\mathsf{H}_{0,k}(\delta)$ and $\mathsf{H}_{0,\mathrm{CS}}(\delta)$ being false does not logically imply that $\mathsf{H}_0(\delta)$ is false and, therefore, does not logically imply a Wald CI centered at $\hat{\psi}_1$ is invalid).

On the other hand, $\widehat{\chi}_k^{(2)}(z_{\alpha^\dagger/2}, \delta)$ can be a powerless test for $\mathsf{H}_{0,\mathrm{CS}}(\delta)$ under certain laws P_θ: even when $\widehat{\chi}_k^{(2)}(z_{\alpha^\dagger/2}, \delta)$ fails to reject $\mathsf{H}_{0,k}(\delta)$ with (conditional) probability 1, $\mathsf{H}_{0,\mathrm{CS}}(\delta)$ may still be false.

Furthermore, $\widehat{\chi}_k^{(2)}(z_{\alpha^\dagger/2}, \delta)$ is not an asymptotically valid level $\alpha^\dagger$ test of $\mathsf{H}_0(\delta)$. However, if we assume Condition Faithfulness(δ'), then $\widehat{\chi}_k^{(2)}(z_{\alpha^\dagger/2}, \delta)$ is an asymptotically valid level $\alpha^\dagger$ test of $\mathsf{H}_0(\delta\delta')$. But it can be a powerless test of $\mathsf{H}_0(\delta\delta')$: when $\widehat{\chi}_k^{(2)}(z_{\alpha^\dagger/2}, \delta)$ fails to reject $\mathsf{H}_{0,k}(\delta)$, $\mathsf{H}_0(\delta\delta')$ may still be false even under Condition Faithfulness(δ').

Finally, because $|\mathsf{Bias}_\theta(\hat{\psi}_1)|$ need not exceed $|\mathsf{Bias}_{\theta,k}(\hat{\psi}_1)|$, the concept of upper confidence bound is not particularly useful for $\psi(\theta) = \mathbb{E}_\theta[\mathsf{cov}_\theta(A, Y|X)]$.

REMARK 4.4. We have shown that it is indeed possible to empirically reject the CS null hypothesis $\mathsf{H}_{0,\mathrm{CS}}(\delta)$ by testing $\mathsf{H}_{0,k}(\delta)$ using the two-sided test $\widehat{\chi}_k^{(2)}(\zeta_k, \delta)$. However, it is possible that $\mathsf{H}_{0,k}(\delta)$ is true whereas $\mathsf{H}_{0,\mathrm{CS}}(\delta)$ is false, as we only have $\mathsf{Bias}_{\theta,k}(\hat{\psi}_1) \le \mathsf{CSBias}_\theta(\hat{\psi}_1)$ but do not have control over the gap between these two quantities without making further unverifiable assumptions on the true regression functions b and p and their estimators $\hat{b}$ and $\hat{p}$. This raises the question whether we can test $\mathsf{H}_{0,\mathrm{CS}}(\delta) : \frac{\mathsf{CSBias}_\theta(\hat{\psi}_1)}{\mathsf{s.e.}_\theta(\hat{\psi}_1)} \le \delta$ more directly by instead testing the following surrogate null hypothesis $\mathsf{H}_{0,\mathrm{CS},k}(\delta) : \frac{\mathsf{CSBias}_{\theta,k}(\hat{\psi}_1)}{\mathsf{s.e.}_\theta(\hat{\psi}_1)} \le \delta$ where $\mathsf{CSBias}_{\theta,k}(\hat{\psi}_1) = \mathbb{L}_{2,b,k}\mathbb{L}_{2,p,k}$. We show in Section S7 that it is still possible but we require multiple testing to increase the power to reject $\mathsf{H}_{0,\mathrm{CS},k}(\delta)$ when it is in fact false.

5. TESTING THE VALIDITY OF WALD CIs OF $\hat{\psi}_1$ WITH $k > n$ FOR $\psi(\theta) = \mathbb{E}_\theta[\mathsf{var}_\theta(A|X)]$

The tests developed in the previous sections restrict $k = o(n)$. In this section, we instead consider the case $k \gg n$ yet $k = o(n^2)$. We only consider the parameter $\psi(\theta) = \mathbb{E}_\theta[\mathsf{var}_\theta(A|X)]$.[13] Recall that $\mathsf{Bias}_{\theta,k}(\hat{\psi}_1)$ is nondecreasing in k (see Lemma 2.3) under Condition B. Further, when $k > n$, the variance of $\widehat{\mathbb{IF}}_{22,k}$ is always of order k/n^2, and thus increases with k and exceeds the order of $\mathsf{var}_\theta(\hat{\psi}_1)$. We exploit this bias-variance trade-off below. Although $\psi(\theta) = \mathbb{E}_\theta[\mathsf{cov}_\theta(A, Y|X)]$ does not have a bias nondecreasing in k, results we obtained concerning $\psi(\theta) = \mathbb{E}_\theta[\mathsf{var}_\theta(A|X)]$ can be extended to the parameter $\mathsf{CSBias}_\theta(\hat{\psi}_1)$ discussed above and in Section S7, although we omit the details. We continue to assume that Ω_k^{-1} is known.

If $k_0 = o(n)$, then for $\psi(\theta) = \mathbb{E}_\theta[\mathsf{var}_\theta(A|X)]$, we may always prefer to report a Wald CI centered at $\hat{\psi}_{2,k_0}$[14] than one centered at $\hat{\psi}_1$ for the following reason: we know $\mathsf{Bias}_\theta(\hat{\psi}_{2,k_0}) \le \mathsf{Bias}_\theta(\hat{\psi}_1)$ and yet the variances of $\hat{\psi}_{2,k_0}$ and $\hat{\psi}_1$ are close (i.e., of the same order). This choice naturally raises the question as to whether $\hat{\psi}_{2,k_0} \pm z_{\alpha/2}\widehat{\mathsf{s.e.}}[\hat{\psi}_{2,k_0}]$ covers $\psi(\theta)$ at its nominal level, which we operationalize as the null hypothesis $\mathsf{H}_{0,2,k_0}(\delta) : \frac{\mathsf{Bias}_\theta(\hat{\psi}_{2,k_0})}{\mathsf{s.e.}_\theta(\hat{\psi}_{2,k_0})} \le \delta$.

If $\mathsf{H}_{0,2,k_0}(\delta)$ is rejected, we may choose to report $\hat{\psi}_{2,k}$ for some $k > n$ to further reduce bias at the the price of inflating the variance $\mathsf{var}_\theta(\hat{\psi}_{2,k}) \asymp k/n^2$ whose order then exceeds $\mathsf{var}_\theta(\hat{\psi}_1) \asymp 1/n$. Our goal is to find the values of k for which we do not have empirical evidence that the Wald CI centered at $\hat{\psi}_{2,k}$ undercovers. We operationalize this goal as testing the null hypotheses in the following set, with bounded cardinality J,

$$\left\{\mathsf{H}_{0,2,k}(\delta) : \frac{\mathsf{Bias}_\theta(\hat{\psi}_{2,k})}{\mathsf{s.e.}_\theta(\hat{\psi}_{2,k})} = \frac{\mathsf{TB}_{\theta,k}(\hat{\psi}_1)}{\mathsf{s.e.}_\theta(\hat{\psi}_{2,k})} \le \delta, k \in \mathcal{K}_J\right\}, \tag{5.1}$$

where

$$\mathcal{K}_J := \{k_0 < n < k_1 < \cdots < k_{J-1} = o(n^2) : k_0 = o(n), k_{j-1} = o(k_j), j = 1, \ldots, J-1\}.$$

Note that the hypotheses in the above set are ordered: for any $k_1 < k_2 \in \mathcal{K}_J$, $\mathsf{H}_{0,2,k_1}(\delta) \Rightarrow \mathsf{H}_{0,2,k_2}(\delta)$ because $\mathsf{Bias}_\theta(\hat{\psi}_{2,k_1}) \ge \mathsf{Bias}_\theta(\hat{\psi}_{2,k_2})$ whereas $\mathsf{s.e.}_\theta(\hat{\psi}_{2,k_1}) \ll \mathsf{s.e.}_\theta(\hat{\psi}_{2,k_2})$. Hence if for each $k \in \mathcal{K}_J$ we have a level $\alpha_k^\dagger$ test, the following sequential test protects the level for each hypothesis $\mathsf{H}_{0,2,k}(\delta)$. See Rosenbaum (2008), Proposition 1, for the proof.

DEFINITION 5.1. Given a sequence of desired levels $\{0 < \alpha_k^\dagger \le \frac{1}{2}, k \in \mathcal{K}_J\}$. For $j = 0, \ldots, J-1$, at $k = k_j$:

- If the level $\alpha_k^\dagger$ test of $\mathsf{H}_{0,2,k}(\delta)$ rejects, set $k = k_{j+1}$ and repeat.
- Otherwise, we declare failure to reject $\mathsf{H}_{0,2,k_{j'}}(\delta)$ for all $j' \ge j$ and stop.

[13] The variance of $\widehat{\mathbb{IF}}_{22,k}$ is of order k/n^2 when $k \gg n$.

[14] Without loss of generality, we assume $\mathsf{var}_\theta[\widehat{\mathbb{IF}}_{22,k_0}] \asymp k_0/n^2$.

In particular, for any $j = 0, 1, \ldots, J-2$, we define the following test of $\mathsf{H}_{0,2,k_j}(\delta)$, given the desired level $\alpha_{k_j}^{\dagger}$,

$$\widehat{\chi}_{2,k_j}(z_{\alpha_{k_j}^{\dagger}}, \delta) := \max\{\widehat{\chi}_{2,k_j \to k'}(z_{\alpha_{k_j}^{\dagger}/(J-j-1)}, \delta), \; k' \in \mathcal{K}_J^{-j} := \mathcal{K}_J \setminus \{k_0, \ldots, k_j\}\}, \tag{5.2}$$

where[15]

$$\widehat{\chi}_{2,k_j \to k'}(z_{\alpha_{k_j}^{\dagger}/(J-j-1)}, \delta) := \mathbb{1}\left\{\frac{\widehat{\mathbb{IF}}_{22,k'} - \widehat{\mathbb{IF}}_{22,k_j}}{\widehat{\mathsf{s.e.}}(\hat{\psi}_{2,k_j})} - z_{\alpha_{k_j}^{\dagger}/(J-j-1)} \frac{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k'}]}{\widehat{\mathsf{s.e.}}(\hat{\psi}_{2,k_j})} > \delta\right\}. \tag{5.3}$$

$\widehat{\chi}_{2,k_j}(z_{\alpha^{\dagger}}, \delta)$ implicitly tests $J - j - 1$ surrogate hypotheses[16] associated with the actual null hypothesis of interest $\mathsf{H}_{0,2,k_j}(\delta)$. We choose the cutoff $z_{\alpha^{\dagger}/(J-j-1)}$ in $\widehat{\chi}_{2,k_j \to k'}(z_{\alpha^{\dagger}/(J-j-1)}, \delta)$ to protect the level of $\widehat{\chi}_{2,k_j}(z_{\alpha^{\dagger}}, \delta)$ by adjusting for multiple testing.

REMARK 5.2. We use Figure 2 to visually illustrate the *sequential test* given in Definition 5.1 using $\widehat{\chi}_{2,k_j}(z_{\alpha^{\dagger}}, \delta)$. We use the same level $\alpha^{\dagger}$ for each k_j in this example. Figure 2 displays one hypothetical dataset drawn from P_θ. Reading from the top ($j' = 0$) to the bottom panel ($j' = 2$):

1. The y-values of the points are $\frac{\hat{\psi}_{2,k_j}}{\widehat{\mathsf{s.e.}}[\hat{\psi}_{2,k_{j'}}]} - \frac{\delta}{2}$ for $j = j' + 1, \ldots, J-1$. As shown in the plot, any given point moves closer to 0 from top ($j' = 0$) to bottom ($j' = 2$) because $\widehat{\mathsf{s.e.}}(\hat{\psi}_{2,k_0}) \ll \widehat{\mathsf{s.e.}}(\hat{\psi}_{2,k_1}) \ll \widehat{\mathsf{s.e.}}(\hat{\psi}_{2,k_2})$ when $k_0 \ll k_1 \ll k_2$.
2. The length of the error bar associated with k_j is $z_{\alpha^{\dagger}/(J-j'-1)} \frac{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k_j}]}{\widehat{\mathsf{s.e.}}[\hat{\psi}_{2,k_{j'}}]}$, which decreases as we go from the top ($j' = 0$) to the bottom ($j' = 2$) panel. This reflects the fact that $\widehat{\mathsf{s.e.}}(\hat{\psi}_{2,k_0}) \ll \widehat{\mathsf{s.e.}}(\hat{\psi}_{2,k_1}) \ll \widehat{\mathsf{s.e.}}(\hat{\psi}_{2,k_2})$ when $k_0 \ll k_1 \ll k_2$ while $z_{\alpha^{\dagger}/(J-1)} \asymp z_{\alpha^{\dagger}/(J-2)} \asymp z_{\alpha^{\dagger}/(J-3)}$.

The *sequential test* for this example proceeds as follows:

- The upper panel of Figure 2 corresponds to be the test of $\mathsf{H}_{0,2,k_0}(\delta)$. The length of the error bar at each k_j is $z_{\alpha^{\dagger}/(J-1)} \frac{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k_j}]}{\widehat{\mathsf{s.e.}}[\hat{\psi}_{2,k_0}]}$. The upper end of each error bar is $\frac{\hat{\psi}_{2,k_j}}{\widehat{\mathsf{s.e.}}[\hat{\psi}_{2,k_0}]} - \frac{\delta}{2} + z_{\alpha^{\dagger}/(J-1)} \frac{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k_j}]}{\widehat{\mathsf{s.e.}}[\hat{\psi}_{2,k_0}]}$. If the point at k_0 (blue colored) lies outside at least one of the error bars to its right, we reject $\mathsf{H}_{0,2,k_0}(\delta)$. This corresponds to the test $\widehat{\chi}_{2,k_0}(z_{\alpha^{\dagger}}, \delta)$ (see equation (5.2)). We choose the cutoff $z_{\alpha^{\dagger}/(J-1)}$ to adjust for the $J-1$ multiple comparisons. As shown in the plot, we reject $\mathsf{H}_{0,2,k_0}(\delta)$ because the blue point at k_0 is outside the error bar at k_{J-2} (purple).
- As $\mathsf{H}_{0,2,k_0}(\delta)$ is rejected, we next test $\mathsf{H}_{0,2,k_1}(\delta)$, as shown in the middle panel of Figure 2. To test $\mathsf{H}_{0,2,k_1}(\delta)$, we follow the above procedure. In the middle panel, the upper end of the error bars for a given k_j equals $\frac{\hat{\psi}_{2,k_j}}{\widehat{\mathsf{s.e.}}[\hat{\psi}_{2,k_1}]} - \frac{\delta}{2} + z_{\alpha^{\dagger}/(J-2)} \frac{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k_j}]}{\widehat{\mathsf{s.e.}}[\hat{\psi}_{2,k_1}]}$, $j = 2, \ldots, J-1$. When $\frac{\hat{\psi}_{2,k_1}}{\widehat{\mathsf{s.e.}}[\hat{\psi}_{2,k_1}]} - \frac{\delta}{2}$ (the leftmost green point) lies outside at least one of the error bars to its right, we reject $\mathsf{H}_{0,2,k_1}(\delta)$. This corresponds to the test $\widehat{\chi}_{2,k_1}(z_{\alpha^{\dagger}}, \delta)$ (see equation (5.2)). We reject $\mathsf{H}_{0,2,k_1}(\delta)$ because the green point $\frac{\hat{\psi}_{2,k_1}}{\widehat{\mathsf{s.e.}}[\hat{\psi}_{2,k_1}]} - \frac{\delta}{2}$ at k_1 is outside the error bar at k_{J-2} (purple).
- We continue to test $\mathsf{H}_{0,2,k_2}(\delta)$, as shown in the lower panel of Figure 2. The upper end of the error bar for a given k_j equals $\frac{\hat{\psi}_{2,k_j}}{\widehat{\mathsf{s.e.}}[\hat{\psi}_{2,k_2}]} - \frac{\delta}{2} + z_{\alpha^{\dagger}/(J-3)} \frac{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k_j}]}{\widehat{\mathsf{s.e.}}[\hat{\psi}_{2,k_2}]}$ for $j = 3, \ldots, J-1$. We fail to reject $\mathsf{H}_{0,2,k_2}(\delta)$ because $\frac{\hat{\psi}_{2,k_2}}{\widehat{\mathsf{s.e.}}(\hat{\psi}_{2,k_2})} - \frac{\delta}{2}$ (the leftmost black point) is covered by all the error bars to its right.
- We thus terminate the sequential test and declare failure to reject $\mathsf{H}_{0,2,k}(\delta)$ for all $k \geq k_2$.

The result below shows that the sequential test given in Definition 5.1 using $\widehat{\chi}_{2,k_j}(z_{\alpha_{k_j}^{\dagger}}, \delta)$ protects the desired level for each null hypothesis $\mathsf{H}_{0,2,k_j}(\delta)$ in the set given in equation (5.1). It follows from Proposition 5.5 below.

PROPOSITION 5.3. *Under Condition W, for every $k_j \in \mathcal{K}_J$, $\widehat{\chi}_{2,k_j}(z_{\alpha_{k_j}^{\dagger}}, \delta)$ is an asymptotic level $\alpha_{k_j}^{\dagger}$ test of the null hypothesis $\mathsf{H}_{0,2,k_j}(\delta)$. Consequently, the sequential test defined in Definition 5.1 using $\widehat{\chi}_{2,k_j}(z_{\alpha_{k_j}^{\dagger}}, \delta)$ is an asymptotically level $\alpha_{k_j}^{\dagger}$ test for every individual null hypothesis $\mathsf{H}_{0,2,k_j}(\delta)$ in $\mathcal{K}_J$.*

REMARK 5.4. We have assumed that J is bounded for technical reasons: we need the joint conditional asymptotic normality of $\widehat{\mathbb{IF}}_{22,k}$ for $k \in \mathcal{J}$, which is not guaranteed if $J \to \infty$ as $n \to \infty$. It is possible to relax the boundedness assumption on J using exponential inequalities for U-statistics rather than normality to set critical values. But to do so requires that we estimate the constants in the exponential inequalities, which is left for future work.

[15] We can choose $\widehat{\mathsf{s.e.}}(\hat{\psi}_{2,k_0}) = \widehat{\mathsf{s.e.}}(\hat{\psi}_1)$ (as we have assumed $\mathsf{s.e.}_\theta(\widehat{\mathbb{IF}}_{22,k_0}) \asymp \sqrt{k_0}/n \ll n^{-1/2}$ in footnote 14) and $\widehat{\mathsf{s.e.}}(\hat{\psi}_{2,k}) = \widehat{\mathsf{s.e.}}(\widehat{\mathbb{IF}}_{22,k})$ for any $k \gg n$, where $\widehat{\mathsf{s.e.}}(\widehat{\mathbb{IF}}_{22,k})$ is given in Theorem 2.6 (as $\mathsf{s.e.}_\theta(\widehat{\mathbb{IF}}_{2,k}) \asymp \sqrt{k}/n \gg n^{-1/2}$).

[16] We explain why we test multiple surrogate hypotheses instead of single hypothesis in Remark S8.3.

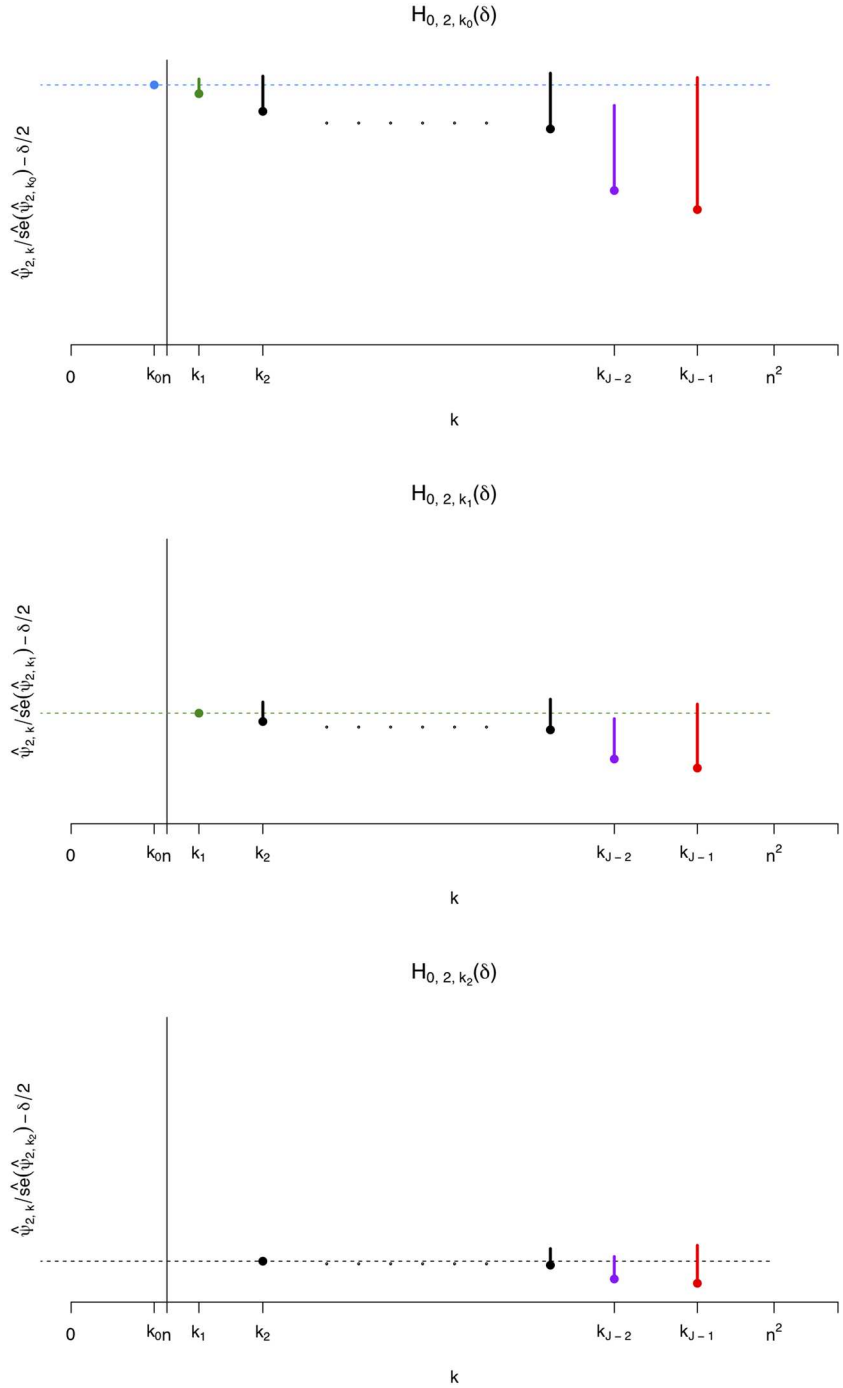

FIG. 2. *An illustration of the sequential test. Depicted is a hypothetical data (one realization from the true data generating law P_θ in which the sequential test rejects both $H_{0,2,k_0}(\delta)$ and $H_{0,2,k_1}(\delta)$ but fails to reject $H_{0,2,k_2}(\delta)$. The error bars and points are defined in Remark 5.2.*

The following result, which is a consequence of Proposition S8.2, summarizes the asymptotic power of the test $\widehat{\chi}_{2,k_j}(z_{\alpha^\dagger_{k_j}}, \delta)$ when the null hypothesis $\mathsf{H}_{0,2,k_j}(\delta)$ is false, for any given $k_j \in \mathcal{K}_J$.

PROPOSITION 5.5. *Under Condition W, for a given $j = 0, \ldots, J-1$, let $k = k_j$. Given any $\delta > 0$, suppose that $\frac{\mathsf{Bias}_\theta(\hat{\psi}_{2,k})}{\mathsf{s.e.}_\theta(\hat{\psi}_{2,k})} = \gamma$ for some (sequence) $\gamma \equiv \gamma(n)$ and $\frac{\mathsf{Bias}_{\theta,k'}(\hat{\psi}_{2,k})}{\mathsf{s.e.}_\theta(\hat{\psi}_{2,k})} = \gamma_{k'}$ for some (sequence) $\gamma_{k'} \equiv \gamma_{k'}(n)$,[17] $\widehat{\chi}_{2,k}(z_{\alpha^\dagger_k}, \delta)$ rejects $\mathsf{H}_{0,2,k}(\delta)$ with probability that lies in the following interval*:

$$
\left[\begin{array}{l}
\max\left\{1 - \Phi\left(z_{\alpha^\dagger/(J-j-1)}\right.\right. \\
\quad \left.\left. - \lim_{n\to\infty}(\gamma_{k'} - \delta)\dfrac{\mathsf{s.e.}_\theta(\hat{\psi}_{2,k})}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k'}]}\right), k' \in \mathcal{K}_J^{-j}\right\}, \\
\min\left\{\displaystyle\sum_{k'\in\mathcal{K}_J^{-j}} 1 - \Phi\left(z_{\alpha^\dagger/(J-j-1)}\right.\right. \\
\quad \left.\left. - \lim_{n\to\infty}(\gamma_{k'} - \delta)\dfrac{\mathsf{s.e.}_\theta(\hat{\psi}_{2,k})}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k'}]}\right), 1\right\}
\end{array}\right] \tag{5.4}
$$

as $n \to \infty$. In particular, under the following alternative to $\mathsf{H}_{0,2,k}(\delta)$: if there exists $k' \in \mathcal{K}_J^{-j}$ such that $\gamma_{k'} = \delta + c$ with $c \gg \sqrt{\frac{k'}{\max\{k,n\}}}$, then the test $\widehat{\chi}_{2,k}(z_{\alpha^\dagger_k}, \delta)$ rejects the null with probability approaching 1, as $n \to \infty$.

REMARK 5.6. Proposition 5.5 follows from Proposition S8.2 (analogous to Theorem 3.2) and the definition of $\widehat{\chi}_{2,k}(z_{\alpha^\dagger_k}, \delta)$ in equation (5.2).

In Proposition S8.2, we prove that for $k = k_j$, $\widehat{\chi}_{2,k\to k'}(z_{\alpha^\dagger/(J-j-1)}, \delta)$ rejects the null hypothesis $\mathsf{H}_{0,2,k\to k'}(\delta) : \frac{\mathsf{Bias}_\theta(\hat{\psi}_{2,k_j}) - \mathsf{Bias}_\theta(\hat{\psi}_{2,k'})}{\mathsf{s.e.}_\theta(\hat{\psi}_{2,k_j})} \le \delta$ with probability

$$
1 - \Phi\left(z_{\alpha^\dagger/(J-j-1)} - \lim_{n\to\infty}(\gamma_{k'} - \delta)\frac{\mathsf{s.e.}_\theta(\hat{\psi}_{2,k})}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k'}]}\right).
$$

Here,

$$
\mathsf{Bias}_\theta(\hat{\psi}_{2,k_j}) - \mathsf{Bias}_\theta(\hat{\psi}_{2,k'}) = \mathbb{E}_\theta[\widehat{\mathbb{IF}}_{22,k'} - \widehat{\mathbb{IF}}_{22,k_j}] \ge 0.
$$

$\mathsf{H}_{0,2,k_j\to k'}(\delta)$ is the surrogate null hypothesis associated with $\mathsf{H}_{0,2,k_j}(\delta)$ in the following sense (see also Lemma S8.1): $\mathsf{H}_{0,2,k_j}(\delta) \Rightarrow \mathsf{H}_{0,2,k_j\to k'}(\delta)$ for all $k' \in \mathcal{K}_J^{-j}$, therefore, if one of $\mathsf{H}_{0,2,k_j\to k'}(\delta)$ is false, $\mathsf{H}_{0,2,k_j}(\delta)$ is false.

Under $\mathsf{H}_{0,2,k_j\to k'}(\delta)$, $\widehat{\chi}_{2,k\to k'}(z_{\alpha^\dagger/(J-j-1)}, \delta)$ rejects $\mathsf{H}_{0,2,k\to k'}(\delta)$ no more than $\alpha^\dagger/(J-j-1)$. Under the following alternative $\gamma_{k'} - \delta \gg \sqrt{\frac{k'}{\max\{k_j,n\}}}$,[18] $\widehat{\chi}_{2,k_j\to k'}(z_{\alpha^\dagger/(J-j-1)}, \delta)$ rejects $\mathsf{H}_{0,2,k_j\to k'}(\delta)$ with probability approaching 1.

[17] $\gamma \ge \gamma_{k'}$ for any $k' \in \mathcal{K}_J^{-j}$

[18] The need for a diverging alternative is a consequence of the variance of the statistic $\frac{\widehat{\mathbb{IF}}_{22,k'}}{\widehat{\mathsf{s.e.}}(\hat{\psi}_{2,k_j})}$ being of order $k'/\max\{k_j, n\}$.

6. CONCLUDING REMARKS

We conclude by mentioning some open problems:

- We did not consider how to optimally select the basis functions $\bar{\mathsf{Z}}_k$ from a dictionary of $K > k$ basis functions. Data driven basis selection in the training sample has the potential of markedly increased power.
- As mentioned in Section 1 (also see Section S3 in Liu, Mukherjee and Robins, 2020), for unknown Ω_k^{-1}, we lack theoretical guarantees as to the statistical properties of the estimators/tests that performed the best in our simulation studies.

Once these open problems are solved, we would suggest that testing the undercoverage of Wald confidence intervals centered at DRML estimators would become routine.

APPENDIX: ESTIMATORS FOR $\mathsf{Bias}_{k,\theta}(\hat{\psi}_1)$ WHEN Ω_k^{-1} IS UNKNOWN

In this appendix, we describe the data-adaptive test and the upper confidence bound used in the simulation studies of Section 1 when Ω_k^{-1} is unknown:

$$
\begin{aligned}
&\widehat{\chi}_k^{(1)}(\widehat{\Omega}_k^{-1}; \zeta_k, \delta) \\
&= \mathbb{1}\left\{\frac{\widehat{\mathbb{IF}}_{22,k}(\widehat{\Omega}_k^{-1})}{\widehat{\mathsf{s.e.}}(\hat{\psi}_1)}\right. \\
&\qquad \left. -\zeta_k\frac{\widehat{\mathsf{s.e.}}(\widehat{\mathbb{IF}}_{22,k}(\widehat{\Omega}_k^{-1}))}{\widehat{\mathsf{s.e.}}(\hat{\psi}_1)} > \delta\right\} \quad \text{(see Table 1)},
\end{aligned} \tag{A.1}
$$

$$
\begin{aligned}
&\mathsf{UCB}^{(1)}(\widehat{\Omega}_k^{-1}; \alpha, \alpha^\dagger) \\
&:= \mathsf{TC}_\alpha\left(\left[\frac{\widehat{\mathbb{IF}}_{22,k}(\widehat{\Omega}_k^{-1}) - z_{\alpha^\dagger}\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\widehat{\Omega}_k^{-1})]}{\widehat{\mathsf{s.e.}}[\hat{\psi}_1]}\right]\right) \\
&\qquad \text{(see Figure S1)}.
\end{aligned} \tag{A.2}
$$

Both statistics depend on a data-adaptive estimator $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Omega}_k^{-1})$, which we next define. At a given k, $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Omega}_k^{-1})$ is equal to either $\widehat{\mathbb{IF}}_{22,k}([\widehat{\Omega}_k^{\mathsf{shrink}}]^{-1})$ or $\widehat{\mathbb{IF}}_{22,k}^{\mathsf{quasi}}([\widehat{\Omega}_k^{\mathsf{est}}]^{-1})$, defined as follows:

$$
\begin{aligned}
&\widehat{\mathbb{IF}}_{22,k}([\widehat{\Omega}_k^{\mathsf{shrink}}]^{-1}) \\
&:= \frac{(n-2)!}{n!} \\
&\qquad \times \sum_{1\le i_1\ne i_2\le n} [\hat{\varepsilon}_b\bar{\mathsf{Z}}_k]_{i_1}^\top[\widehat{\Omega}_k^{\mathsf{shrink}}]^{-1}[\bar{\mathsf{Z}}_k\hat{\varepsilon}_p]_{i_2},
\end{aligned} \tag{A.3}
$$

$$
\begin{aligned}
&\widehat{\mathbb{IF}}_{22,k}^{\mathsf{quasi}}([\widehat{\Omega}_k^{\mathsf{est}}]^{-1}) \\
&:= \frac{(n-2)!}{n!}
\end{aligned} \tag{A.4}
$$

$$\times \sum_{1\le i_1\ne i_2\le n} [\hat{\varepsilon}_b \bar{\mathsf{Z}}_k]_{i_1}^\top Q\big([\widehat{\Omega}_k^{\mathsf{est}}]^{-1}, \bar{\mathsf{Z}}_{k,i_1}, \bar{\mathsf{Z}}_{k,i_2}\big) \times [\bar{\mathsf{Z}}_k \hat{\varepsilon}_p]_{i_2},$$

where

$$Q\big([\widehat{\Omega}_k^{\mathsf{est}}]^{-1}, \bar{\mathsf{Z}}_{k,1}, \bar{\mathsf{Z}}_{k,2}\big) := [\widehat{\Omega}_k^{\mathsf{est}}]^{-1} + \frac{1}{n}[\widehat{\Omega}_k^{\mathsf{est}}]^{-1}(\bar{\mathsf{Z}}_{k,1}\bar{\mathsf{Z}}_{k,1}^\top + \bar{\mathsf{Z}}_{k,2}\bar{\mathsf{Z}}_{k,2}^\top)[\widehat{\Omega}_k^{\mathsf{est}}]^{-1},$$

$$\widehat{\Omega}_k^{\mathsf{est}} := \frac{1}{n}\sum_{i\in\mathsf{est}} \bar{\mathsf{Z}}_{k,i}\bar{\mathsf{Z}}_{k,i}^\top$$

and $\widehat{\Omega}_k^{\mathsf{shrink}}$ is the nonlinear shrinkage covariance matrix estimator developed in Ledoit and Wolf (2012), computed from the training sample. We briefly describe below how we choose between $\widehat{\mathbb{IF}}_{22,k}^{\mathsf{quasi}}([\widehat{\Omega}_k^{\mathsf{est}}]^{-1})$ and $\widehat{\mathbb{IF}}_{22,k}([\widehat{\Omega}_k^{\mathsf{shrink}}]^{-1})$. More details can be found in Section S3, Section S6 and Section S9. Their variance estimators are described in Remark S5.1 and Remark S5.2 respectively.

- In simulations, for every k, $\widehat{\mathbb{IF}}_{22,k}^{\mathsf{quasi}}([\widehat{\Omega}_k^{\mathsf{est}}]^{-1})$ is always numerically stable. We know that $\mathsf{Bias}_{\theta,k}(\hat{\psi}_1)$ increases with k. In contrast, although $\widehat{\mathbb{IF}}_{22,k}^{\mathsf{quasi}}([\widehat{\Omega}_k^{\mathsf{est}}]^{-1})$ initially increases with k, we observe that after some k^*, it begins to decrease. Our adaptive estimator switches to $\widehat{\mathbb{IF}}_{22,k^*}([\widehat{\Omega}_{k^*}^{\mathsf{shrink}}]^{-1}))$ at this k^*, if the variance estimator of $\widehat{\mathbb{IF}}_{22,k^*}([\widehat{\Omega}_k^{\mathsf{shrink}}]^{-1}))$ does not blow up. Empirically, $\widehat{\mathbb{IF}}_{22,k}([\widehat{\Omega}_k^{\mathsf{shrink}}]^{-1}))$ performs well as an estimator of $\mathsf{Bias}_{\theta,k}(\hat{\psi}_1)$ when its variance estimator does not blow up.
- In our simulation study, at each k, the empirical probability of either choosing $\widehat{\mathbb{IF}}_{22,k}^{\mathsf{quasi}}([\widehat{\Omega}_k^{\mathsf{est}}]^{-1})$ or $\widehat{\mathbb{IF}}_{22,k}([\widehat{\Omega}_k^{\mathsf{shrink}}]^{-1}))$ is 1. Thus we do not need to take into account the above data-driven selection step in estimating the variance of the data-adaptive estimator $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Omega}_k^{-1})$.

We leave the problem of unknown Ω_k^{-1} with $k > n$ to future work, because estimation of Ω_k^{-1} with $k > n$ requires additional assumptions on the distribution of X outside those in Condition W that may not hold.

ACKNOWLEDGMENTS

We would like to thank the editor Cun-Hui Zhang, the Associate Editor and the anonymous referee for their constructive comments which significantly improved our paper. We would also like to thank Thomas M. Kolokotrones (Harvard University), Weiming Li (Shanghai University of Finance and Economics), Thomas S. Richardson (University of Washington), Linbo Wang (University of Toronto) and Michael Wolf (ETH Zurich) for valuable discussions. Lin Liu and James M. Robins were supported by the U.S. Office of Naval Research Grant N000141912446. Rajarshi Mukherjee's research was partially supported by NSF Grant EAGER-1941419.

SUPPLEMENTARY MATERIAL

Supplement to "On Nearly Assumption-Free Tests of Nominal Confidence Interval Coverage for Causal Parameters Estimated by Machine Learning" (DOI: 10.1214/20-STS786SUPP; .pdf). In the Supplementary Materials (Liu, Mukherjee and Robins, 2020), we discuss estimators/tests when Ω_k^{-1} is unknown, other technical details, the details of the simulation reported in Table 1 and other simulation studies.

REFERENCES

AYYAGARI, R. (2010). Applications of influence functions to semiparametric regression models. Ph.D. thesis, Harvard Univ. MR2813909

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

BELLONI, A., CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *J. Econometrics* **186** 345–366. MR3343791 https://doi.org/10.1016/j.jeconom.2015.02.014

BHATTACHARYA, R. N. and GHOSH, J. K. (1992). A class of U-statistics and asymptotic normality of the number of k-clusters. *J. Multivariate Anal.* **43** 300–330. MR1193616 https://doi.org/10.1016/0047-259X(92)90038-H

BOUCHERON, S., LUGOSI, G. and MASSART, P. (2013). *Concentration Inequalities*: *A Nonasymptotic Theory of Independence*. Oxford Univ. Press, Oxford. MR3185193 https://doi.org/10.1093/acprof:oso/9780199535255.001.0001

BREIMAN, L. (2001). Random forests. *Mach. Learn.* **45** 5–32.

CHAKRABORTTY, A. and CAI, T. (2018). Efficient and adaptive linear regression in semi-supervised settings. *Ann. Statist.* **46** 1541–1572. MR3819109 https://doi.org/10.1214/17-AOS1594

CHAPELLE, O., SCHÖLKOPF, B. and ZIEN, A. (2010). *Semi-Supervised Learning*. *Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA.

CHERNOZHUKOV, V., CHETVERIKOV, D., DEMIRER, M., DUFLO, E., HANSEN, C., NEWEY, W. and ROBINS, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econom. J.* **21** C1–C68. MR3769544 https://doi.org/10.1111/ectj.12097

CORTES, C. and VAPNIK, V. (1995). Support-vector networks. *Mach. Learn.* **20** 273–297.

FREUND, Y. and SCHAPIRE, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. System Sci.* **55** 119–139. MR1473055 https://doi.org/10.1006/jcss.1997.1504

KRIZHEVSKY, A., SUTSKEVER, I. and HINTON, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* 1097–1105.

KUCHIBHOTLA, A. K. and CHAKRABORTTY, A. (2018). Moving beyond sub-Gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. Preprint. Available at arXiv:1804.02605.

LEDOIT, O. and WOLF, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Ann. Statist.* **40** 1024–1060. MR2985942 https://doi.org/10.1214/12-AOS989

LIU, L., MUKHERJEE, R. and ROBINS, J. (2020). Supplement to "On nearly assumption-free tests of confidence interval coverage of parameters estimated by machine learning." https://doi.org/10.1214/20-STS786SUPP

MUKHERJEE, R., NEWEY, W. K. and ROBINS, J. M. (2017). Semiparametric efficient empirical higher order influence function estimators. Preprint. Available at arXiv:1705.07577.

NEWEY, W. K. and ROBINS, J. M. (2018). Cross-fitting and fast remainder rates for semiparametric estimation. Preprint. Available at arXiv:1801.09138.

RITOV, Y., BICKEL, P. J., GAMST, A. C. and KLEIJN, B. J. K. (2014). The Bayesian analysis of complex, high-dimensional models: Can it be CODA? *Statist. Sci.* **29** 619–639. MR3300362 https://doi.org/10.1214/14-STS483

ROBINS, J. M. and ROTNITZKY, A. (2001). Comments on "Inference for semiparametric models: Some questions and an answer." *Statist. Sinica* **11** 920–936.

ROBINS, J., LI, L., TCHETGEN, E. and VAN DER VAART, A. (2008). Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and Statistics*: *Essays in Honor of David A. Freedman* 335–421. IMS, Beachwood, OH.

ROBINS, J., TCHETGEN TCHETGEN, E., LI, L. and VAN DER VAART, A. (2009). Semiparametric minimax rates. *Electron. J. Stat.* **3** 1305–1321. MR2566189 https://doi.org/10.1214/09-EJS479

ROBINS, J. M., ZHANG, P., AYYAGARI, R., LOGAN, R., TCHETGEN TCHETGEN, E., LI, L., LUMLEY, T. and VAN DER VAART, A. (2013). New statistical approaches to semiparametric regression with application to air pollution research. Research Report 175, Health Effects Institute.

ROBINS, J. M., LI, L., MUKHERJEE, R., TCHETGEN TCHETGEN, E. and VAN DER VAART, A. (2017). Minimax estimation of a functional on a structured high-dimensional model. *Ann. Statist.* **45** 1951–1987. MR3718158 https://doi.org/10.1214/16-AOS1515

ROSENBAUM, P. R. (2008). Testing hypotheses in order. *Biometrika* **95** 248–252. MR2409727 https://doi.org/10.1093/biomet/asm085

RUBIN, H. and VITALE, R. A. (1980). Asymptotic distribution of symmetric statistics. *Ann. Statist.* **8** 165–170. MR0557561

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

SCHARFSTEIN, D. O., ROTNITZKY, A. and ROBINS, J. M. (1999b). Rejoinder. *J. Amer. Statist. Assoc.* **94** 1135–1146.

SCHICK, A. (1986). On asymptotically efficient estimation in semiparametric models. *Ann. Statist.* **14** 1139–1151. MR0856811 https://doi.org/10.1214/aos/1176350055

TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. *Springer Series in Statistics*. Springer, New York. MR2724359 https://doi.org/10.1007/b13794

VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247 https://doi.org/10.1017/CBO9780511802256

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*: *With Applications to Statistics*. *Springer Series in Statistics*. Springer, New York. MR1385671 https://doi.org/10.1007/978-1-4757-2545-2

VERSHYNIN, R. (2018). *High-Dimensional Probability*: *An Introduction with Applications in Data Science*. *Cambridge Series in Statistical and Probabilistic Mathematics* **47**. Cambridge Univ. Press, Cambridge. MR3837109 https://doi.org/10.1017/9781108231596

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