ON ASSUMPTION-FREE TESTS AND CONFIDENCE INTERVALS FOR CAUSAL EFFECTS ESTIMATED BY MACHINE LEARNING

LIN LIU, RAJARSHI MUKHERJEE, AND JAMES M. ROBINS

ABSTRACT. For many causal effect parameters $\psi$ of interest doubly robust machine learning (DR-ML) (Chernozhukov et al., 2018a) estimators $\hat{\psi}_1$ are the state-of-the-art, incorporating the benefits of the low prediction error of machine learning (ML) algorithms; 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, when the vector of potential confounders is high dimensional, the associated $(1 - \alpha)$ Wald confidence intervals $\hat{\psi}_1 \pm z_{\alpha/2}\hat{\text{se}}[\hat{\psi}_1]$ may still undercover even in large samples, because the bias of the estimator may be of the same or even larger order than its standard error of order $n^{-1/2}$.

In this paper, we introduce novel tests that (i) can have the power to detect whether the bias of $\hat{\psi}_1$ is of the same or even larger order than its standard error of order $n^{-1/2}$, (ii) can provide a lower confidence limit on the degree of under coverage of the interval $\hat{\psi}_1 \pm z_{\alpha/2}\hat{\text{se}}[\hat{\psi}_1]$ and (iii) strikingly, are valid under essentially no assumptions whatsoever. We also introduce an estimator $\hat{\psi}_2 = \hat{\psi}_1 - \hat{\text{IF}}_{22}$ with bias generally less, and often much less, than that of $\hat{\psi}_1$, yet whose standard error is not much greater than $\hat{\psi}_1$’s. The tests, as well as the estimator $\hat{\psi}_2$, are based on a U-statistic $\hat{\text{IF}}_{22}$ that is the second-order influence function for the parameter that encodes the estimable part of the bias of $\hat{\psi}_1$. For the definition and theory of higher order influence functions see Robins et al. (2008, 2017). When the covariance matrix of the potential confounders is known, $\hat{\text{IF}}_{22}$ is an unbiased estimator of its parameter. When the covariance matrix is unknown, we propose several novel estimators of $\hat{\text{IF}}_{22}$ that perform almost as well as the known covariance case in simulation experiments.

Our impressive 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 can be consistent [without making additional assumptions (e.g. smoothness or sparsity) that may be incorrect]. Furthermore the above claims only apply to parameters in a particular class. For the others, our results are unavoidably less sharp and require more careful interpretation.

Keywords. Higher-order influence functions, Assumption-free inference, Confidence intervals, Valid inference

1. INTRODUCTION

Valid inference (i.e. valid tests and confidence intervals) for causal effects are of importance in many subject matter areas. For example, in medicine it is critical to evaluate whether a non-null 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 a necessary condition 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 – as a component – the building of non-causal 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 (usually) 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 $\hat{\psi}_1 \pm z_{\alpha/2}\hat{\text{se}}[\hat{\psi}_1]$ at sample size $n$ for a parameter $\psi$ 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 \( \hat{se}[\hat{\psi}_1] \). A nominal \((1 - \alpha)\) interval is said to be valid if the actual coverage rate under repeated sampling is greater than or equal to \((1 - \alpha)\). 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 clear that, in general, so-called machine learning algorithms (e.g. deep neural nets (Krizhevsky et al., 2012), support vector machines (Cortes and Vapnik, 1995), boosting (Freund and Schapire, 1997), regression trees and random forests (Breiman, 2001, 2017), etc., especially when combined with algorithm selection by cross-validation) can do a much better job of prediction than traditional parametric or non-parametric approaches (e.g. kernel or orthogonal 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 (Chernozhukov et al., 2018a) estimators have been developed that can be nearly unbiased for the causal effect \( \psi \), even when both of the above approaches fail. DR-ML estimators employ ML estimators of both the outcome regression \( b(x) \) and the propensity score \( p(x) \). DR-ML estimators are the current state of the art for estimation of causal effects, combining the benefits of sample splitting, machine learning, and double robustness (Scharfstein et al., 1999a,b; 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 machine learning estimators \( \hat{b}(x) \) and \( \hat{p}(x) \) of the outcome regression and the propensity score are fit using the training sample data. The estimator \( \hat{\psi}_1 \) of our causal parameter \( \psi \) of interest is computed from the estimation sample treating the machine learning regression 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 1.3 below, the efficiency lost due to sample splitting can be recovered by cross-fitting. The cross-fit estimator \( \hat{\psi}_{\text{cross-fit, 1}} \) 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 the need for Donsker conditions has a long history (Schick, 1986; van der Vaart, 1998, Page 391), although the idea of explicitly combining cross-fitting with machine learning 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 of papers that emphasized the theoretical and finite sample advantages of DR-ML estimators. See also Belloni et al. (2012).

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

An important point is that the aforementioned ML prediction algorithms have largely unknown statistical properties and, in that sense, can be viewed as statistical black boxes. The ML algorithms such as random forests (Athey et al., 2016; Wager and Athey, 2018) or neural networks (Farrell et al., 2018) for which statistical properties have been proved are either very simplified versions of the actual algorithms used in practice and/or the statistical properties are proved under restrictive modeling assumptions (such as smoothness or sparsity) that may not hold in practice.

The discussion above leads to the following questions: Can \( \omega \)-level tests be developed that have the ability to detect whether the bias of a DR-ML estimator \( \hat{\psi}_1 \) or \( \hat{\psi}_{\text{cross-fit, 1}} \) is of the same or greater order than its standard error? In particular, can we provide a lower confidence bound on the degree of under-coverage of a nominal \((1 - \alpha)\) confidence interval \( \hat{\psi}_1 \pm z_{\alpha/2} \hat{se}[\hat{\psi}_1] \). 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) requiring any assumptions at all (aside from a few standard, quite weak assumptions given later) - in particular, without making any assumptions about the smoothness or sparsity of the true outcome regression or propensity score function?

In this paper, we show that, perhaps surprisingly, the answer to these questions can be “yes” by using higher-order influence function tests and estimators introduced in Robins et al. (2008, 2017); Mukherjee et al. (2017).

However this claim needs to be tempered in several important ways. First the claims in the preceding paragraph only apply to parameters of interest that are in a particular class that we characterize in Section 2.1. For other parameters, our results have a similar flavor, but are unavoidably less sharp and require more careful interpretation as we discuss in Section 2.1. Second as explained in Remark 1.1, 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 DR-ML estimator is negligible compared to its standard error can be consistent [without making additional assumptions (e.g. smoothness or sparsity) that may be incorrect about the true but unknown propensity score and outcome regression functions]. Thus, when our \(\omega\)-level test rejects the null for \(\omega\) small, we can have strong evidence that the estimators \(\hat{\psi}_1\) and \(\hat{\psi}_{\text{cross-fit,1}}\) 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, because we make (essentially) no assumptions whatsoever we can never empirically rule out that the bias of \(\hat{\psi}_1\) and \(\hat{\psi}_{\text{cross-fit,1}}\) is as large as order 1 and thus \(n^{1/2}\) times greater than \(\text{se}[\hat{\psi}_1]\)!

Put another way, because we make essentially no assumptions, no methodology can (non-trivially) upper bound the bias of any estimator or lower bound the coverage of any confidence interval. What is novel about our methodology is that it can reject the claim that a particular DR-ML estimator \(\hat{\psi}_1\) has bias small relative to its standard error under the actual law generating the data (without claiming that the bias-corrected estimator \(\hat{\psi}_2\) has bias less than order 1.)

We now describe our approach at a high level. Throughout, we let \(A\) denote the treatment indicator, \(Y\) the outcome of interest, and \(X\) the vector of potential confounders with compact support. Again let \(\hat{\psi}_1\) and \(\psi_1 \pm z_{\alpha/2}\text{se}[\hat{\psi}_1]\) denote a DR-ML estimator of and associated \((1 - \alpha)\) Wald confidence interval for a particular parameter \(\psi\). In this paper, for didactic purposes only, we will choose \(\psi\) to be (components) of the so-called variance-weighted-average causal effect of a binary treatment \(A\) on a binary outcome \(Y\) given a vector \(X\) of confounding variables. However, 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 non-causal parameters. In fact, our results extend straightforwardly to the entire class of functionals with the mixed bias property of Rotnitzky et al. (2019a,b). This class of functionals strictly contain both the class of doubly robust functionals covered in Robins et al. (2008) and Chernozhukov et al. (2018b).

We assume that we have 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 that they outputted.

All DR-ML estimators are based on the first order influence function of the parameter \(\psi\) (van der Vaart, 1998, 2002). Our proposed approach begins by computing a second order influence function estimator \(\hat{\theta}_{22,k}\) of estimable part of the conditional bias \(\mathbb{E}[\hat{\psi}_1 - \psi | \bar{\theta}]\) of \(\hat{\psi}_1\) given the training sample data denoted by \(\bar{\theta}\). Our bias corrected estimator is \(\hat{\psi}_{2,k} \equiv \hat{\psi}_1 - \hat{\theta}_{22,k}\). The statistic \(\hat{\theta}_{22,k}\) is a second-order U-statistic that depends on a choice of \(k\) (with \(k = o(n^2)\) for reasons explained in Remark 2.7), a vector of basis functions \(Z_k \equiv z_k(X) \equiv (z_1(X), \ldots, z_k(X))\) of the high dimensional vector of potential confounders \(X\) and an estimator \(\hat{\Omega}_{k}^{-1}\) of the inverse expected outer product \(\Omega_{k}^{-1} \equiv \{\mathbb{E}[z_k(X)z_k(X)']\}^{-1}\). Both \(\hat{\psi}_{2,k}\) and \(\hat{\theta}_{22,k}\) will be asymptotically normal when, as in our asymptotic set-up, \(k = k(n) \to \infty\) and \(k = o(n^2)\) as \(n \to \infty\) (If \(k\) did not increase with \(n\), the asymptotic distribution of \(\hat{\theta}_{22,k}\) would be the so-called
Gaussian chaos distribution (Rubin and Vitale, 1980)). Furthermore, the variance of \( \hat{F}_{22,k} \) and \( \hat{\psi}_{2,k} \) are of order \( k/n^2 \) and \( 1/n + k/n^2 \) respectively.

The degree of the bias corrected by \( \hat{F}_{22,k} \) depends critically on (i) the choice of \( k \), (ii) the accuracy of the estimator \( \hat{\Omega}_k^{-1} \) of \( \Omega_k^{-1} \), and (iii) the particular \( k \)-vector of (basis) functions \( \hat{Z}_k \equiv \hat{z}_k(X) \) selected from a much larger, possibly countably infinite, dictionary of candidate functions. Data adaptive choices of \( k \) and \( \hat{\Omega}_k^{-1} \) are discussed in Appendix I. We have developed several data-adaptive methods to choose \( \hat{Z}_k \) that will be the subject of a separate paper.

Occasionally one 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 \( s \) 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 \( \Omega_k \) by the empirical covariance matrix based on all subjects; and then treat \( \Omega_k \) and \( \hat{\Omega}_k^{-1} \) as known in an analysis based on the \( n \) subjects with complete data (Chapelle et al., 2010; Chakrabortty and Cai, 2018). Thus with \( X \)-semisupervised data, the discussion of the estimators \( \hat{\Omega}_k^{-1} \) of Section 3 is irrelevant.

For further motivation and before going into technical details, we now summarize some finite sample results from simulation studies that are described in full detail in Appendix K. Depending on the study, we either simulated 200 or 1000 estimation samples each with sample size \( n = 2500 \). The same training sample, also of size 2500, and thus the same estimates of the regression functions were used in each simulation study. Thus the results are conditional on that training sample. However, additional unconditional simulation studies wherein we simulated both 200 estimation and training samples reported in Table 11 and Table 12 of Appendix K.3 produced similar results. We took \( k \) to be significantly less than the sample size \( n \) for the following three reasons: \( k < n \) is necessary i) for bias-corrected confidence intervals (CI) centered at \( \hat{\psi}_{2,k} \equiv \hat{\psi}_1 - \hat{F}_{22,k} \) to have length approximately equal to CIs centered at \( \hat{\psi}_1 \), ii) for \( \hat{F}_{22,k} \)’s standard error of order \( k^{1/2}/n \) to be smaller than the order \( n^{-1/2} \) of the standard error of \( \hat{\psi}_1 \), thereby creating the possibility of detecting, for any given \( \delta > 0 \), that the (absolute value) of the ratio of the bias of \( \hat{\psi}_1 \) to its standard error exceeds \( \delta \), provided the sample size \( n \) is sufficiently large and iii) to be able to estimate \( \Omega_k^{-1} \) accurately without imposing the additional (possibility incorrect) smoothness or sparsity assumptions required for accurate estimation when \( k \geq n \).

In all our simulation studies we chose a class of data generating processes 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 to insure that the DR-ML estimator \( \hat{\psi}_1 \) had significant asymptotic bias. Furthermore, in most of our simulation studies, 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 nearly as well as any other DR-ML estimators. The basis functions \( \hat{z}_k(x) \) were chosen to be particular Daubechies wavelets that Robins et al. (2008, 2009, 2017) showed to be minimax optimal for estimation of \( \psi \) by \( \hat{\psi}_{2,k} \) for the chosen smoothness classes. Thus, in our simulations, we used optimal versions of both \( \hat{\psi}_1 \) and \( \hat{\psi}_{2,k} \) to ensure a fair comparison. [Out of interest, in Table 13 and Table 14 of Appendix K.3, we also report additional simulation results in which the propensity score and outcome regression were estimated with convolutional neural networks (Farrell et al., 2018). Qualitatively similar results were obtained.]

Table 1 reports results from one of the simulation studies to demonstrate the empirical behavior of \( \hat{F}_{22,k} \) and \( \hat{\psi}_{2,k} \). We examined the empirical behavior of our data adaptive estimator \( \hat{\Omega}_k^{-1} \) of \( \Omega_k^{-1} \) as \( k \) varies by comparing the estimators \( \hat{F}_{22,k}(\hat{\Omega}_k^{-1}) \) and \( \hat{\psi}_{2,k}(\hat{\Omega}_k^{-1}) \) that use \( \hat{\Omega}_k^{-1} \) to the oracle estimators \( \hat{F}_{22,k}(\Omega_k^{-1}) \) and \( \hat{\psi}_{2,k}(\Omega_k^{-1}) \) that use the true inverse covariance matrix \( \Omega_k^{-1} \). The target parameter \( \psi \) of this simulation study is in the class for which our results are sharpest as explained later in Section 2.1.

Note the 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, \( \hat{F}_{22,k=0}(\Omega_k^{-1}) \) and \( \hat{F}_{22,k=0}(\Omega_k^{-1}) \) are zero. As seen in row 1, column 2 of Table 1, nominal 90 % Wald confidence intervals centered at \( \hat{\psi}_1 = \hat{\psi}_{2,k=0} \) had empirical coverage of 0% in 100
Table 1.

| \( k \) | \( \hat{\mathbb{P}}_{22, k}(\Omega_k^{-1}) \) | MC Coverage | \( \hat{\psi}_{2, k}(\Omega_k^{-1}) \) 90% Wald CI | Bias(\( \hat{\psi}_{2, k}(\Omega_k^{-1}) \)) | \( \hat{\psi}_{2, k}(\Omega_k^{-1}) \) |MC Coverage |
|---|---|---|---|---|---|---|
| 0 | 0 (0) | 0 | 0.197 (0.034) | 0 | 0 | 0.191 (0.034) | 0 |
| 8 | 0.0063 (0.0034) | 0 | 0.191 (0.034) | 0 | 0.0063 (0.0034) | 0 | 0.191 (0.034) | 0 |
| 256 | 0.081 (0.013) | 0.034 | 0.116 (0.036) | 96.5% | 0.081 (0.013) | 0.094 (0.012) | 0.116 (0.036) | 96.5% |
| 512 | 0.094 (0.023) | 0.146 | 0.103 (0.041) | 97.4% | 0.094 (0.023) | 0.102 (0.021) | 0.103 (0.040) | 98.9% |
| 1024 | 0.150 (0.037) | 0.157 | 0.147 (0.050) | 100% | 0.150 (0.037) | 0.149 (0.032) | 0.147 (0.046) | 99.2% |
| 2048 | 0.191 (0.062) | 0.988 | 0.006 (0.071) | 99.9% | 0.191 (0.062) | 0.991 (0.057) | 0.006 (0.070) | 96.6% |

We reported the MCav of point estimates and standard errors (first column in each panel) of \( \hat{\mathbb{P}}_{22, k}(\Omega_k^{-1}) \) and \( \hat{\mathbb{P}}_{22, k}(\Omega_k^{-1}) \), together with the coverage probability of 90% confidence intervals (second column in each panel) of \( \hat{\psi}_{2, k}(\Omega_k^{-1}) \) and \( \hat{\psi}_{2, k}(\Omega_k^{-1}) \), the MCav of the bias and standard errors (third column in each panel) of \( \hat{\psi}_{2, k}(\Omega_k^{-1}) \) and \( \hat{\psi}_{2, k}(\Omega_k^{-1}) \) and the empirical rejection rate based on \( \hat{\chi}_k(\Omega_k^{-1}; \zeta, \delta = 3/4) \) and \( \hat{\chi}_k(\Omega_k^{-1}; \zeta, \delta = 3/4) \) with \( \zeta = z_{\omega=0.10} = 1.28 \) (fourth column in each panel).

For complete details, see Appendix K.1.

Simulations! However, as seen in column 2 of both the left and right panels of the last row, 90% Wald confidence intervals centered at either \( \hat{\psi}_{2, k}(\Omega_k^{-1}) \) or \( \hat{\psi}_{2, k}(\Omega_k^{-1}) \) at \( k = 2048 \) had empirical coverage exceeding 90%, even though their standard errors did not greatly exceed that of \( \hat{\psi}_1 \).

In more detail, the left panel of Table 1 displays the Monte Carlo average (MCav) of the point estimates and estimated standard errors (in parentheses) of \( \hat{\mathbb{P}}_{22, k}(\Omega_k^{-1}) \) in the first column; the empirical probability that a nominal 90% Wald confidence interval centered at \( \hat{\psi}_{2, k}(\Omega_k^{-1}) \) covered the true parameter value in the second column; the MC bias (i.e., MCav of \( \hat{\psi}_{2, k} - \psi \)) and MCav of the estimated standard errors of \( \hat{\psi}_{2, k}(\Omega_k^{-1}) \) in the third column; and, in the fourth column, the empirical rejection rate of a one-sided \( \omega = 0.10 \) level test \( \hat{\chi}_k(\Omega_k^{-1}; z_{\omega=0.10}, \delta = 3/4) \) (defined in eq. (2.6) of Section 2) of the null hypothesis that the bias of \( \hat{\psi}_1 \) is smaller than \( \delta = 3/4 \) of its standard error. The test rejects when the ratio \( \hat{\mathbb{P}}_{22, k}(\Omega_k^{-1})/\hat{\mathbb{S}}[\hat{\psi}_1] \) is large. Similarly, the right panel displays these same summary statistics but with the data adaptive estimator \( \tilde{\Omega}_k^{-1} \) in place of \( \Omega_k^{-1} \). Thus, for example the difference between the MC bias of \( \hat{\psi}_{2, k}(\tilde{\Omega}_k^{-1}) \) and \( \hat{\psi}_{2, k}(\Omega_k^{-1}) \) is an estimate of the additional bias due to the estimation of \( \Omega_k^{-1} \) by \( \tilde{\Omega}_k^{-1} \). (The uncertainty in the estimate of the bias itself is not given in the table but it is negligible as it approximately equals \((1/1000)^{1/2} \times \) 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 0.197. The MC bias of \( \hat{\psi}_{2, k}(\Omega_k^{-1}) \) and \( \hat{\psi}_{2, k}(\tilde{\Omega}_k^{-1}) \) decreased with increasing \( k \), becoming nearly zero at \( k = 2048 \). The observation that the bias decreases as \( k \) increases is predicted by the theory developed in Section 2. The decrease in bias reflects the increase in the absolute value of \( \hat{\mathbb{P}}_{22, k}(\Omega_k^{-1}) \) with \( k \). Note further that both the MCavs of \( \hat{\mathbb{P}}_{22, k}(\Omega_k^{-1}) \) and \( \hat{\mathbb{P}}_{22, k}(\tilde{\Omega}_k^{-1}) \) are quite close, as are the MCavs of their estimated standard errors implying that our estimator \( \tilde{\Omega}_k^{-1} \) performs similarly to the true \( \Omega_k^{-1} \). The actual coverages of 90% Wald confidence intervals centered at \( \hat{\psi}_{2, k}(\Omega_k^{-1}) \) and \( \hat{\psi}_{2, k}(\tilde{\Omega}_k^{-1}) \) both increase from 0% at \( k = 0 \) to over 95% at \( k = 2048 \). From simulation experiments, we show in Appendix H that the conservative coverage rate at \( k = 2048 \) is due to the estimated standard errors of \( \hat{\mathbb{P}}_{22, k}(\Omega_k^{-1}) \) and \( \hat{\mathbb{P}}_{22, k}(\tilde{\Omega}_k^{-1}) \) being somewhat upwardly biased. In Appendix H.1, we demonstrate that bootstrap resampling method can provide better variance estimator empirically, with additional computational cost by computing \( B \gg 1 \) bootstrapped \( \hat{\mathbb{P}}_{22, k}(\Omega_k^{-1}) \). Also, reading from the third column, we see that the standard error (0.070) of \( \hat{\psi}_{2, k=2048}(\tilde{\Omega}_{k=2048}^{-1}) \) is only 2 times the standard error (0.034) of \( \hat{\psi}_1 \), confirming that the dramatic difference in coverage rates of their associated confidence intervals 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 \( \hat{\chi}_{k}(\Omega_k^{-1}; z_{\omega=0.10}, \delta = 3/4) \) and \( \hat{\chi}_{k}(\tilde{\Omega}_k^{-1}; z_{\omega=0.10}, \delta = 3/4) \) are already above 96% when \( k = 256 \), indicating that the bias of \( \hat{\psi}_1 \) is much greater than 3/4 of its standard error. Indeed, reading from row 1 of column 3, we see that the ratio of the MCav 0.197 of the bias of \( \hat{\psi}_1 = \hat{\psi}_{2, k=0} \) to the MCav 0.034 of its estimated standard error is nearly 6! In Remark 2.3 of Section 2, we show that this ratio is close to that predicted by theory.
Figure 3 of Appendix L.1 provides a histogram over the 1000 estimation samples of \( (1 - \omega) \) upper confidence bounds \( UCB(1)(\Omega_k^{-1};\alpha,\omega) \) (defined in eq. (2.9) of Section 2) and \( UCB(1)(\tilde{\Omega}_k^{-1};\alpha,\omega) \) (defined in Section 3) for the actual conditional asymptotic coverage of the nominal \((1 - \alpha)\) interval \( \hat{\psi}_1 \pm z_{\alpha/2}\hat{\text{se}}[\hat{\psi}_1] \). To clarify the meaning of \( UCB(1)(\Omega_k^{-1};\alpha,\omega) \), define the conditional actual coverage \( \text{cactcov}(\alpha) \), given the training sample to be

\[
\text{cactcov}(\alpha) = P(\psi \in \{\hat{\psi}_1 \pm z_{\alpha/2}\hat{\text{se}}[\hat{\psi}_1]\}|\hat{\theta}).
\]

Then, by definition, a \((1 - \omega)\) conditional upper confidence bound \( UCB(1)(\Omega_k^{-1};\alpha,\omega) \) is a random variable satisfying

\[
P\left\{ \text{cactcov} \leq UCB(1)(\Omega_k^{-1};\alpha,\omega)|\hat{\theta}\right\} \geq 1 - \omega
\]

For example if \( UCB(1)(\Omega_k^{-1};\alpha,\omega) = 0.14 \) for \( \omega = 0.10, \alpha = 0.10 \), then the actual coverage of the nominal 90% interval \( \hat{\psi}_1 \pm 1.64\hat{\text{se}}[\hat{\psi}_1] \) is no more than 14% with confidence at least 1 - \( \omega = 0.90 \). More precisely, the random interval \([0, UCB(1)(\Omega_k^{-1};\alpha = 0.10, \omega = 0.10)]\) is guaranteed to include the actual coverage of \( \hat{\psi}_1 \pm 1.64\hat{\text{se}}[\hat{\psi}_1] \) at least 90% of the time in repeated sampling of the estimation sample with the training sample fixed. Recall from row 1, column 2 of the right panel of Table 1, that the actual Monte Carlo coverage of the nominal 90% interval \( \hat{\psi}_1 \pm 1.64\hat{\text{se}}[\hat{\psi}_1] \) was 0%. Hence our nominal 90% upper confidence bounds \( UCB(1)(\Omega_k^{-1};\alpha,\omega) \) and \( UCB(1)(\tilde{\Omega}_k^{-1};\alpha,\omega) \) are trivially conservative. More interestingly, they were both very close to 0% (more precisely, less than 1%) in more than 15% of the 1000 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 \( \hat{\psi} \), the state of the art DR-ML estimators, and the statistical properties of these estimators.

In Section 2, we study the theoretical statistical properties of the oracle estimator \( \hat{\Omega}_{22,k}(\Omega_k^{-1}) \) of the bias of the DR-ML estimator \( \hat{\psi}_1 \), the biased-corrected estimator \( \hat{\psi}_{2,k}(\Omega_k^{-1}) \), the \( \omega \)-level oracle test \( \hat{\chi}_{k}^{(1)}(\Omega_k^{-1};z_\omega,\delta) \) and the \((1 - \omega)\) upper confidence bound \( UCB(1)(\Omega_k^{-1};\alpha,\omega) \) for the actual conditional asymptotic coverage of the nominal \((1 - \alpha)\) Wald interval associated with \( \hat{\psi}_1 \).

In Section 3, we define several estimators \( \tilde{\Omega}_k^{-1} \) of \( \Omega_k^{-1} \) and, for each, compare both the finite sample properties (through simulation) and aspects of the theoretical asymptotic behavior of \( \hat{\Omega}_{22,k}(\tilde{\Omega}_k^{-1}) \) to those of the oracle \( \hat{\Omega}_{22,k}(\Omega_k^{-1}) \). To choose among these candidate estimators, we construct a data-adaptive estimator \( \hat{\Omega}_{22,k}^{\text{adapt}} \) (see Appendix I) where the candidate choice depends on both \( k \) and the data. \( \hat{\Omega}_{22,k}^{\text{adapt}} \) was the estimator \( \hat{\Omega}_{22,k}(\tilde{\Omega}_k^{-1}) \) used in the simulation study reported in the right panel in Table 1.

Although motivated by asymptotic results, our data-adaptive estimator \( \hat{\Omega}_{22,k}^{\text{adapt}} \) was ultimately chosen based on its finite sample performance in simulations. This choice reflects the fact that for certain estimators \( \hat{\Omega}_{22,k}(\tilde{\Omega}_k^{-1}) \) with known theoretical properties, we found that, at values of \( k \) required by the oracle \( \hat{\psi}_{2,k}(\Omega_k^{-1}) \) to correct most of the bias of \( \hat{\psi}_1 \), the Monte Carlo variance and mean of \( \hat{\Omega}_{22,k}(\tilde{\Omega}_k^{-1}) \) exploded, becoming orders of magnitude larger than that predicted by the asymptotics. (A proposed explanation for this discrepancy is given in Appendix B.) For this reason, it is likely that future theoretical development and simulation experiments may lead us to somewhat modify \( \hat{\Omega}_{22,k}^{\text{adapt}} \).

Section 4 considers the case where \( k \), rather than being less than \( n \), includes \( k > n \) but with \( k = o(n^2) \). We leave the unknown \( \Omega_k^{-1} \) case to a separate paper, because estimation of \( \Omega_k^{-1} \) with \( k > n \) requires additional assumptions that may not hold. The motivation to study the \( k \geq n \) case is as follows. As discussed above, the estimator \( \hat{\psi}_{2,k}(\Omega_k^{-1}) = \hat{\psi}_1 - \hat{\Omega}_{22,k}(\Omega_k^{-1}) \) with \( k \) less than but near \( n \) has standard error not much larger than the standard error of \( \hat{\psi}_1 \), but can have much smaller bias. This suggests foregoing the estimation of an upper bound on the actual coverage of a nominal \((1 - \alpha)\) Wald confidence interval centered on \( \hat{\psi}_1 \); rather always report, when \( \Omega_k^{-1} \) is known, the nominal \((1 - \alpha)\) Wald confidence interval

\[
\hat{\psi}_{2,k}(\Omega_k^{-1}) \pm z_{\alpha/2}\hat{\text{se}}[\hat{\psi}_{2,k}(\Omega_k^{-1})]
\]

for \( k = k_0 \) with \( k_0 \) just less than \( n \). However doing so naturally raises
the question of whether the interval \( \hat{\psi}_{2,k}({\Omega}^{-1}_k) \pm z_{\alpha/2} \delta \) with \( k = k_0 \) covers \( \psi \) at its nominal 1 - \( \alpha \) rate. Like for \( \hat{\psi}_1 \), we examine this by testing the null hypothesis that the ratio of the bias of \( \hat{\psi}_{2,k} \) to its standard error is smaller than a fraction \( \delta \). In Section 4.1, we construct such a test for any \( k = o(n^2) \). In addition, we develop a sequential multiple testing procedure that tests this null hypothesis at level \( \alpha \) for each of \( J \) different \( k \) with \( k_0 < n < k_1 < \cdots < k_{j-1} < n^2 \) with \( k_{j-1} = o(n^2) \). Our procedure tests this null hypothesis sequentially beginning with \( k_0 \) and stops at the first \( k_j \) for which the test does not reject and then accepts the null for \( k_{j+1}, \ldots, k_{j-1} \). The sequential procedure protects the level of all \( J \) tests (Rosenbaum, 2008). (Since all of our tests are inconsistent for reasons described earlier, “accepting the null” simply means protecting the level of the test if, in fact, the null were indeed true.)

In Section 5, we will conclude by discussing several open problems.

The following common asymptotic notations are used throughout the paper: \( x \lesssim y \) (equivalently \( x = O(y) \)) denotes that there exists some constant \( C > 0 \) such that \( x \leq C y \). \( x \gg y \) 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 \( \lim_{x,y \rightarrow \infty} \frac{x}{y} = 0 \). 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 \( \lim_{n \rightarrow \infty} P(|X_n/a_n| \geq \epsilon) = 0 \) for every positive real number \( \epsilon \).

### 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 (defined below) for a binary treatment \( A \) and binary outcome \( Y \) given a vector \( X \) of baseline covariates. To perform inference, we shall have access to \( N \) samples from the joint distribution of \((Y, A, X)\). We parametrize the joint distribution \( P_\theta \) of \((Y, A, X)\) by the variation independent parameters \( \theta \equiv (b, p, f_X, OR_{Y|A=X=x}) \), 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 \( OR_{Y|A=X=x} \) is the conditional odds ratio. Often, \( p(X) \) is referred to as the propensity score. Throughout the paper, we use \( \mathbb{E}_\theta \), \( \text{var}_\theta \) and \( \text{cov}_\theta \) with subscript \( \theta \) to indicate the expectation, variance, and covariance over the probability measure \( P_\theta \) indexed by \( \theta \). We assume a non-parametric infinite dimensional model \( \mathcal{M}(\Theta) := \{P_\theta, \theta \in \Theta \} \) where \( \Theta \) indexes all possible \( \theta \) subject to weak regularity conditions given later.

Under the assumption that the vector \( X \) of measured covariates suffices to control confounding, the variance-weighted average treatment effect \( \tau(\theta) \) is identified as \( \tau(\theta) := \mathbb{E}_{\theta}[\gamma_\theta(X) \text{var}_\theta(A|X)]/\mathbb{E}_{\theta}[\text{var}_\theta(A|X)] \) 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 \( \text{var}_\theta(A|X) = p(X)(1-p(X)) \).

Some algebra shows that

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

Henceforth, we shall restrict attention to the estimation of

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

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

We shall see that the statistical guarantees of our bias correction methodology differ depending on whether the parameter of interest is \( \mathbb{E}_{\theta}[\text{cov}_\theta(Y,A|X)] \) versus \( \mathbb{E}_{\theta}[\text{var}_\theta(A|X)] \). 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 \( DR-ML \) estimator \( \hat{\psi}_1 \) and \( \hat{\psi}_{\text{cross-fit}} \). 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 1.1. The methods in Robins et al. (2009) and Ritov et al. (2014) can be straightforwardly combined to show that, without further unverifiable assumptions, for some $\sigma > 0$, no consistent $\alpha$-level test of the null hypothesis $\mathbb{E}_\theta[\text{cov}_\theta(Y,A|X)] = 0$ versus the alternative $\mathbb{E}_\theta[\text{cov}_\theta(Y,A|X)] = \sigma$ exists, whenever some components of $X$ have a continuous distribution. See Remark 4.1 for a heuristic non-technical explanation. An analogous result holds for $\mathbb{E}_\theta[\text{var}_\theta(A|X)]$.

1.2. State-of-the-art estimators $\hat{\psi}_1$ and $\hat{\psi}_{\text{cross-fit},1}$ and their asymptotic properties. The state-of-the-art DR-ML 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-fit estimator $\hat{\psi}_{\text{cross-fit},1}$ is a DR-ML estimator that can recover the information lost by $\hat{\psi}_1$ due to sample splitting, provided that $\hat{\psi}_1$ is asymptotically unbiased.

The following algorithm defines $\hat{\psi}_1$ and $\hat{\psi}_{\text{cross-fit},1}$:

(i) The $N$ study subjects are randomly split into 2 parts: an estimation sample of size $n$ and a training (nuisance) sample of size $n_{tr} = N - n$ 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

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

from $n$ subjects in the estimation sample and

$$\hat{\psi}_{\text{cross-fit},1} = \left( \hat{\psi}_1 + \overline{\psi} \right) / 2$$

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

1.3. Asymptotic properties of $\hat{\psi}_1$ and $\hat{\psi}_{\text{cross-fit},1}$. The following theorem (Theorem 1.2) gives the asymptotic properties of $\hat{\psi}_1$, conditional on the training sample.

**Theorem 1.2.** Conditional on the training sample, $\hat{\psi}_1$ is asymptotically normal with conditional bias

$$cBias_\theta(\hat{\psi}_1) := \mathbb{E}_\theta \left[ \hat{\psi}_1 - \psi(\theta) | \hat{\theta} \right] = \mathbb{E}_\theta \left\{ \left\{ b(X) - \hat{b}(X) \right\} \{ p(X) - \hat{p}(X) \} \right\} | \hat{\theta}$$

where $\hat{\theta}$ to the right of the conditioning bar indicates conditioning on the training sample through the estimated components $(\hat{b}, \hat{p})$ of $\theta$.

**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 $cBias_\theta(\hat{\psi}_1)$ is the conditional bias. \hfill $\square$

We note that $\hat{\psi}_1$ is, by definition, doubly robust (Bang and Robins, 2005) because $cBias_\theta(\hat{\psi}_1) = 0$ if either $b(X) = \hat{b}(X)$ or $p(X) = \hat{p}(X)$ with $P_\theta$-probability 1. Finally, before proceeding, we summarize unconditional statistical properties of the DR-ML estimator in the following theorem (Theorem 1.3).

Recall that $cBias_\theta(\hat{\psi}_1)$ is random only through its dependence on the training sample data via $\hat{b}$ and $\hat{p}$.

**Theorem 1.3.** If a) $cBias_\theta(\hat{\psi}_1)$ is $o_{P_\theta}(n^{-1/2})$ and b) $\hat{b}(x)$ and $\hat{p}(x)$ converge to $b(x)$ and $p(x)$ in $L_2(P_\theta)$, then
Remark 1.5. Nonetheless, $M$

Remark 1.4. We could have divided the whole sample of size $N$ into $M \geq 2$ random samples, say 5, with $\hat{\psi}_1^{(m)}$, $m = 1, 2, \ldots, M$, calculated using sample $m$ as estimation sample and the remaining $M - 1$ samples as training sample. Then $\hat{\psi}_{\text{cross-fit,1}} = \frac{1}{M} \sum_{m=1}^{M} \hat{\psi}_1^{(m)}$. When, for each $m$, $c\text{Bias}^{(m)}_{\theta}(\hat{\psi}_1)$ is of smaller order than its standard error, the asymptotic distribution of $\hat{\psi}_{\text{cross-fit,1}}$ does not depend on $M$. Nonetheless, $M = 5$ or 10 generally performs better than $M = 2$ in finite samples. We restrict to $M = 2$ for notational convenience only.

Remark 1.5. Had we chosen the average treatment effect $\mathbb{E}_\theta[\mathbb{E}_a[Y|A = 1, X] - \mathbb{E}_a[Y|A = 0, X]]$ rather than the variance weighted average treatment effect as our parameter of interest, the outcome regression function appearing in the first order influence function would be $\mathbb{E}_\theta[Y|A = a, X]$ for $a = 0, 1$ rather than $\mathbb{E}_\theta[Y|X]$.

Remark 1.6 (Training sample squared error loss cross-validation). How can we choose among the many (say, J) available machine learning algorithms if our goal is to minimize the conditional mean squared error $\mathbb{E}_\theta \left[ (b(X) - \hat{b}(X))^2 \right]$? 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 $n_v/S$. For each subsample $s$, we fit the J ML algorithms to the other $S - 1$ subsamples to obtain outputs $\hat{b}_s^{(j)}(\cdot)$, for $j = 1, \ldots, J$. Next we calculate, for each $j$, the squared error loss $CV^{(j)} = \sum_{s=1}^{S} CV_s^{(j)}$ with $CV_s^{(j)} = \sum_{i \in s} \left( Y_i - \hat{b}_s^{(j)}(X_i) \right)^2$, and finally select the ML algorithm $\hat{j}_{\text{select}}$ that minimizes $CV^{(j)}$.

Although a standard result, Theorem 1.3 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 reject either the hypothesis $c\text{Bias}_\theta(\hat{\psi}_1) = O_P(n^{-1/2})$ or the hypothesis that $n^{1/2}(\hat{\psi}_1 - \psi(\theta))$ is asymptotically normal with mean zero. Rather, as discussed in Section 1, our interest, instead, lies in testing and rejecting hypotheses such as, at the actual sample size $n$, the actual asymptotic coverage of the interval $\hat{\psi}_1 \pm z_{\alpha/2} \hat{SE}(\hat{\psi}_1)$, conditional on the training sample, is less than a fraction $\varrho < 1$ of its nominal coverage.
Second, since the ML estimators $\hat{b}(x)$ and $\hat{p}(x)$ have unknown statistical properties and we are interested in essentially assumption-free inference, our inferential statements regard the training sample as fixed rather than random. In particular (with the exception of Section 3), the only randomness referred to in any theorem is that of the estimation sample. In fact, our inferences rely on being in ‘asymptopia’, only to be able to posit that, at our sample size of $n$, the quantiles of the finite sample distribution of a conditionally asymptotically normal statistic (e.g. $\hat{F}_{22,k}(\Omega^{-1}_k)$) are close to the quantiles of a normal. Indeed, by using finite sample tail inequalities, our reliance on asymptotics could be totally eliminated at the expense of decreased power and increased confidence interval width.

**Remark 1.7.** In Appendix K.3, we report results of an unconditional simulation study, (i.e. both the training and estimation sample were redrawn 200 times). The study gave similar results to ones in which the initial training sample was held fixed and only the estimation sample redrawn. This similarity only serves as evidence that the initial training sample was not an outlier. It provides no evidence concerning the order in probability of the conditional bias $cBias_\theta(\hat{\psi}_1)$.

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

**Notations.** For a (random) vector $V$ $\|V\|_\theta = \mathbb{E}_\theta[|V^T V|^\theta]^{1/2}$ denotes its $L_2(P_\theta)$ norm conditioning on the training sample, $\|V\|_2 \equiv (\sum V^T V)^{1/2}$ denotes its $\ell_2$ norm and $\|V\|_\infty$ denotes its $L_\infty$ norm. For any matrix $A$, $\|A\|$ will be used for its operator norm. Given a $k$, the random vector $\bar{Z}_k = \bar{z}_k(X)$, $\Pi [\cdot | \bar{Z}_k]$ denotes the population linear projection operator onto the space spanned by $\bar{Z}_k$ conditioning on the training sample: with $\Omega_k := \mathbb{E}_\theta[\bar{Z}_k \bar{Z}_k^T | \tilde{\theta}]$, $\Pi [\cdot | \bar{Z}_k] = I - \Pi [\cdot | \bar{Z}_k]$ is the projection onto the orthogonal complement of $\bar{Z}_k$ in the Hilbert space $L_2(f_X)$. Hence, for a random variable $V = v(X)$, 
$$
\Pi [V | \bar{Z}_k] = \bar{Z}_k^T \Omega_k^{-1} \mathbb{E}_\theta[\bar{Z}_k V | \tilde{\theta}], \quad \Pi [V | \bar{Z}_k^\perp] = V - \Pi [V | \bar{Z}_k].
$$

By defining $\Omega_k$ conditionally, we can allow the vector $\bar{Z}_k$ to depend on the training sample data. $\hat{\Omega}_k$ denotes a generic estimator of $\Omega_k^{-1}$. When referring to a particular estimator of $\Omega_k^{-1}$, an identifying superscript will often be attached.

We also denote the following commonly used residuals as $\hat{e}_{b,i} := y_i - \hat{b}(x_i)$, $\hat{e}_{p,i} := A_i - \hat{p}(x_i)$, $\hat{e}_{b,i} := b(x_i) - \hat{b}(x_i)$, and $\tilde{e}_{p,i} := p(x_i) - \hat{p}(x_i)$ for $i = 1, 2, \ldots, n$, where $\hat{b}$ and $\hat{p}$ are estimated from the training sample.

If $\bar{Z}_{k_1}$ and $\bar{Z}_{k_2}$ 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 $\bar{Z}_{k_1}$ is a subspace of the space spanned by $\bar{Z}_{k_2}$.

**Remark 1.8.** For example, when choosing the basis functions $\bar{Z}_k$ from a dictionary $\mathcal{V}$ of (candidate) functions greedily, **Condition B** holds.

2. The Projected Conditional Bias and an Oracle Test and Estimator

It is useful to first consider an oracle procedure that would only be implementable if $\Omega_k^{-1}$ were known. To this end 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 the countable set of all $m$-th powers and $m$-way interactions of $X$ for $m = 1, 2, \ldots$. Alternatively, if we restrict to $m < m_{\text{max}}$, then $\mathcal{V}$ will be finite. We could also replace monomial bases with tensor products of Fourier, spline or wavelet bases (or the union of all three types) in defining $\mathcal{V}$.

We decompose $b(X) - \hat{b}(X) = \Pi[b(X) - \hat{b}(X)|\bar{Z}_k] + \Pi[b(X) - \hat{b}(X)|\bar{Z}_k^\perp]$, 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 $\bar{Z}_k$ and the second term is the projection onto the orthocomplement $\bar{Z}_k^\perp$. Subsequently, the Pythagorean theorem allows us to decompose

$$
cBias_\theta(\hat{\psi}_1) = cBias_{\theta,k}(\hat{\psi}_1) + cTB_{\theta,k}(\hat{\psi}_1)
$$
into the sum of two terms \( cBias_{\theta,k}(\tilde{\psi}_1) \) and \( cTB_{\theta,k}(\tilde{\psi}_1) \) where

\[
\begin{align*}
  cBias_{\theta,k}(\tilde{\psi}_1) &= \mathbb{E}_\theta \left\{ \Pi [b(X) - \tilde{b}(X)] | Z_k \} \right\} \{\Pi [p(X) - \tilde{p}(X)|Z_k] \} | \theta \}, \\
  cTB_{\theta,k}(\tilde{\psi}_1) &= \mathbb{E}_\theta \left\{ \Pi [b(X) - \tilde{b}(X)] | Z_k^k \} \right\} \{\Pi [p(X) - \tilde{p}(X)|Z] \} | \theta \}.
\end{align*}
\]  

(2.2)

It follows from Remark 1.1 above that in the absence of further assumptions, \( cTB_{\theta,k}(\tilde{\psi}_1) \) could be of order 1 and cannot be consistently estimated. However, since one can also write \( cBias_{\theta,k}(\tilde{\psi}_1) \) as

\[
cBias_{\theta,k}(\tilde{\psi}_1) = \mathbb{E}_\theta \left\{ b(X) - \tilde{b}(X) \right\} \Omega_{k}^{-1} \mathbb{E}_\theta \left\{ Z_k (p(X) - \tilde{p}(X)) \right\} | \theta ,
\]

it is immediate that the oracle second-order U-statistic estimator \( \tilde{\psi}_1 \) is consistent when \( \theta_k \rightarrow \theta \in \text{in Robins et al. (2008)} \).

\( \tilde{\psi}_1 \) by \( \theta_k \rightarrow \theta \in \text{in Robins et al. (2008)} \).

Remark 2.1. The definition of \( \tilde{\psi}_2,k(\Omega_k^{-1}) \) in Robins et al. (2008) differs from that in the current paper in the sign; thus \( \tilde{\psi}_2,k(\Omega_k^{-1}) \equiv \tilde{\psi}_1 - \tilde{\psi}_2,k(\Omega_k^{-1}) \) would be \( \tilde{\psi}_1 + \tilde{\psi}_2,k(\Omega_k^{-1}) \) in Robins et al. (2008). We reversed the sign because it seems didactically useful to have \( \tilde{\psi}_2,k(\Omega_k^{-1}) \) be an unbiased estimator of \( cBias_{\theta,k}(\tilde{\psi}_1) \).

Before proceeding to the main results of this section (Section 2), we first compare certain properties of the parameters \( \mathbb{E}_\theta[\text{cov}_\theta[Y, A|X]] = \mathbb{E}_\theta[Y - b(X)] \{A - p(X) \} \) and \( \mathbb{E}_\theta[\text{var}_\theta[A|X]] = \mathbb{E}_\theta[(A - p(X))^2] \), where we note that all the earlier results and definitions concerning \( \mathbb{E}_\theta[\text{cov}_\theta[Y, A|X]] \) apply with \( \psi \equiv \psi(\theta) = \mathbb{E}_\theta[\text{var}_\theta[A|X]] \) when we everywhere substitute \( A, p, \tilde{p} \) for \( Y, b, \tilde{b} \). Given Condition B, key differences between the parameters are collected in the following lemma (Lemma 2.2), whose proof is trivial once we note that for \( \mathbb{E}_\theta[\text{var}_\theta[A|X]] \), in contrast with \( \mathbb{E}_\theta[\text{cov}_\theta[Y, A|X]] \), \( cBias_{\theta}(\tilde{\psi}_1) = \mathbb{E}_\theta[p(X) - \tilde{p}(X))^2 \theta], cBias_{\theta,k}(\tilde{\psi}_1) = \mathbb{E}_\theta[(p(X) - \tilde{p}(X)|Z_k\|^2\theta] \) and \( \tilde{\psi}_2,k(\tilde{\psi}_1) = \mathbb{E}_\theta[(p(X) - \tilde{p}(X)|Z_k\|^2\theta] \) are all non-negative. We thus have the following lemma:

Lemma 2.2. The followings are true for \( \psi(\theta) = \mathbb{E}_\theta[\text{var}_\theta[A|X]] \) but not for \( \psi(\theta) = \mathbb{E}_\theta[\text{cov}_\theta[Y, A|X]] \):

(i) \( cBias_{\theta,k}(\tilde{\psi}_1) \) is non-decreasing in \( k \) (since, by Condition B, the space spanned by \( Z_k \) increases with \( k \)) and, thus, \( cTB_{\theta,k}(\tilde{\psi}_1) \) is non-increasing in \( k \). That is, for \( k_2 > k_1 \)

\[
0 < cBias_{k_1,\theta}(\tilde{\psi}_1) \leq cBias_{k_2,\theta}(\tilde{\psi}_1) \leq cBias_{\theta}(\tilde{\psi}_1),
\]

\[
0 < cTB_{k_1,\theta}(\tilde{\psi}_1) \leq cTB_{k_2,\theta}(\tilde{\psi}_1) \leq 0.
\]

(ii) \( \tilde{\psi}_2,k(\tilde{\psi}_1) \) of \( \tilde{\psi}_2,k(\Omega_k^{-1}) \) is always less than or equal to the conditional bias \( cBias_{\theta}(\tilde{\psi}_1) \) of \( \tilde{\psi}_1 \).
(iii) For any $\delta > 0$, consider the null hypotheses

$$H_0(\delta) : \frac{|c\text{Bias}_\theta(\hat{\psi}_1)|}{\text{se}_\theta[\hat{\psi}_1]} = \frac{|c\text{Bias}_{\theta,k}(\hat{\psi}_1) + cTB_{\theta,k}(\hat{\psi}_1)|}{\text{se}_\theta[\hat{\psi}_1]} < \delta$$

and

$$H_{0,k}(\delta) : \frac{|c\text{Bias}_{\theta,k}(\hat{\psi}_1)|}{\text{se}_\theta[\hat{\psi}_1]} < \delta.$$  

If $H_0(\delta)$ (2.4) is true then $H_{0,k}(\delta)$ (2.5) is true. Hence rejection of $H_{0,k}(\delta)$ (2.5) implies rejection of $H_0(\delta)$ (2.4).

The null hypothesis $H_0(\delta)$ (2.4) states that the conditional bias $c\text{Bias}_\theta(\hat{\psi}_1)$ of $\hat{\psi}_1$ is less than a fraction $\delta$ of its standard error. In Theorem 2.9 below, we construct an $\omega$-level test for the null hypothesis $H_{0,k}(\delta)$ (2.5), which by (iii) is also an $\omega$-level test of $H_0(\delta)$ for $\psi(\theta) = \mathbb{E}_\theta[\text{var}_\theta[A|X]]$ but not for $\psi(\theta) = \mathbb{E}_\theta[\text{var}_\theta[Y,A|X]]$. Furthermore Lemma 2.2(ii) implies that for $\mathbb{E}_\theta[\text{cov}_\theta[Y,A|X]]$, unlike $\mathbb{E}_\theta[\text{var}_\theta[A|X]]$, we cannot guarantee that the estimator $\hat{\psi}_{2,k}(\Omega_k^{-1})$ is preferred to $\hat{\psi}_1$. Thus, one might reasonably ask whether our methods are useful for inference concerning the parameter $\mathbb{E}_\theta[\text{cov}_\theta[Y,A|X]]$, a question to which we return in the next Section 2.1.

**Remark 2.3.** The simulation study reported in Table 1 was for the parameter $\psi(\theta) = \mathbb{E}_\theta[\text{var}_\theta[A|X]]$. Were it not, our claim that the observation that the bias of $\hat{\psi}_{2,k}(\Omega_k^{-1})$ decreases as $k$ increases as predicted by the theory developed in Section 2 would have been false. Similarly, our claim that the test $\hat{\chi}_k^{(1)}(\Omega_k^{-1}; z, \delta)$ is an $\omega$-level test of $H_0(\delta)$ (2.4) would also have been false.

In apparent contradiction to these statements, in our simulation studies for the parameter $\mathbb{E}_\theta[\text{cov}_\theta[Y,A|X]]$ reported in Table 11 and Table 12 of Appendix K.3, the results were qualitatively the same as those in Table 1 (e.g the MCav of $\hat{\psi}_{2,k}(\Omega_k^{-1})$ increased with $k$). However this was due to the particular data generating process used and is not true in general for $\psi(\theta) = \mathbb{E}_\theta[\text{cov}_\theta[Y,A|X]]$.

An additional point in regard to studies reported in Table 1, the ratio of the MCav of the bias 0.197 of $\hat{\psi}_1$ for $\psi = \mathbb{E}_\theta[\text{var}_\theta[A|X]]$ to the MCav 0.034 of its estimated standard error was approximately 6. The theoretical prediction based on rates of convergence, ignoring constants, was close, being equal to 3.7, calculated as follows. In the simulation, $p(x)$ had a Hölder exponent $\beta_p$ of 0.251 and therefore the conditional bias $\mathbb{E}_\theta[(\hat{p}(X) - p(X))^2]$ was of order $n^{-2\beta_p}/(2\beta_p + 1) = n^{-1/3}$, because we used a rate minimax estimator $\hat{p}(x)$. 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 $n = 2500$ gives 3.7.

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 $\Omega_k$ are bounded away from 0 and $\infty$;
2. $\|\mathbb{Z}_k^* \mathbb{Z}_k\|_\infty \leq B_k$ for some constant $B > 0$;
3. $\|rac{d\mathbb{Z}_k^*}{dx}\|_\infty$ is bounded away from 0 and $\infty$;
4. The number of selected bases $k = k(n)$ in $\mathbb{Z}_k$ is $o(n^2)$; furthermore $k \to \infty$ as $n \to \infty$;
5. The residuals $\hat{\varepsilon}_b \equiv Y - \hat{b}(X)$ and $\hat{\varepsilon}_p \equiv A - \hat{p}(X)$ are bounded with probability 1.

**Remark 2.4.** Condition W(2) will only be needed in Section 3 below. Condition W(2) holds for Fourier, Daubechies wavelets and B-spline bases (Belloni et al., 2015). Although it does not hold in general for polynomial bases (Belloni et al., 2015), it will hold if we apply a bounded monotone transformation to each monomial. Condition W(5) might be considered a strong assumption by some. We impose it so we can focus on the issues of importance to us. However no important changes in our results would occur if we replaced (5) by the assumption that $\hat{\varepsilon}_b$ and $\hat{\varepsilon}_p$ are sub-Gaussian random variables.

Then we have the following result regarding the statistical properties of the oracle estimator $\hat{\Omega}_{22,k}(\Omega_k^{-1})$ of the projected bias $c\text{Bias}_{\theta,k}(\hat{\psi}_1)$ and of the oracle biased-corrected estimator $\hat{\psi}_{2,k}(\Omega_k^{-1})$. 


Theorem 2.5. Under Condition W, for $k = o(n^2)$

(i) Conditional on the training sample, $\mathbb{P}_{22,k}(\Omega_k^{-1})$ is unbiased for $cBias_{\theta,k}(\psi_1)$ with variance of order $k/n^2$ and $\hat{\psi}_{2,k}(\Omega_k^{-1}) = \hat{\psi}_1 - \mathbb{P}_{22,k}(\Omega_k^{-1})$ has mean $\psi(\theta) + cTB_{\theta,k}(\hat{\psi}_1)$ and variance of order $\max\{k,n\}/n^2$.

(ii) Conditional on the training sample, $(n^2/k)^{1/2}(\mathbb{P}_{22,k}(\Omega_k^{-1}) - cBias_{\theta,k}(\hat{\psi}_1))$ converges in law to a normal distribution with mean 0 and variance consistently estimated by $(n^2/k)\hat{\var}(\mathbb{P}_{22,k}(\Omega_k^{-1}))$ with $\hat{\var}(\mathbb{P}_{22,k}(\Omega_k^{-1})) \equiv \{\hat{\var}(\mathbb{P}_{22,k}(\Omega_k^{-1}))\}^2$ defined in eq. (H.1).

Furthermore, conditional on the training sample,

$$\left\{ \left( \frac{\max\{k,n\}}{n^2} \right) \right\}^{-1/2} \left( \hat{\psi}_{2,k}(\Omega_k^{-1}) - (\hat{\psi}(\theta) + cTB_{\theta,k}(\hat{\psi}_1)) \right)$$

converges in law to a normal distribution with mean 0 and variance that can be consistently estimated by $(n^2/\max\{k,n\})\hat{\var}(\hat{\psi}_{2,k}(\Omega_k^{-1}))$, where $\hat{\var}(\hat{\psi}_{2,k}(\Omega_k^{-1})) = \hat{\var}(\hat{\psi}_1) + \hat{\var}(\mathbb{P}_{22,k}(\Omega_k^{-1}))$.

(iii) $\mathbb{P}_{22,k}(\Omega_k^{-1}) \pm z_{\omega/2}\hat{\var}(\mathbb{P}_{22,k}(\Omega_k^{-1}))$ is a $(1-\omega)$ asymptotic conditional CI for $cBias_{\theta,k}(\hat{\psi}_1)$ with length of order $(k/n^2)^{1/2}$.

$\mathbb{P}_{22,k}(\Omega_k^{-1}) \pm z_{\omega/2}\hat{\var}(\mathbb{P}_{22,k}(\Omega_k^{-1}))$, $\infty$ is a $(1-\omega)$ asymptotic conditional one-sided CI for $cBias_{\theta,k}(\hat{\psi}_1)$.

(iv) $\hat{\psi}_{2,k}(\Omega_k^{-1}) \pm z_{\alpha/2}\hat{\var}(\hat{\psi}_{2,k}(\Omega_k^{-1}))$ is a $(1-\alpha)$ asymptotic conditional CI for $\psi(\theta) + cTB_{\theta,k}(\hat{\psi}_1)$.

$\hat{\psi}_{2,k}(\Omega_k^{-1}) \pm z_{\alpha}\hat{\var}(\hat{\psi}_{2,k}(\Omega_k^{-1}))$, $\infty$ is a $(1-\alpha)$ asymptotic conditional one-sided CI for $\psi(\theta) + cTB_{\theta,k}(\hat{\psi}_1)$.

Proof. Except for asymptotic normality, Theorem 2.5 follows from Robins et al. (2008). When $k = o(n^2)$ and $k \to \infty$ as $n \to \infty$, the (conditional) asymptotic normality of $\sqrt{n^2/k} \left\{ \mathbb{P}_{22,k}(\Omega_k^{-1}) - cBias_{\theta,k}(\hat{\psi}_1) \right\}$ follows directly from Bhattacharya and Ghosh (1992, Corollary 1.2), by verifying the conditions (1) through (5) therein. When $k > n$ and $k = o(n^2)$, Robins et al. (2016, Theorem 1) can also be used to establish the (conditional) asymptotic normality of $\sqrt{n^2/k} \left\{ \mathbb{P}_{22,k}(\Omega_k^{-1}) - cBias_{\theta,k}(\hat{\psi}_1) \right\}$. \(\square\)

Remark 2.6. The left panel of the qqplots in Figure 4 (see Appendix L.2) provides strong empirical evidence that, in our simulation experiments, the distribution of $\mathbb{P}_{22,k}(\Omega_k^{-1})$ at the sample size of 2500 described in Appendix K.1 is nearly normal.

Remark 2.7. When $k$ is of order greater than or equal to $n^2$, the asymptotic normality of $\sqrt{n^2/k} \{ \mathbb{P}_{22,k}(\Omega_k^{-1}) - cBias_{\theta,k}(\hat{\psi}_1) \}$ does not hold. Moreover, when $k \gg n^2$, $\var\{ \mathbb{P}_{22,k}(\Omega_k^{-1}) | \theta \} = O(\frac{n}{k})$ is then of order greater than 1, and therefore $\mathbb{P}_{22,k}(\Omega_k^{-1})$ cannot help estimate $cBias_{\theta,k}(\hat{\psi}_1)$ even if $cBias_{\theta,k}(\hat{\psi}_1)$ is of order 1.

Remark 2.8 (Undersmoothing, three-way sample splitting, and $\hat{\psi}_{2,k}(\Omega_k^{-1})$). In this remark, we provide a heuristic understanding of the relationship between $\hat{\psi}_1$ and $\hat{\psi}_{2,k}(\Omega_k^{-1})$ by considering the relationship of each to the undersmoothed, triple sample splitting estimator $\hat{\psi}_{1, NR}$ of Newey and Robins (2018). We begin by comparing $\hat{\psi}_1$ with $\hat{\psi}_{1, NR}$.

Recall that the bias of $\hat{\psi}_1 = \frac{1}{n} \sum_{i=1}^{n} \left\{ A_i - \hat{\psi}(X_i) \right\}^2$ as an estimator of $\psi := E_\theta[\var\theta[A|X]]$ is $E_\theta[\{p(X) - \hat{\psi}(X)\}^2]$. If $p(x)$ lies in a Hölder ball with exponent $\beta_p$, then if, as we assume, the density of the $d$-dimensional vector $X$ is known, arguments analogous to those in Stone (1980, 1982) shows that for any estimator $\tilde{p}(x)$ of $p(x)$ of $E_\theta\left\{ (p(X) - \hat{\psi}(X))^2 \right\}$ is at best $O_{P_0}\left(n^{-\frac{2d}{2d+\beta_p}}\right)$, which can be achieved by the estimator $\tilde{p}(x) = \hat{\alpha}_k^T \hat{z}_k(x)$ with $\hat{\alpha}_k = \frac{1}{n} \left\{ \sum_i A_i \hat{z}_k(X_i)^T \right\} \Omega_k^{-1}$, with $k = n^{\frac{d}{2d+\beta_p}}$ and $\hat{z}_k(x)$ the first $k$ bases of a suitably chosen spline or wavelet orthonormal basis for $L_2(P_0)$. This implies that $\beta_p > d/2$ is needed for the bias of $\hat{\psi}_1$ to be $E_\theta\left\{ (p(X) - \hat{\psi}(X))^2 \right\} = o_{P_0}(n^{-1/2})$. Furthermore, the bias of $\hat{\psi}_1$ increases if we use an undersmoothed estimator $\tilde{p}(x)$ obtained by choosing $k \gg n^{\frac{d}{2d+\beta_p}}$ orthonormal bases. (Note choosing $k = n^{\frac{d}{2d+\beta_p}}$ equalizes the order $k/n$ of the variance of $\tilde{p}(x)$ and the order $k^{-2\beta_p/d}$ of the square of the approximation bias $E_\theta[\var\theta[p(X)|Z_k]]\theta).
However, suppose as in Newey and Robins (2018), we replaced \( \widehat{\psi}_1 \) defined in Section 1.2 by

\[
\widehat{\psi}_{1,NR} = \frac{1}{n} \sum_{i=1}^{n} \left\{ A_i - \widehat{p}_1(X_i) \right\} \left\{ A_i - \widehat{p}_2(X_i) \right\}
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} A_i^2 - A_i \left\{ \widehat{p}_1(X_i) + \widehat{p}_2(X_i) \right\} + \widehat{p}_1(X_i) \widehat{p}_2(X_i)
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} A_i^2 - (A_i - \widehat{p}_2(X_i)) \widehat{p}_1(X_i) - (A_i - \widehat{p}_1(X_i)) \widehat{p}_2(X_i) - \widehat{p}_1(X_i) \widehat{p}_2(X_i)
\]

where the training sample is itself randomly split into two subsamples \( I_1 \) and \( I_2 \) of equal size and the regression coefficients in \( \widehat{p}_1(x) = \hat{\alpha}_1^T \zeta_k(x) \) and \( \widehat{p}_2(x) = \hat{\alpha}_2^T \zeta_k(x) \) are computed from subjects in subsamples \( I_1 \) and \( I_2 \) respectively. Hence \( \widehat{\psi}_{1,NR} \) is computed from three independent samples and uses the true covariance matrix \( \Omega_k \) and its inverse. Newey and Robins (2018) show that unconditionally the (random) bias \( \mathbb{E}_n[\{p(X) - \widehat{p}_1(X)\} \{p(X) - \widehat{p}_2(X)\}] \) of \( \widehat{\psi}_{1,NR} \) is of order \( (k/n^2 + k^{-4\beta_p/d})^{1/2} \) in probability which is minimized by choosing \( k = n^{2d/(d+4\beta_p)} \), for which \( \mathbb{E}_n[\{p(X) - \widehat{p}_1(X)\} \{p(X) - \widehat{p}_2(X)\}] \) is of order \( n^{-4\beta_p/d} \).

The bias of order \( n^{-4\beta_p/d} \) is \( o_p(n^{-1/2}) \) if \( \beta_p > 1/4 \). Note also that \( n^{-4\beta_p/d} > n^{-2\beta_p/d} \), so, unlike with \( \widehat{\psi}_1 \), it is optimal to undersmooth the estimators \( \widehat{p}_1(x) \) and \( \widehat{p}_2(x) \) to minimize the bias of \( \widehat{\psi}_{1,NR} \). Hence the secret sauce behind the much better performance of \( \widehat{\psi}_{1,NR} \) compared to \( \widehat{\psi}_1 \) is a combination of three-way (rather than two-way) sample splitting combined with undersmoothing in the estimation of \( \widehat{p}_1(x) \) and \( \widehat{p}_2(x) \).

We now turn to a comparison of \( \widehat{\psi}_{2,k}(\Omega_k^{-1}) \) and \( \widehat{\psi}_{1,NR} \). For didactic purposes, it will be useful to first consider the case in which \( \widehat{p}(x) \) is artificially chosen to be identically zero; then \( \widehat{\psi}_1 = \frac{1}{n} \sum_{i=1}^{n} A_i^2 \),

\[
\widehat{\mathbb{F}}_{22,k}(\Omega_k^{-1}) = \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} \left\{ A_i \zeta_k(X_{i1})^T \Omega_k^{-1} \right\} \zeta_k(X_{i2}) A_{ij},
\]

and \( \widehat{\psi}_{2,k}(\Omega_k^{-1}) = \frac{1}{n} \sum_{i=1}^{n} A_i^2 - \widehat{\mathbb{F}}_{22,k}(\Omega_k^{-1}) \).

We now show that \( \widehat{\psi}_{1,NR} \) and \( \widehat{\psi}_{2,k}(\Omega_k^{-1}) \) have the mean and the same order of variance \( 1/n \), with the variance of \( \widehat{\psi}_{2,k}(\Omega_k^{-1}) \) strictly less than that of \( \widehat{\psi}_{1,NR} \). Since both \( \widehat{\psi}_{1,NR} \) and \( \widehat{\psi}_{2,k}(\Omega_k^{-1}) \) have the common term \( \frac{1}{n} \sum_{i=1}^{n} A_i^2 \), our goal becomes to compare \( \widehat{\psi}_{1,NR} - \frac{1}{n} \sum_{i=1}^{n} A_i^2 \) with \( \widehat{\mathbb{F}}_{22,k}(\Omega_k^{-1}) \). Rather than doing so directly, let us consider the following third-order U-statistic that substitutes the unbiased estimator \( \zeta_k(X_{i3}) \zeta_k(X_{i3})^T \) for \( \Omega_k \) to give

\[
\widehat{\mathbb{F}}_{22,k}(\Omega_k^{-1}) = \frac{1}{n(n-1)(n-2)} \sum_{1 \leq i \neq j \neq k \leq n} \left\{ A_i \zeta_k(X_{i1})^T \Omega_k^{-1} \right\} \zeta_k(X_{i3}) \zeta_k(X_{i3})^T \Omega_k^{-1} \left\{ \zeta_k(X_{i2}) A_{ij} \right\} \Omega_k^{-1} \left\{ \zeta_k(X_{i2}) A_{ij} \right\}.
\]

It thus follows that \( \widehat{\mathbb{F}}_{22,k}(\Omega_k^{-1}) \) and \( \widehat{\mathbb{F}}_{22,k}(\Omega_k^{-1}) \) have the same mean and the same order of variance \( 1/n \), with the variance of \( \widehat{\mathbb{F}}_{22,k}(\Omega_k^{-1}) \) being smaller. Denote the estimation sample of size \( n \) as \( I_3 \), and the sample sizes of \( I_1 \) and \( I_2 \) as \( n_1 \) and \( n_2 \) respectively. Then note that \( \widehat{\mathbb{F}}_{22,k}(\Omega_k^{-1}) \) and

\[
\frac{1}{n} \sum_{i=1}^{n} \widehat{p}_1(X_i) \widehat{p}_2(X_i) = \frac{1}{n n_1 n_2} \sum_{i_1 \in I_1, i_2 \in I_2, i_3 \in I_3} \left\{ A_i \zeta_k(X_{i1})^T \Omega_k^{-1} \right\} \zeta_k(X_{i3}) \zeta_k(X_{i3})^T \Omega_k^{-1} \left\{ \zeta_k(X_{i2}) A_{ij} \right\}
\]

have identical kernels

\[
\left\{ A_i \zeta_k(X_{i1})^T \Omega_k^{-1} \right\} \zeta_k(X_{i3}) \zeta_k(X_{i3})^T \Omega_k^{-1} \left\{ \zeta_k(X_{i2}) A_{ij} \right\}.
\]
and thus identical expectations. They differ only in that $\hat{\Pi}_{22,k}^*(\Omega_k^{-1})$ is a third-order $U$-statistic while $\hat{\Pi}_{22,k}^\dagger(\Omega_k^{-1})$ splits the sample into three subsets $I_1, I_2$ and $I_3$. Both $n^{-1} \sum_{i=1}^n \tilde{p}_1(X_i)\tilde{p}_2(X_i)$ and $\hat{\Pi}_{22,k}^\dagger(\Omega_k^{-1})$ have variance of order $1/n$ although the constants will differ with $\hat{\Pi}_{22,k}^*(\Omega_k^{-1})$ having the smaller variance. Furthermore $\hat{\psi}_{1,N,R} - \frac{1}{n} \sum_{i=1}^n A_i^2$ has the two additional mean zero terms $-(A_i - \tilde{p}_2(X_i))\tilde{p}_1(X_i) - (A_i - \tilde{p}_1(X_i))\tilde{p}_2(X_i)$ which both have variance of order $1/n$.

In summary, $\hat{\psi}_{2,k}^*(\Omega_k^{-1}) := \frac{1}{n} \sum_{i=1}^n A_i^2 - \hat{\Pi}_{22,k}^*(\Omega_k^{-1})$ and $\hat{\psi}_{1,N,R}$ have the same kernel and thus the same mean as $\hat{\psi}_{2,k}^*(\Omega_k^{-1})$. It also follows from the above arguments that $\hat{\psi}_{1,N,R}, \hat{\psi}_{2,k}^*(\Omega_k^{-1})$ and $\hat{\psi}_{2,k}^*(\Omega_k^{-1})$ have the same order of variance with the variance of $\hat{\psi}_{2,k}^*(\Omega_k^{-1})$ being the smallest. Although $\hat{\psi}_{1,N,R}$ has larger variance than $\hat{\psi}_{2,k}^*(\Omega_k^{-1})$, nonetheless when $\beta_p > 1/4$ (and thus both are consistent and asymptotically normal), the cross fit versions will have the same variance equal to the semiparametric variance bound.

Lastly consider the case where, as earlier, we have a preliminary machine learning estimator $\hat{p}(x)$ computed from a second independent (training) sample in the case of $\hat{\psi}_{2,k}^*(\Omega_k^{-1})$ and from a fourth sample $I_4$ in the case of $\hat{\psi}_{1,N,R}$. If, in defining $\hat{\psi}_{1,N,R}$, we redefine $\hat{\alpha}_j$ as $\frac{1}{n_i} \left\{ \sum_{X_i \in I_j} \{ A_i - \hat{p}(X_i)\} \tilde{z}_k(X_i)^\top \right\} \Omega_k^{-1}$ with $j = 1, 2$, and use $\hat{p}(x)$ in $\hat{\psi}_{2,k}^*(\Omega_k^{-1})$ as earlier, the relationships between $\hat{\psi}_{1,N,R}$ and $\hat{\psi}_{2,k}^*(\Omega_k^{-1})$ remain as above.

Based on the statistical properties of $\hat{\psi}_1$ and $\hat{\Pi}_{22,k}^*(\Omega_k^{-1})$ summarized in Theorem 1.3 and Theorem 2.5, for the parameter expected conditional variance $\psi := \mathbb{E}_\delta[\text{var}_\theta[A,Y]]$ (resp. expected conditional covariance $\psi = \mathbb{E}_\vartheta[\text{cov}_\vartheta[A,Y|X]]$, we now consider the properties of the following one-sided test $\hat{\chi}_k^1(\Omega_k^{-1};\zeta_k, \delta)$ (resp. two-sided test $\hat{\chi}_k^2(\Omega_k^{-1};\zeta_k, \delta)$) of $H_{0,k}(\delta)$ (2.5):

$$\hspace{1cm}
\hat{\chi}_k^1(\Omega_k^{-1};\zeta_k, \delta) := I \left\{ \frac{\hat{\Pi}_{22,k}^*(\Omega_k^{-1})}{\hat{\psi}_1} - \zeta_k \left[ \frac{\hat{\Pi}_{22,k}^*(\Omega_k^{-1})}{\hat{\psi}_1} \right] > \delta \right\},$$

and $\hat{\chi}_k^2(\Omega_k^{-1};\zeta_k, \delta) := I \left\{ \frac{\hat{\Pi}_{22,k}^*(\Omega_k^{-1})}{\hat{\psi}_1} - \zeta_k \left[ \frac{\hat{\Pi}_{22,k}^*(\Omega_k^{-1})}{\hat{\psi}_1} \right] > \delta \right\},$ for user-specified $\zeta_k, \delta > 0$. $\zeta_k$ determines the level of the test $\hat{\chi}_k^1(\Omega_k^{-1};\zeta_k, \delta)$ (resp. $\hat{\chi}_k^2(\Omega_k^{-1};\zeta_k, \delta)$).

In Appendix A, we prove the following Theorem 2.9 that characterizes the asymptotic level and power of the oracle one-sided test $\hat{\chi}_k^1(\Omega_k^{-1};\zeta_k, \delta)$ (resp. two-sided test $\hat{\chi}_k^2(\Omega_k^{-1};\zeta_k, \delta)$) of $H_{0,k}(\delta)$ (2.5) for the parameter expected conditional variance (resp. expected conditional covariance). For example, Theorem 2.9 implies that $\hat{\chi}_k^1(\Omega_k^{-1};\zeta_k, \delta)$ (resp. $\hat{\chi}_k^2(\Omega_k^{-1};\zeta_k, \delta)$) is an $\omega$ level one-sided test (resp. two-sided test) of $H_{0,k}(\delta)$ (2.5).

**Theorem 2.9.** Under Condition W, when $k$ increases with $n$ but $k = o(n)$, for any given $\delta, \zeta_k > 0$, suppose that $\frac{\hat{\alpha}_k(\hat{\psi}_1)}{\text{seq}[\hat{\psi}_1]} \gamma = \gamma$ for some (sequence) $\gamma = \gamma(n)$ (where $\gamma(n)$ can diverge with $n$), then the rejection probability of $\hat{\chi}_k^1(\Omega_k^{-1};\zeta_k, \delta)$ (resp. $\hat{\chi}_k^2(\Omega_k^{-1};\zeta_k, \delta)$) is asymptotically equivalent to $1 - \Phi(\zeta_k - \sqrt{\frac{k}{n}} \vartheta(\gamma - \delta))$ (resp. $2 - \Phi(\zeta_k - \sqrt{\frac{k}{n}} \vartheta(\gamma - \delta)) - \Phi(\zeta_k + \sqrt{\frac{k}{n}} \vartheta(\gamma + \delta))$) where $\vartheta > 0$ and $\vartheta := \sqrt{\frac{k}{n} \text{seq}[\hat{\psi}_1]}$, as $n \to \infty$. In particular,

1. under $H_{0,k}(\delta)$ (2.5), $\hat{\chi}_k^1(\Omega_k^{-1};\zeta_k, \delta)$ (resp. $\hat{\chi}_k^2(\Omega_k^{-1};\zeta_k, \delta)$) is an asymptotically conservative level $1 - \Phi(\zeta_k)$ one-sided test (resp. $2 - \Phi(\zeta_k) - \Phi(\zeta_k + 2 \sqrt{\frac{k}{n}} \vartheta) \downarrow 1 - \Phi(\zeta_k)$ two-sided test), as $n \to \infty$;
2. under the alternative to $H_{0,k}(\delta)$ (2.5), i.e. $|\gamma| = |\gamma(n)| > \delta$,
   (i) if $|\gamma| - \delta = C \sqrt{\frac{k}{n}}$ for some $C > 0$, then $\hat{\chi}_k^1(\Omega_k^{-1};\zeta_k, \delta)$ (resp. $\hat{\chi}_k^2(\Omega_k^{-1};\zeta_k, \delta)$) has rejection probability asymptotically equivalent to $1 - \Phi(\zeta_k - C \vartheta)$ (resp. $2 - \Phi(\zeta_k - C \vartheta) - \Phi(\zeta_k + 2 \sqrt{\frac{k}{n}} \vartheta + C \vartheta) \downarrow 1 - \Phi(\zeta_k - C \vartheta)$), as $n \to \infty$.

---

1. The local alternative regime $\gamma - \delta = \sqrt{\frac{k}{n}}$ is a consequence of the variance of the statistic $\frac{\hat{\alpha}_k(\hat{\psi}_1)}{\text{seq}[\hat{\psi}_1]}$ (resp. $\frac{\hat{\alpha}_k(\hat{\psi}_1)}{\text{seq}[\hat{\psi}_1]}$) in the definition of the test 2.6 (resp. 2.7) of order $k/n$. 

(ii) if \(|\gamma| - \delta = o(\sqrt{n})\), then \(\hat{\chi}_k^{(1)}(\Omega^{-1}_k; \zeta, \delta)\) (resp. \(\hat{\chi}_k^{(2)}(\Omega^{-1}_k; \zeta, \delta)\)) has rejection probability asymptotically equivalent to the level of the test \(1 - \Phi(\zeta)\) (resp. \(2 - \Phi(\zeta) - \Phi(\zeta + 2\sqrt{\frac{\delta}{k}})\)) \((1 - \Phi(\zeta))\), as \(n \to \infty\);
(iii) if otherwise, then \(\hat{\chi}_k^{(1)}(\Omega^{-1}_k; \zeta, \delta)\) (resp. \(\hat{\chi}_k^{(2)}(\Omega^{-1}_k; \zeta, \delta)\)) has rejection probability converging to 1, as \(n \to \infty\).

In summary, on one hand, for any fixed \(\delta > 0\), \(\hat{\chi}_k^{(1)}(\Omega^{-1}_k; \zeta, \delta)\) (resp. \(\hat{\chi}_k^{(2)}(\Omega^{-1}_k; \zeta, \delta)\)) is an asymptotically conservative level \(\omega\) one-sided test (resp. two-sided test) of \(H_{0,k}(\delta)\) \((2.5)\). The asymptotic power of \(\hat{\chi}_k^{(1)}(\Omega^{-1}_k; \zeta, \delta)\) (resp. \(\hat{\chi}_k^{(2)}(\Omega^{-1}_k; \zeta, \delta)\)) as a test of \(H_{0,k}(\delta)\) \((2.5)\), on the other hand, can be divided into three different regimes: (1) local alternative regime as in Theorem 2.9(2.i) – the asymptotic power is 1; (2) under the local alternative regime as in Theorem 2.9(2.ii) – the asymptotic power is the asymptotic level; and (3) above the local alternative regime as in Theorem 2.9(2.iii) – the asymptotic power is 1.

We now show how to obtain the upper confidence bounds on the actual asymptotic coverage of the Wald CI associated with \(\hat{\psi}_1\) by the duality between hypothesis testing and confidence intervals. We start with defining the following function, for some \(\delta > 0\),

\[
TC_\alpha(\delta) := \Phi(z_{\alpha/2} - \delta) - \Phi(-z_{\alpha/2} - \delta),
\]

the utility of which is summarized in Lemma 2.10 below:

**Lemma 2.10.** Suppose an estimator \(\hat{\mu}\) is asymptotic normal under \(P_0\) with centering by the mean \(E_0[\hat{\mu}]\) and scaling by the standard error \(se[\hat{\mu}] = \sqrt{\text{var}[\hat{\mu}]}\).

(i) Then for any constant \(\rho\), \(TC_\alpha(\frac{\rho}{\text{se}[\hat{\mu}]})\) is the actual asymptotic coverage probability of \(E_0[\hat{\mu}] + \rho\) by the nominal \((1 - \alpha)\) Wald interval \(\hat{\mu} \pm z_{\alpha/2} se[\hat{\mu}]\) where \(se[\hat{\mu}]\) is a consistent estimator of \(se[\hat{\mu}]\).

Suppose now \(\rho\) is unknown but we have an estimator \(\hat{\rho}\) that is asymptotic normal with mean \(\rho\) and standard error \(se[\hat{\rho}]\).

(ii) Then \(TC_\alpha(\frac{\hat{\rho} - \zeta_se[\hat{\rho}]}{se[\hat{\rho}]}\) (resp. \(TC_\alpha(\frac{\hat{\rho} - \zeta_se[\hat{\rho}]}{se[\hat{\rho}]}\)) is an asymptotic \(\Phi(\zeta)\) (resp. \(\Phi(\zeta) + \Phi(\zeta + \frac{2|\rho|}{\text{se}[\hat{\rho}]})\)) \((1 - \Phi(\zeta)) = 2\Phi(\zeta) - 1\) upper confidence bound for the actual asymptotic coverage \(TC_\alpha(\frac{\hat{\rho}}{\text{se}[\hat{\rho}]})\) (resp. \(TC_\alpha(\frac{\hat{\rho}}{\text{se}[\hat{\rho}]})\)) of \(E_0[\hat{\mu}] + \rho\) by the Wald interval \(\hat{\mu} \pm z_{\alpha/2} se[\hat{\mu}]\) where \(se[\hat{\rho}]\) is a consistent estimator of \(se[\hat{\rho}]\) when the sign of \(\rho\) is a priori known to be nonnegative (resp. when the sign of \(\rho\) is a priori unknown). In particular, for any fixed \(\rho\), \(TC_\alpha(\frac{\hat{\rho} - \zeta_se[\hat{\rho}]}{se[\hat{\rho}]})\) is an asymptotic \(\Phi(\zeta)\) upper confidence bound for the actual asymptotic coverage \(TC_\alpha(\frac{\hat{\rho}}{\text{se}[\hat{\rho}]})\) of \(E_0[\hat{\mu}] + \rho\) by the Wald interval \(\hat{\mu} \pm z_{\alpha/2} se[\hat{\mu}]\) when the sign of \(\rho\) is a priori unknown.

By the above Lemma 2.10, \(TC_\alpha(\delta)\) is the actual asymptotic coverage of a two-sided \((1 - \alpha)\) Wald CI centered on \(\hat{\psi}_1\) when \(\hat{\psi}_1\) is asymptotically normally distributed and its bias is a fraction \(\delta\) of its standard error. Notice that \(TC_\alpha(\delta)\) is decreasing in \(\delta\). A picture of \(TC_\alpha(\delta)\) for several commonly used \(\alpha\) is shown in Figure 1. Given the mapping \(TC_\alpha(\delta)\) between the ratio \(\delta\) (of the bias to the standard error of \(\hat{\psi}_1\)) and the actual asymptotic coverage of the Wald CI, we can find a \((1 - \omega)\) upper confidence bound on the actual asymptotic coverage of the Wald CI associated with \(\hat{\psi}_1\) for the truncated parameter \(\hat{\psi}_1(\theta) \equiv \psi(\theta) + cTB_{\theta,k}(\hat{\psi}_1)\) by plugging into \(TC_\alpha(\delta)\) the smallest \(\delta\) such that the one-sided test \(\hat{\chi}_k^{(1)}(\Omega^{-1}_k; \zeta, \omega, \delta)\) (resp. the two-sided test \(\hat{\chi}_k^{(2)}(\Omega^{-1}_k; \zeta, \omega, \delta)\)) fails to reject, when \(\psi(\theta) = E_0[\text{var}_0[A|X]]\) (resp. \(\psi(\theta) = E_0[\text{cov}_0[A,Y|X]]\)).

We thus have the following result, to state which we need to define

\[
UCB^{(1)}(\Omega^{-1}_k; \alpha, \omega) := TC_\alpha\left(\frac{\Pi_{22,k}(\Omega^{-1}_k) - z_\omega se[\Pi_{22,k}(\Omega^{-1}_k)]}{se[\hat{\psi}_1]}\right).
\]

and

\[
UCB^{(2)}(\Omega^{-1}_k; \alpha, \omega) := TC_\alpha\left(\frac{||\Pi_{22,k}(\Omega^{-1}_k) - z_\omega se[\Pi_{22,k}(\Omega^{-1}_k)]||}{se[\hat{\psi}_1]}\right).
\]

Then we have
Corollary 2.11. Under the conditions of Theorem 2.9, suppose that \( \frac{cBias_{\theta,k}(\hat{\psi}_1)}{se[\hat{\psi}_1]} = \gamma \) for some fixed \( \gamma \), if \( \psi(\theta) = \mathbb{E}_{\theta}[\text{var}_{\theta}[A|X]] \) (resp. \( \psi(\theta) = \mathbb{E}_{\theta}[\text{cov}_{\theta}[A,Y|X]] \)), then \( UCB^{(1)}(\hat{\psi}_1, \alpha, \omega) \) (resp. \( UCB^{(2)}(\hat{\psi}_1, \alpha, \omega) \)) is a valid asymptotic \((1 - \omega)\) upper confidence bound for the actual asymptotic coverage of the \((1 - \omega)\) Wald confidence interval \( \hat{\psi}_1 \pm z_{\alpha/2} \overline{s}[\hat{\psi}_1] \) for the truncated parameter \( \psi(\theta) \equiv \psi(\theta) + cTB_{\theta,k}(\hat{\psi}_1) \).

To this point, assuming knowing \( \Omega_{k}^{-1} \), we have established the conditional asymptotic properties of the oracle one-sided test \( \hat{\chi}^{(1)}_{k}(\Omega_{k}^{-1}; \zeta, \delta) \) (resp. two-sided test \( \hat{\chi}^{(2)}_{k}(\Omega_{k}^{-1}; \zeta, \delta) \)) of the null hypothesis \( H_{0,k}(\delta) \) \((2.5)\) and the corresponding upper confidence bound \( UCB^{(1)}(\hat{\psi}_1, \alpha, \omega) \) (resp. \( UCB^{(2)}(\hat{\psi}_1, \alpha, \omega) \)) on the actual conditional asymptotic coverage of the \((1 - \omega)\) Wald CI associated with \( \hat{\psi}_1 \) for the truncated parameter \( \psi(\theta) \equiv \psi(\theta) + cTB_{\theta,k}(\hat{\psi}_1) \), where \( \psi \) is the expected conditional variance (resp. expected conditional covariance).

Furthermore, when the parameter of interest \( \psi \) is the expected conditional variance and thus \( cTB_{\theta,k}(\hat{\psi}_1) \) is nonincreasing in \( k \) by Lemma 2.2, the constructed upper confidence bound \( UCB^{(1)}(\hat{\psi}_1, \alpha, \omega) \) is also a \((1 - \omega)\) upper confidence bound for the actual conditional asymptotic coverage for \( \psi \) by the interval the \( \hat{\psi}_1 \pm z_{\alpha/2} \overline{s}[\hat{\psi}_1] \).

2.1. Inference for the expected conditional covariance. The discussion in this section applies not only to \( \mathbb{E}_{\theta}[\text{cov}_{\theta}[A,Y|X]] \), but also to any parameter \( \psi \) with a unique first order influence function depending on unknown nuisance functions for which the absolute value \( cTB_{\theta,k}(\hat{\psi}_1) \) of the truncation bias need not be a non-increasing function of \( k \). In particular it applies to the class of functionals with the mixed bias property of Rotnitzky et al. (2019a). Such parameters cover most causal parameters, including the average treatment effect and the effect of treatment on the treated, as well as many non-causal parameters. It is the class of parameters mentioned in the Section 1 for which our results are unavoidably less sharp and require more careful interpretation. Parameters for which \( cTB_{\theta,k}(\hat{\psi}_1) \) is guaranteed to be non-increasing in \( k \), however, constitute the class of parameters for which we obtain much sharper results. This class includes not only \( \mathbb{E}_{\theta}[\text{var}_{\theta}[A|X]] \) but also \( \mathbb{E}_{\theta}[\{b(X) - \hat{b}(X)\}^2 \delta \mathbb{E}_{\theta}[\{p(X) - \hat{p}(X)\}^2 \delta] \), the square of “Cauchy-Schwarz bias”, as discussed below.

Theorem 2.9 implies that inference concerning \( cBias_{\theta,k}(\hat{\psi}_1) \) is the same for \( \psi = \mathbb{E}_{\theta}[\text{var}_{\theta}[A|X]] \) and \( \psi = \mathbb{E}_{\theta}[\text{cov}_{\theta}[A,Y|X]] \) except, in the covariance case, two-sided rather than one-sided tests are needed, because it may be either positive or negative. Suppose that \( \hat{\psi}_{2,k} = \hat{\psi}_1 - \mathbb{P}_{22,k}(\Omega_{k}^{-1}) \) differed in magnitude from \( \hat{\psi}_1 \) by many times the standard error of \( \hat{\psi}_1 \), so our two-sided test \( \hat{\chi}^{(2)}_{k}(\Omega_{k}^{-1}; \zeta, \delta = 1.64, \delta) \) rejects \( H_{0,k}(\delta) \) \((2.5)\) for \( \delta = 6 \), say. Since \( \mathbb{E}_{\theta}[\hat{\psi}_{2,k} - \hat{\psi}] = cTB_{\theta,k}(\hat{\psi}_1) \) and \( \mathbb{E}_{\theta}[\hat{\psi}_1 - \hat{\psi}] = cTB_{\theta,k}(\hat{\psi}_1) + cBias_{\theta,k}(\hat{\psi}_1) \),
we would prefer \( \hat{\psi}_{2,k} \) to \( \hat{\psi}_1 \) and would reject \( H_0(\delta) \) (2.4) as well if \( cTB_{\theta,k}(\hat{\psi}_1) \) and \( cBias_{\theta,k}(\hat{\psi}_1) \) were of the same sign, as then the absolute value of the bias of \( \hat{\psi}_1 \) would exceed that of \( \hat{\psi}_{2,k} \). However, for \( \mathbb{E}_\theta[\text{cov}_\theta(A,Y|X)] \), we cannot know the sign of \( cTB_{\theta,k}(\hat{\psi}_1) \). Thus, we will not be able to make as strong inferential statements as for \( \mathbb{E}_\theta[\text{var}_\theta(A|X)] \).

In fact, 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 1.3 as, in current literature, authors often report a nominal (1 − \( \alpha \)) Wald confidence interval \( \hat{\psi}_1 \pm z_{\alpha/2}\hat{\text{se}}[\hat{\psi}_1] \), or, more commonly \( \hat{\psi}_{\text{cross-fit},1} \pm z_{\alpha/2}\hat{\text{se}}[\hat{\psi}_{\text{cross-fit},1}] \), based on a DR-ML estimator \( \hat{\psi}_1 \) and then appeal to Theorem 1.3 to support a validity claim that the actual coverage is no less than nominal. Specifically Theorem 1.3 implies validity under the null hypothesis \( cBias_\theta(\hat{\psi}_1) = o_P(n^{-1/2}) \) that the bias of \( \hat{\psi}_1 \) for \( \psi \) is \( o_P(n^{-1/2}) \). The use of the asymptotic \( o_P(n^{-1/2}) \) is implicitly justified by the tacit assumption that, at their sample size of \( N = 2n = 2n_{\text{tr}} \), they are nearly in asymptopia not only, like us, in regards to convergence to conditional normality in the estimation sample, but unlike us, in regards to the ratio \( cBias_\theta(\hat{\psi}_1)/\hat{\text{se}}[\hat{\psi}_1|\theta] \) being close to its asymptotic limit of 0.

However most authors fail to quantify or operationalize the last clause. In line with the approach of this paper, whenever a null hypothesis is defined in terms of an asymptotic rate of convergence such as \( o_P(n^{-1/2}) \) in the training sample data, we will (1) ask the authors to specify a positive number \( \delta_{op} = \delta_{op}(N) \) possibly depending on the actual sample size \( N \) of their study and (2) then operationalize the asymptotic null hypothesis \( cBias_\theta(\hat{\psi}_1) = o_P(n^{-1/2}) \) as the null hypothesis \( H_0(\delta_{op}) \). That is, we have the operationalized pair

\[
NH_0 : cBias_\theta(\hat{\psi}_1) = o_P(n^{-1/2})
\]

\[
H_0(\delta_{op}) : \frac{|cBias_\theta(\hat{\psi}_1)|}{\hat{\text{se}}[\hat{\psi}_1|\theta]} < \delta_{op}
\]

by which we mean that if \( H_0(\delta_{op}) \) is rejected (accepted), we, by convention, will declare \( NH_0 \) rejected (accepted). The authors’ choice of \( \delta_{op} \) depends on the degree of under coverage they are willing to tolerate. For example, if 88% is the minimum actual coverage they would tolerate for a 90% nominal two-sided Wald interval, then under normality they would select \( \delta_{op} = 0.3 \), as \( TC_{\alpha=0.1}(0.3) = 0.88 \) (see eq. (2.8) and Lemma 2.10).

Similarly, we have the operationalized pair

\[
NH_{0,k} : cBias_{\theta,k}(\hat{\psi}_1) = o_P(n^{-1/2})
\]

\[
H_{0,k}(\delta_{op}) : \frac{|cBias_{\theta,k}(\hat{\psi}_1)|}{\hat{\text{se}}[\hat{\psi}_1|\theta]} < \delta_{op}
\]

Suppose now the authors of a research paper agree that in reporting \( \hat{\psi}_1 \pm z_{\alpha/2}\hat{\text{se}}[\hat{\psi}_1] \) as a (1 − \( \alpha \)) confidence interval for \( \psi = \mathbb{E}_\theta[\text{cov}_\theta(Y,A|X)] \), their implicit or explicit null hypothesis is that the bias \( cBias_\theta(\hat{\psi}_1) \) of their estimator is \( o_P(n^{-1/2}) \). Further suppose the test \( \hat{\chi}_k^{(2)}(\Omega_k^{-1}; z_\omega, \delta_{op}) \) rejects \( H_{0,k}(\delta_{op}) \), equivalently \( NH_{0,k} \). However, unlike for \( \mathbb{E}_\theta[\text{var}_\theta(A|X)] \), rejecting \( NH_{0,k} \) does not logically imply rejecting \( NH_0 \).

What, if anything, can be done? One approach is to adopt an additional “faithfulness” assumption under which rejection of \( NH_{0,k} \) logically implies rejection of \( NH_0 \).

**Faithfulness Assumption.** Given a fixed \( k \), if \( cBias_{\theta,k}(\hat{\psi}_1) \) is not \( o_P(n^{-1/2}) \) then \( \frac{cTB_{\theta,k}(\hat{\psi}_1)}{cBias_{\theta,k}(\hat{\psi}_1)} \) is not \( o_P(1) \).

One might find this assumption rather natural because it holds unless \( cTB_{\theta,k}(\hat{\psi}_1) \) and \( cBias_{\theta,k}(\hat{\psi}_1) \) are of the same order and their leading constants sum to zero, which seems highly unlikely to be the case. However, this assumption is asymptotic and is without a clear finite sample operationalization, at least
to us. Therefore we shall consider an alternative assumption under which rejection of \( NH_{0,k} \) should, in our opinion, convince most authors to refuse to make assumption \( NH_0 \).

Cauchy-Schwarz Bias. We shall assume that the implicit or explicit goal in using a machine learning algorithm to predict the regression functions \( b(x) \) and \( p(x) \) is to construct predictors \( \hat{b}(x) \) and \( \hat{p}(x) \) that (nearly) minimize the conditional mean square errors \( \mathbb{E}_\theta[(b(X) - \hat{b}(X))^2|\theta] \) and \( \mathbb{E}_\theta[(p(X) - \hat{p}(X))^2|\theta] \) 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 confidence interval \( \hat{\psi}_1 \pm z_{\alpha/2}\hat{s}[\hat{\psi}_1] \) or \( \hat{\psi}_{\text{cross-fit},1}\pm z_{\alpha/2}\overline{s}[\hat{\psi}_{\text{cross-fit},1}] \), based on a DR-ML estimator \( \hat{\psi}_1 \) for \( \psi = \mathbb{E}_\theta[\text{cov}_\theta[A,Y|X]] \) should naturally appeal to the following Cauchy-Schwarz (CS) null hypothesis \( NH_{0,CS} \) and its operationalization \( H_{0,CS}(\delta_{op}) \):

\[
\begin{align*}
NH_{0,CS} : & \quad \text{CSBias}_\theta(\hat{\psi}_1) := \{|\mathbb{E}_\theta[(b(X) - \hat{b}(X))^2|\theta]|\mathbb{E}_\theta[(p(X) - \hat{p}(X))^2|\theta]|\}^{1/2} = o_{P_\theta}(n^{-1/2}), \\
H_{0,CS}(\delta_{op}) : & \quad \text{CSBias}_\theta(\hat{\psi}_1)/\text{sec}_\theta(\hat{\psi}_1) < \delta_{op}
\end{align*}
\]

(2.11)

\( H_{0,CS}(\delta_{op}) \) to support a validity claim that the interval’s actual coverage of \( \psi \) is (within the tolerance level set by \( \delta_{op} \)) nominal. The CS-null hypothesis \( NH_{0,CS} \) is the hypothesis that the Cauchy-Schwarz (CS) bias, \( \text{CSBias}_\theta(\hat{\psi}_1) \), is \( o_{P_\theta}(n^{-1/2}) \). Indeed, the truth of \( NH_{0,CS} \) does imply validity of the Wald intervals because, by the CS inequality, \( NH_{0,CS} \) implies \( NH_0 : c\text{Bias}_{\theta,k}(\hat{\psi}_1) \equiv \mathbb{E}_\theta[\{b(X) - \hat{b}(X)\}\{p(X) - \hat{p}(X)\}] = o_{P_\theta}(n^{-1/2}) \). However the converse is false: \( NH_0 \) may be true (and thus, by Theorem 1.3 the above the Wald intervals associated with \( \hat{\psi}_1 \) are valid) even when the CS-null hypothesis is false. But remember the goal in choosing to use a DR-ML estimator was precisely to minimize the CS bias. Thus if the CS-null hypothesis is rejected, although logically \( NH_0 \) may be true, there seems, to us, neither a substantive nor a philosophical reason to assume \( NH_0 \) to be true. In Bayesian language, our (subjective) posterior probability that \( NH_0 \) is true conditional on \( NH_{0,CS} \) being false is small. Thus we will make the following

CS Assumption. If the CS-null hypothesis \( NH_{0,CS} \) is false, one should refuse to support claims whose validity rests on the truth of \( NH_0 \); in particular, the claims that the intervals \( \hat{\psi}_1 \pm z_{\alpha/2}\hat{s}[\hat{\psi}_1] \) or \( \hat{\psi}_{\text{cross-fit},1}\pm z_{\alpha/2}\overline{s}[\hat{\psi}_{\text{cross-fit},1}] \) have actual coverage greater than or equal to their nominal.

Clearly the CS Assumption will allow meaningful inferences regarding \( \psi = \mathbb{E}_\theta[\text{cov}_\theta[A,Y|X]] \) only if it is possible to empirically reject the CS null hypothesis \( NH_{0,CS} \). In fact it is possible; indeed, rejection of \( NH_{0,k} : c\text{Bias}_{\theta,k}(\hat{\psi}_1) = o_{P_\theta}(n^{-1/2}) \) implies rejection of the CS null hypothesis \( NH_{0,CS} \). To see this define

\[
\text{CSBias}_{\theta,k}(\hat{\psi}_1) := \left\{ \mathbb{E}_\theta \left[ \left\{ \Pi[b(X) - \hat{b}(X)|\overline{Z}_k] \right\}^2 \mid \theta \right] \mathbb{E}_\theta \left[ \left\{ \Pi[p(X) - \hat{p}(X)|\overline{Z}_k] \right\}^2 \mid \theta \right] \right\}^{1/2}.
\]

We then have

\[
|c\text{Bias}_{\theta,k}(\hat{\psi}_1)| \leq \text{CSBias}_{\theta,k}(\hat{\psi}_1) \leq \text{CSBias}_\theta(\hat{\psi}_1),
\]

where we used the CS inequality followed by the fact that a projection never increases the \( L_2 \) norm.

Suppose we fail to reject the null hypothesis \( NH_{0,k} \). Is it still possible to reject the CS null hypothesis \( NH_{0,CS} \) by finding a direct empirical test of the hypothesis \( NH_{0,CS} \)? We next show the answer is yes. Because \( \text{CSBias}_\theta(\hat{\psi}_1) \) is not a smooth functional, we instead construct a higher order influence function test for the smooth functional \( \text{CSBias}_\theta^{(2)}(\hat{\psi}_1) \equiv \{ \text{CSBias}_\theta(\hat{\psi}_1) \}^2 \). Specifically we consider the operationalized pair

\[
\begin{align*}
NH_{0,k}^{(2)} & : \quad \text{CSBias}_\theta^{(2)} \equiv \mathbb{E}_\theta \left[ \left\{ b(X) - \hat{b}(X) \right\}^2 \mid \theta \right] \mathbb{E}_\theta \left[ \left\{ p(X) - \hat{p}(X) \right\}^2 \mid \theta \right] = o_{P_\theta}(n^{-1}), \\
H_{0,CS}^{(2)}(\delta_{op}) & : \quad \frac{\text{CSBias}_\theta^{(2)}}{\text{var}_\theta[\hat{\psi}_1|\theta]} < \delta_{op}^2.
\end{align*}
\]
Next define
\[ CSBias^{(2)}_{\theta,k} := E_\theta \left[ \left\{ \Pi[b(X) - \hat{b}(X)|Z_k] \right\}^2 | \theta \right] E_\theta \left[ \left\{ \Pi[p(X) - \hat{p}(X)|Z_k] \right\}^2 | \theta \right] \]
and the operationalized pair
\[ NH^{(2)}_{0,CS,k} : CSBias^{(2)}_{\theta,k} = o_P(n^{-1}) \]
\[ H^{(2)}_{0,CS,k}(\delta_{op}) : \frac{CSBias^{(2)}_{\theta,k}}{\text{var}_\theta[\psi_1|\theta]} < \delta_{op}^2. \]

Write \( CSBias^{(2)}_{\theta} = CSBias^{(2)}_{\theta,k} + CSTB^{(2)}_{\theta,k} \). Since \( CSBias^{(2)}_{\theta,k} \leq CSBias^{(2)}_{\theta} \) rejection of \( NH^{(2)}_{0,CS} \) and thus rejection of \( NH_{0,CS} \). We construct a conservative 1 - \( \omega \) level one-sided test \( \hat{\chi}_{CS,k}^{(2)}(\zeta_k, \delta) \) of \( H^{(2)}_{0,CS,k} \) based on the 4th order U-statistic
\[
\hat{\Pi}_{44,k} = \frac{(n - 4)!}{n!} \sum_{1 \leq i_1 \neq i_2 \neq i_3 \neq i_4 \leq n} \hat{IF}_{44,k,(i_1,i_2,i_3,i_4)}
\]
where
\[
\hat{IF}_{44,k,(i_1,i_2,i_3,i_4)} = \tilde{\varepsilon}_{i_1,i_2} \tilde{z}_k(X_{i_1})^\top \Omega_k^{-1} \tilde{z}_k(X_{i_2}) \tilde{\varepsilon}_{i_3,i_4} z_k(X_{i_3})^\top \Omega_k^{-1} z_k(X_{i_4}) z_{i_3,i_4}
\]
for which we have the following

**Theorem 2.12.** Under Condition \( W \), when \( k \) increases with \( n \) but \( k = o(n^2) \)

(i) Conditional on the training sample, \( \hat{\Pi}_{44,k} \) is unbiased for \( CSBias^{(2)}_{\theta,k} \) with variance of order \( k^2/n^4 \);

(ii) Conditional on the training sample, \( (n^4/k^2)^{1/2} \{ \hat{\Pi}_{44,k} - CSBias^{(2)}_{\theta,k} \} \) converges in law to a normal distribution with mean 0 and variance consistently estimated by \( (n^4/k^2) \hat{\text{var}}[\hat{\Pi}_{44,k}] \) with \( \hat{\text{var}}[\hat{\Pi}_{44,k}] \equiv \{ \hat{\varepsilon}[\hat{\Pi}_{44,k}] \}^2 \) defined analogously as in Theorem 2.5 or Appendix H.

(iii) \( \hat{\Pi}_{44,k} \pm \hat{\varepsilon}[\hat{\Pi}_{44,k}]\sqrt{2} / 2 \) is a conditional \( (1 - \omega) \) asymptotic conditional two-sided CI for \( CSBias^{(2)}_{\theta,k} \) with length of order \( (k^2/n^4)^{1/2} \).

Proof. Except for asymptotic normality, Theorem 2.12 follows from Robins et al. (2008). When \( k = o(n^2) \) and \( k \to \infty \) as \( n \to \infty \), the (conditional) asymptotic normality of
\[
\left( \frac{n^4}{k^2} \right)^{1/2} \left\{ \hat{\Pi}_{44,k} - CSBias^{(2)}_{\theta,k} \right\}
\]
follows directly from Bhattacharya and Ghosh (1992, Theorem 1.1) by verifying the conditions (A1) through (A5) therein under Condition \( W \).}

Based on the statistical properties of \( \hat{\Pi}_{44,k} \) summarized above, we now consider the properties of the following one-sided test \( \hat{\chi}_{CS,k}^{(2)}(\zeta_k, \delta_{op}) \) of the null hypothesis \( H^{(2)}_{0,CS,k}(\delta_{op}) : \frac{CSBias^{(2)}_{\theta,k}}{\text{var}_\theta[\psi_1|\theta]} < \delta_{op}^2 \):
\[
\hat{\chi}_{CS,k}^{(2)}(\zeta_k, \delta_{op}) := I \left\{ \frac{\hat{\Pi}_{44,k}}{\text{var}[\psi_1]} - \zeta_k \frac{\hat{\varepsilon}[\hat{\Pi}_{44,k}]}{\text{var}[\psi_1]} > \delta_{op}^2 \right\},
\]
for user-specified \( \zeta_k, \delta_{op} > 0 \). \( \zeta_k \) determines the level of the test \( \hat{\chi}_{CS,k}^{(2)}(\zeta_k, \delta_{op}) \).

The conditional asymptotic level and power of the oracle one-sided test \( \hat{\chi}_{CS,k}^{(2)}(\zeta_k, \delta_{op}) \) of \( H^{(2)}_{0,CS,k}(\delta_{op}) : \frac{CSBias^{(2)}_{\theta,k}}{\text{var}_\theta[\psi_1|\theta]} < \delta_{op}^2 \) are directly analogous to those of the one-sided test in Theorem 2.9. Specifically,

**Theorem 2.13.** Under Condition \( W \), when \( k \) increases with \( n \) but \( k = o(n) \), for any given \( \delta_{op}, \zeta_k > 0 \), suppose that
\[
\frac{CSBias^{(2)}_{\theta,k}}{\text{var}[\psi_1|\theta]} = \gamma \text{ for some (sequence) } \gamma = \gamma(n),
\]
then the rejection probability of \( \hat{\chi}_{CS,k}^{(2)}(\zeta_k, \delta_{op}) \)
converges to
\[
1 - \Phi(\zeta_k - \sqrt{\frac{n^2}{k^2} \vartheta(\gamma - \delta_{op}^2)}) \quad \text{where } \vartheta > 0 \text{ and } \vartheta := \sqrt{\frac{k^2}{n^2} \frac{\text{var}[\psi_1|\theta]}{\text{var}[\psi_1|\theta]}}, \text{ as } n \to \infty. \]
(1) under $H_{0,CS,k}^{(2)}(\delta_{op})$, $\hat{\chi}_{CS,k}^{(2)}(\zeta_k,\delta_{op})$ is an asymptotically conservative level $1 - \Phi(\zeta_k)$ one-sided test, and specifically $\hat{\chi}_{CS,k}^{(2)}(\zeta_\omega,\delta_{op})$ is an asymptotically conservative $\omega$ level one-sided test, as $n \to \infty$;

(2) under the alternative to $H_{0,CS,k}^{(2)}(\delta_{op})$, i.e. $\gamma \equiv \gamma(n) > \delta_{op}^2$

(i) if $\gamma - \delta_{op}^2 = C \sqrt{k \log(n)}$ for some $C > 0^2$, then $\hat{\chi}_{CS,k}^{(2)}(\zeta_k,\delta_{op})$ has rejection probability asymptotically equivalent to $1 - \Phi(\zeta_k - C\delta)$, as $n \to \infty$;

(ii) if $\gamma - \delta_{op}^2 = o(\sqrt{k \log(n)})$, then $\hat{\chi}_{CS,k}^{(2)}(\zeta_k,\delta_{op})$ has rejection probability asymptotically equivalent to the level of the test $1 - \Phi(\zeta_k)$, as $n \to \infty$;

(iii) if otherwise, then $\hat{\chi}_{CS,k}^{(2)}(\zeta_k,\delta_{op})$ has rejection probability converging to 1, as $n \to \infty$.

Remark 2.14 (A troubling unlikely scenario). Consider the troubling scenario that for all $k < n$, $\hat{\Pi}_{44,k}$ is strictly increasing in $k$ and $H_{0,CS,k}^{(2)}(\delta_{op})$ is, while $H_{0,k}(\delta_{op})$ is not, rejected at level $\omega$. Then NH$_{0,k}$ has been rejected, whereas NH$_{0,CS}$ has been accepted for every $k$. But recall that if NH$_{0,k}$ is ever rejected then so is NH$_{0,CS}$. Hence $H_{0,CS,k}^{(2)}(\delta_{op})$ and $H_{0,k}(\delta_{op})$ are in a sense in conflict at every $k$ in regards to the rejection of NH$_{0,CS}$. If we follow the CS assumption, we should not support claims that depend on the truth of NH$_0$ as NH$_{0,CS}$ has been rejected. On the other hand the failure of $H_{0,k}(\delta_{op})$ and thus NH$_{0,k}(\delta_{op})$ to be rejected at any $k$ begins to seem like reasonable evidence that perhaps NH$_0$ is true, especially when one recalls that rejection of NH$_{0,CS}$ does not logically imply rejection of NH$_0$. It is not clear to us what attitude one should take in this setting about the likely truth of the (empirically untestable) null hypothesis NH$_0$ of actual interest.

One might hope that this scenario is mathematically impossible, at least for $n$ sufficiently large, so it need not trouble us. But one can show that it can occur. However, in the limit as $n \to \infty$, the number of equality restrictions on the joint distribution that must prevail are so great that it is essentially impossible for it to occur in any real data set. Therefore it seems most prudent to continue to accept the CS assumption.

3. On the choice of estimators when $\Omega_k^{-1}$ is unknown

Heretofore we have assumed that $\Omega_k^{-1}$ is known. Outside the X-semisupervised case, this assumption is almost always untenable and $\Omega_k^{-1}$ must be estimated from data. To resolve this issue, one approach is to construct an estimator of the density $f_X$ of $X$ (Robins et al., 2008, 2017). But when dimension of the covariates $X$ is large, accurate density estimation is problematic. More recently, in the regime $k = o(n)$, Mukherjee et al. (2017) proposed to replace $\Omega_k$ by $\hat{\Omega}_k$ in $\hat{\Pi}_{22,k}(\Omega_k^{-1})$, where $\hat{\Omega}_k = n^{-1} \sum_{i \in \Omega_k} z_k(X_i) z_k(X_i)^\top$ is the sample covariance matrix estimator from the training sample. They show that $\hat{\Pi}_{22,k}(\hat{\Omega}_k^{-1})$ is a biased estimator of $cBias_{\theta_\psi}(\psi_1)$ with estimation bias $EB_{2,k}(\hat{\Omega}_k^{-1}) \equiv \mathbb{E}_0[\hat{\Pi}_{22,k}(\hat{\Omega}_k^{-1}) - cBias_{\theta_\psi}(\psi_1)(\theta)]$ of order $O(\theta_{\Pi} / \theta_{\Pi} \hat{\Pi}) \| \Pi \Pi \hat{\Pi} \|_{\theta} \sqrt{\theta_{\Pi} / \theta_{\Pi} \hat{\Pi} / n}$ under Condition W.

[Note $EB_{2,k}(\hat{\Omega}_k^{-1})$ is also the bias of $\hat{\psi}_{2,k}(\hat{\Omega}_k^{-1})$ as an estimator of $\hat{\psi}_k(\theta) = \psi(\theta) + cTB_{\theta_\psi}(\psi_1)$. It follows that the bias in estimating $cBias_{\theta_\psi}(\psi_1)$ converges to zero if $k \log(k) = o(n)$ and the projected $L_2(\mathbb{P}_0)$ errors $\Pi \hat{\Pi} \Pi \hat{\Pi} / \Psi \Pi \hat{\Pi} / \Psi$ in estimating $b$ and $p$ are bounded. However, we found in simulation that $\hat{\Pi}_{22,k}(\hat{\Omega}_k^{-1})$ has very unstable finite sample performance when $k$ is relatively large. For example, as shown in the second column of Table 3, $\hat{\Pi}_{22,k}(\hat{\Omega}_k^{-1})$ starts to break down even when $k = 512$, reflected by its MCav of estimated standard error being almost 8 times that of $\hat{\Pi}_{22,k}(\Omega_k^{-1})$. When $k = 1024$, $\hat{\Pi}_{22,k}(\hat{\Omega}_k^{-1})$ and its standard error are more than 1000 times those of $\hat{\Pi}_{22,k}(\Omega_k^{-1})$. This motivates us to find estimators that work better than $\hat{\Pi}_{22,k}(\hat{\Omega}_k^{-1})$ in practice.

3.1. An empirically stable estimator $\hat{\Pi}_{22,k}(\hat{\Omega}_{est}^{-1})$. A natural alternative estimator $\hat{\Pi}_{22,k}(\hat{\Omega}_{est}^{-1})$ is simply to replace $\hat{\Omega}_k$ by the sample covariance matrix $\hat{\Omega}_{est} := n^{-1} \sum_{i \in \Omega_{est}} z_k(X_i) z_k(X_i)^\top$ from the

---

2The local alternative regime $\gamma - \delta_{op}^2 \asymp \sqrt{k \log(n)}$ is a consequence of the variance of the statistic $\hat{\Omega}_{44,k} \var\psi(n_{\var\psi})$ of order $k^2 / n^2$.
estimation sample. Reading from Table 3, we see that \( \hat{IF}_{22,k}(\hat{\Omega}_{k}^{\text{est}})^{-1} \) and its standard error never blows up even for \( k = 2048 \) \((k/n \approx 0.8)\). In Appendix B, we discuss why \( \hat{IF}_{22,k}(\hat{\Omega}_{k}^{\text{est}})^{-1} \) does not suffer from the instability as \( \hat{IF}_{22,k}(\hat{\Omega}_{k}^{\text{tr}})^{-1} \) when \( k \) is large relative to \( n \). However, though numerically stable, the MCav of \( \hat{IF}_{22,k}(\hat{\Omega}_{k}^{\text{est}})^{-1} \), in contrast to that of the oracle \( \hat{IF}_{22,k}(\Omega_{k}^{-1}) \), is not monotone increasing in \( k \) and thus fails to correct nearly as much of the bias of \( \hat{\psi}_1 \) as does \( \hat{IF}_{22,k}(\Omega_{k}^{-1}) \). This can be seen from the third column of Table 3: the MCav of \( \hat{IF}_{22,k}(\hat{\Omega}_{k}^{\text{est}})^{-1} \) at \( k = 256 \) is 0.076 and close to that of \( \hat{IF}_{22,k}(\Omega_{k}^{-1}) \). However the MCav of \( \hat{IF}_{22,k}(\hat{\Omega}_{k}^{\text{est}})^{-1} \) decreases to 0.072 at \( k = 1024 \), while the MCav of \( \hat{IF}_{22,k}(\Omega_{k}^{-1}) \) continues to increase.

These numerical results raise the question whether we can find a stable estimator with MCav closer to that of the oracle \( \hat{IF}_{22,k}(\Omega_{k}^{-1}) \) than is the MCav of \( \hat{IF}_{22,k}(\hat{\Omega}_{k}^{\text{est}})^{-1} \).

We proceeded based on a theoretical analysis of the estimation bias of \( \hat{IF}_{22,k}(\hat{\Omega}_{k}^{\text{tr}})^{-1} \) as an estimator of \( cBias_{\theta,k}(\hat{\psi}_1) \) conditional on the training sample, which we refer to as \( EB_{2,k}(\hat{\Omega}_{k}^{\text{est}})^{-1} := \mathbb{E}_{\theta}[\hat{IF}_{22,k}(\hat{\Omega}_{k}^{\text{est}})^{-1} - cBias_{\theta,k}(\hat{\psi}_1) | \hat{\theta}] \).

As we describe now, this analysis led us to derive a de-biased version, \( \hat{IF}_{22,k}(\hat{\Omega}_{k}^{\text{est}})^{-1} \), defined below in eq. (3.2), of \( \hat{IF}_{22,k}(\hat{\Omega}_{k}^{\text{est}})^{-1} \), with estimation bias

\[
EB_{2,k}(\hat{\Omega}_{k}^{\text{est}})^{-1} := \mathbb{E}_{\theta}[\hat{IF}_{22,k}(\hat{\Omega}_{k}^{\text{est}})^{-1} - cBias_{\theta,k}(\hat{\psi}_1) | \hat{\theta}]
\]

under Condition W of order \( OP_{\theta}(||\Pi[\hat{\xi}_{\theta},Z_k]||_{\vartheta}||\Pi[\hat{\xi}_{p}|Z_k]||_{\vartheta,k\log(k)/n}) \), which is of smaller order than \( EB_{2,k}(\hat{\Omega}_{k}^{\text{tr}})^{-1} \). Specifically our derivation used the following identity (see Appendix C):

\[
EB_{2,k}(\hat{\Omega}_{k}^{\text{est}})^{-1} \equiv \mathbb{E}_{\theta}\left[\hat{IF}_{22,k}(\hat{\Omega}_{k}^{\text{est}})^{-1} - \hat{IF}_{22,k}(\Omega_{k}^{-1})\right] = \mathbb{E}_{\theta}\left[\hat{\xi}_{b,1}\overline{z}_{k}(X_1)^{\top}\left(\hat{\Omega}_{k,-1,-2}^{\text{est}} - \Omega_{k}^{-1}\right)\cdot \overline{z}_{k}(X_2)\hat{\xi}_{p,2}\right]
\]

\[
= \frac{1}{n}\mathbb{E}_{\theta}\left[\hat{\xi}_{b,1}\overline{z}_{k}(X_1)^{\top}\left(\hat{\Omega}_{k,-1,-2}^{\text{est}} - \Omega_{k}^{-1}\right) \cdot \overline{z}_{k}(X_1)\hat{\xi}_{k}^{\text{est}}(X_1)^{\top}\cdot \hat{\Omega}_{k}^{\text{est}} - \hat{\Omega}_{k}^{\text{est}} \cdot \hat{\Omega}_{k}^{\text{est}}\right] \cdot \overline{z}_{k}(X_2)\hat{\xi}_{p,2}\right]
\]

(3.1)

where for any \((i_1, i_2 : 1 \leq i_1 \neq i_2 \leq n)\), we define \( \hat{\Omega}_{k,-i_1,-i_2}^{\text{est}} := \frac{1}{n}\sum_{i \in \text{est} : i \neq i_1, i_2} \overline{z}_{k}(X_i)\overline{z}_{k}(X_i)^{\top} \).

Consider the first term (I) in the last line of the RHS of eq. (3.1). Due to the independence between the three product terms, we show in Appendix C.1 that we can can upper bound (I), up to constant, by

\[
\frac{1}{n^2(n-1)} \sum_{1 \leq i_1 \neq i_2 \leq n} \hat{\xi}_{b,1,1}\overline{z}_{k}(X_1)^{\top}\left(\hat{\Omega}_{k,-i_1,-i_2}^{\text{est}}\right)^{-1} \left(\sum_{i=i_1,i_2} \overline{z}_{k}(X_i)\overline{z}_{k}(X_i)^{\top}\right) \hat{\Omega}_{k}^{\text{est}}^{-1}\overline{z}_{k}(X_{i_2})\hat{\xi}_{p,i_2}.
\]

We cannot write \( \hat{IF}_{22,k}(\hat{\Omega}_{k}^{\text{est}})^{-1} \) in the form of \( \hat{IF}_{22,k}(\Omega_{k}^{-1}) \) because the bias correction on \( \hat{\Omega}_{k}^{\text{est}}^{-1} \) is not common for every pair of subjects in the summation \((i_1, i_2 : 1 \leq i_1 \neq i_2 \leq n)\) and this is reflected in the notation by attaching a superscript “debiased” on \( \hat{IF}_{22,k} \).
Remark 3.2. There are two other estimators of $cBias_{\theta,k}(\hat{\psi})$ that have the same order of estimation bias $O_P(\|\Pi[\hat{\xi}_b|Z_k]\|_2\|\Pi[\hat{\xi}_p|Z_k]\|_2^{k\log(k)}n)$ and variance $k/n^2$ as $EB_{2,k}^{\text{debiased}}([\hat{\Omega}_k^{est}]^{-1})$. The first adds the third order $U$-statistic

$$\hat{F}_{33,k}(\hat{\Omega}_k^{est})^{-1} := \frac{(n-3)^!}{n!} \sum_{1 \leq i_1 \neq i_2 \neq i_3 \leq n} \hat{\xi}_{b,i_1} Z_{k,i_1}^\top \hat{\Omega}_k^{est} \left\{ Z_{k,i_2} Z_{k,i_3}^\top - \hat{\Omega}_k^{est} \right\} \hat{\Omega}_k^{est}^{-1} Z_{k,i_2} \hat{\xi}_{p,i_3},$$

to $\hat{F}_{22,k}(\hat{\Omega}_k^{est})^{-1}$ to reduce the estimation bias $EB_{2,k}([\hat{\Omega}_k^{est}]^{-1})$.

- Specifically it follows from Mukherjee et al. (2017, Theorem 4) that $\hat{F}_{22,k}(\hat{\Omega}_k^{est})^{-1} + \hat{F}_{33,k}(\hat{\Omega}_k^{est})^{-1}$ also has the same order of estimation bias

$$EB_{3,k}(\hat{\Omega}_k^{est})^{-1} := \mathbb{E}_q \hat{F}_{22,k}(\hat{\Omega}_k^{est})^{-1} + \hat{F}_{33,k}(\hat{\Omega}_k^{est})^{-1} - cBias_{\theta,k}(\hat{\psi})|\hat{\theta}$$

as $EB_{2,k}^{\text{debiased}}([\hat{\Omega}_k^{est}]^{-1})^4$. Mukherjee et al. (2017, Theorem 4) also shows that the variance of $\hat{F}_{22,k}(\hat{\Omega}_k^{est})^{-1} + \hat{F}_{33,k}(\hat{\Omega}_k^{est})^{-1}$ is dominated by $\hat{F}_{22,k}(\hat{\Omega}_k^{est})^{-1}$ so $\hat{F}_{22,k}(\hat{\Omega}_k^{est})^{-1} + \hat{F}_{33,k}(\hat{\Omega}_k^{est})^{-1}$ has the same order of variance as $\hat{F}_{22,k}(\hat{\Omega}_k^{est})^{-1}$, $\hat{F}_{22,k}(\hat{\Omega}_k^{est})^{-1}$, and $\hat{F}_{22,k}(\hat{\Omega}_k^{est})^{-1}$. Unfortunately, the finite sample instability of $\hat{F}_{22,k}(\hat{\Omega}_k^{est})^{-1}$ cannot be resolved by correcting its estimation bias by adding $\hat{F}_{33,k}(\hat{\Omega}_k^{est})^{-1}$: when $\hat{F}_{22,k}(\hat{\Omega}_k^{est})^{-1}$ starts to break down at $k = 512$ (see of Table 3), $\hat{F}_{33,k}(\hat{\Omega}_k^{est})^{-1}$ also starts to break down (see the second column of Table 2).

- Another even simpler estimator $\hat{F}_{22,k}(\hat{\Omega}_k^{est})^{-1}$ of $cBias_{\theta,k}(\hat{\psi})$ achieves the same order of estimation bias $EB_{2,k}(\hat{\Omega}_k^{est})^{-1} = O_P(\|\Pi[\hat{\xi}_b|Z_k]\|_2 \cdot \|\Pi[\hat{\xi}_p|Z_k]\|_2 \cdot \frac{k\log(k)}{n})$ as $\hat{F}_{22,k}(\hat{\Omega}_k^{est})^{-1}$ can be constructed by splitting the whole sample into three parts: (1) training sample to estimate $\hat{F}_{33,k}(\hat{\Omega}_k^{est})^{-1}$, defined by replacing $\hat{\Omega}_k^{est}$ in $\hat{F}_{33,k}(\hat{\Omega}_k^{est})^{-1}$ with $\Omega_k^{-1}$, has exact mean 0 as estimation bias of $\hat{F}_{22,k}(\hat{\Omega}_k^{est})^{-1}$ is already zero. This is indeed reflected by the MCav of $\hat{F}_{33,k}(\hat{\Omega}_k^{est})^{-1}$ reported in the first column of Table 2 being close to zero.
nuisance functions \( b \) and \( p \); (2) covariance matrix sample to estimate \( \Omega_k \) by \( \hat{\Omega}_k^{\text{cov}} \), the empirical covariance matrix estimator from the covariance matrix sample; and (3) estimation sample to construct the (second-order) influence functions. However, in our simulation (not shown in this paper), \( \widehat{IF}_{22,k}(\hat{\Omega}_k^{\text{est}})^{-1} \) suffers from the same instability in its finite-sample performance as \( \widehat{IF}_{22,k}(\hat{\Omega}_k^{\text{tr}})^{-1} \).

**Remark 3.3.** Even though \( \widehat{IF}_{22,k}(\hat{\Omega}_k^{\text{est}})^{-1} \) has good theoretical guarantees on its estimation bias and variance, it is extremely difficult to compute it in practice. To compute \( \widehat{IF}_{22,k}(\hat{\Omega}_k^{\text{est}})^{-1} \), in the summation over all \( 1 \leq i_1 \neq i_2 \leq n \), we have to evaluate \( \hat{\Omega}_{k,-i_1,-i_2}^{\text{est}} \) for every different pairs \((i_1, i_2) : 1 \leq i_1 < i_2 \leq n \). Thus one needs to compute \( \binom{n}{2} \) different inverse sample covariance matrices when computing \( \widehat{IF}_{22,k}(\hat{\Omega}_k^{\text{est}})^{-1} \). Moreover, the kernel of \( \widehat{IF}_{22,k}(\hat{\Omega}_k^{\text{est}})^{-1} \) is no longer separable, this is because for every pair \((i_1, i_2) : 1 \leq i_1 < i_2 \leq n \), the kernel also depends on all the other subjects \( \{i \neq i_1, i_2 : 1 \leq i \leq n \} \). We will introduce a computationally-feasible estimator \( \widehat{IF}_{22,k}(\hat{\Omega}_k^{\text{quasi}})^{-1} \) in Section 3.2, which enjoys the stability of \( \widehat{IF}_{22,k}(\hat{\Omega}_k^{\text{est}})^{-1} \) and has smaller estimation bias in simulations than \( \widehat{IF}_{22,k}(\hat{\Omega}_k^{\text{est}})^{-1} \) as shown in the fourth column of Table 3.

3.2. **An easy-to-compute quasi de-biased estimator.** The estimator \( \widehat{IF}_{22,k}(\hat{\Omega}_k^{\text{quasi}})^{-1} \) differs from \( \widehat{IF}_{22,k}(\hat{\Omega}_k^{\text{est}})^{-1} \), only in that \( \hat{\Omega}_{k,-i_1,-i_2}^{\text{est}} \) in \( \widehat{IF}_{22,k}(\hat{\Omega}_k^{\text{est}})^{-1} \) is replaced by \( \hat{\Omega}_k^{\text{quasi}} \):

\[
\widehat{IF}_{22,k}(\hat{\Omega}_k^{\text{quasi}})^{-1} := \frac{1}{n(n-1)} \sum_{1 \leq i_1 \neq i_2 \leq n} \hat{\varepsilon}_{i_1} \hat{\varepsilon}_{i_2} \left( \frac{\hat{\Omega}_k^{\text{est}}}{n \hat{\Omega}_k^{\text{est}}} \right)^{-1} \sum \left( \frac{\hat{\varepsilon}_{i_1} \hat{\varepsilon}_{i_2}}{\hat{\Omega}_k^{\text{est}}} \right)
\]

where\(^5\)

\[
\hat{\Omega}_k^{\text{quasi}} := \hat{\Omega}_k^{\text{est}} - \frac{1}{n \hat{\Omega}_k^{\text{est}}} \sum \left( \frac{\hat{\varepsilon}_{i_1} \hat{\varepsilon}_{i_2}}{\hat{\Omega}_k^{\text{est}}} \right)
\]

In terms of finite sample performance, as shown in column 4 of Table 3 and column 2 of Table 4, \( \widehat{IF}_{22,k}(\hat{\Omega}_k^{\text{quasi}})^{-1} \) did not blow up numerically when \( k = 2048 \). Moreover, the MCav of \( \widehat{IF}_{22,k}(\hat{\Omega}_k^{\text{quasi}})^{-1} \) is closer to that of \( \widehat{IF}_{22,k}(\hat{\Omega}_k^{\text{tr}})^{-1} \) than are the MCavs of either \( \widehat{IF}_{22,k}(\hat{\Omega}_k^{\text{est}})^{-1} \) or \( \widehat{IF}_{22,k}(\hat{\Omega}_k^{\text{quasi}})^{-1} \) for all \( k \) in Table 3. For example, even when \( k = 1024 \), the MCav of \( \widehat{IF}_{22,k}(\hat{\Omega}_k^{\text{quasi}})^{-1} \) (0.119) is still closer to that of \( \widehat{IF}_{22,k}(\hat{\Omega}_k^{\text{est}})^{-1} \) (0.150), compared to the MCav of \( \widehat{IF}_{22,k}(\hat{\Omega}_k^{\text{tr}})^{-1} \) (0.073). Unfortunately, we have not been able to derive a satisfactory upper bound on the estimation bias of \( \widehat{IF}_{22,k}(\hat{\Omega}_k^{\text{quasi}})^{-1} \). However we do show in Appendix G that its variance is also of order \( k/n^2 \).

When \( k \) is relatively small compared to \( n \), we recommend using \( \widehat{IF}_{22,k}(\hat{\Omega}_k^{\text{quasi}})^{-1} \) instead of \( \widehat{IF}_{22,k}(\hat{\Omega}_k^{\text{tr}})^{-1} \) or \( \widehat{IF}_{22,k}(\hat{\Omega}_k^{\text{est}})^{-1} \) because, in our simulation, \( \widehat{IF}_{22,k}(\hat{\Omega}_k^{\text{quasi}})^{-1} \) is stable and has smaller estimation bias than \( \widehat{IF}_{22,k}(\hat{\Omega}_k^{\text{est}})^{-1} \). However, for larger \( k \), \( \widehat{IF}_{22,k}(\hat{\Omega}_k^{\text{est}})^{-1} \) performs poorly. Specifically, as shown in Table 4, the MCav of \( \widehat{IF}_{22,k}(\hat{\Omega}_k^{\text{est}})^{-1} \) decreased from 0.119 at \( k = 1024 \) to 0.061 at \( k = 2048 \), while the MCav of \( \widehat{IF}_{22,k}(\hat{\Omega}_k^{\text{tr}})^{-1} \) increased from 0.150 at \( k = 1024 \) to 0.191 at \( k = 2048 \), as expected for \( \psi = \mathbb{E}_{q}[\text{var}_{\theta}[A|X]] \).

3.3. **Shrinkage covariance matrix estimator.** In this section, we explore whether it is possible to find an estimator \( \hat{\Omega}_k^{\text{shrink}} \) for which the estimation bias of \( \widehat{IF}_{22,k}(\hat{\Omega}_k^{\text{shrink}})^{-1} \) remains small when \( k \) is near \( n \). As \( k \) gets close to \( n \), the relevant asymptotic regime is no longer \( k = o(n) \) but rather \( k/n \to c \) for some \( c \in (0,1) \) as \( n \to \infty \). This motivated us to try a non-linear shrinkage covariance matrix estimator proposed in Ledoit and Wolf (2004, 2012, 2017) for this latter asymptotic regime. In our simulations, we implemented the estimator \( \widehat{IF}_{22,k}(\hat{\Omega}_k^{\text{shrink}})^{-1} \), where \( \hat{\Omega}_k^{\text{shrink}} \) is the nonlinear shrinkage covariance matrix estimator \( \hat{\Omega}_k^{\text{shrink}} \)

\(^5\)Similar to \( \widehat{IF}_{22,k}(\hat{\Omega}_k^{\text{est}})^{-1} \), we cannot write \( \widehat{IF}_{22,k}(\hat{\Omega}_k^{\text{quasi}})^{-1} \) in the form of \( \widehat{IF}_{22,k}(\hat{\Omega}_k^{\text{tr}})^{-1} \) and this is reflected in the notation by attaching a superscript “quasi” on \( \widehat{IF}_{22,k} \).
Table 3. Simulation result for $E_{\theta} \left[ (A - p(X))^2 \right] \equiv 1$, with $\beta f_{X} = 0.4$

| $k$ | $\hat{\Omega}_{22,k}^{(\Omega^{-1})}$ | $\hat{\Omega}_{22,k}^{(\Omega^{-1})}$ | $\hat{\Omega}_{22,k}^{(\Omega^{-1})}$ | $\hat{\Omega}_{22,k}^{(\Omega^{-1})}$ | $\hat{\Omega}_{22,k}^{(\Omega^{-1})}$ |
|-----|--------------------------------|--------------------------------|--------------------------------|--------------------------------|--------------------------------|
| 0   | 0 (0)                          | 0 (0)                          | 0 (0)                          | 0 (0)                          | 0 (0)                          |
| 8   | 0.0063 (0.0034)                | 0.0066 (0.0036)                | 0.0062 (0.0034)                | 0.0063 (0.0034)                |                                |
| 256 | 0.081 (0.013)                  | 0.094 (0.015)                  | 0.076 (0.012)                  | 0.085 (0.012)                  |                                |
| 512 | 0.094 (0.023)                  | 0.155 (0.196)                  | 0.072 (0.021)                  | 0.102 (0.021)                  |                                |
| 1024| 0.150 (0.037)                  | -510.57 (62657.88)            | 0.073 (0.032)                  | 0.119 (0.032)                  |                                |
| 2048| 0.191 (0.062)                  | Blow up (Blow up)             | 0.031 (0.050)                  | 0.061 (0.050)                  |                                |

A comparison between $\hat{\Omega}_{22,k}^{(\Omega^{-1})}$, $\hat{\Omega}_{22,k}^{(\Omega^{-1})}$, and $\hat{\Omega}_{22,k}^{(\Omega^{-1})}$. The numbers in the parentheses are Monte Carlo average of estimated standard errors of the corresponding estimators. For more details on the data generating mechanism, see Appendix K.1.

Table 4. Simulation result for $E_{\theta} \left[ (A - p(X))^2 \right] \equiv 1$, with $\beta f_{X} = 0.4$

| $k$ | $\hat{\Omega}_{22,k}^{(\Omega^{-1})}$ | $\hat{\Omega}_{22,k}^{(\Omega^{-1})}$ | $\hat{\Omega}_{22,k}^{(\Omega^{-1})}$ | $\hat{\Omega}_{22,k}^{(\Omega^{-1})}$ |
|-----|--------------------------------|--------------------------------|--------------------------------|--------------------------------|
| 512 | 0.094 (0.023)                  | 0.102 (0.021)                  | 8390.68 (266082)               |                                |
| 1024| 0.150 (0.037)                  | 0.119 (0.032)                  | 0.154 (0.037)                  |                                |
| 2048| 0.191 (0.062)                  | 0.061 (0.050)                  | 0.204 (0.062)                  |                                |

A comparison between $\hat{\Omega}_{22,k}^{(\Omega^{-1})}$, $\hat{\Omega}_{22,k}^{(\Omega^{-1})}$, and $\hat{\Omega}_{22,k}^{(\Omega^{-1})}$. The numbers in the parentheses are Monte Carlo average of estimated standard errors of the corresponding estimators. For more details on the data generating mechanism, see Appendix K.1.

(Ledoit and Wolf, 2012, 2017) computed from the training sample data. $\hat{\Omega}_{22,k}^{(\Omega^{-1})}$ had very small estimation bias when $k$ is near $n$. In Table 4, even when $k = 2048$ and $n = 2500$ ($k/n \approx 0.8$), the MCav of $\hat{\Omega}_{22,k}^{(\Omega^{-1})}$ (0.204) is still quite close to that of $\hat{\Omega}_{22,k}^{(\Omega^{-1})}$ (0.191) whereas all the other estimators including $\hat{\Omega}_{22,k}^{(\Omega^{-1})}$ do not perform well.

However, $\hat{\Omega}_{22,k}^{(\Omega^{-1})}$ does not work when $k$ is small relative to $n$, as evidenced by the MCav of $\hat{\Omega}_{22,k}^{(\Omega^{-1})}$ (0.204) being more than $10^5$ times that of $\hat{\Omega}_{22,k}^{(\Omega^{-1})}$ when $k = 512$ in Table 4 (where $\hat{\Omega}_{22,k}^{(\Omega^{-1})}$ still performs well). This is not surprising given that the construction of $\hat{\Omega}_{22,k}^{(\Omega^{-1})}$ relies on the validity of the asymptotic characterizations of eigenvectors and eigenvalues of the sample covariance matrix under the asymptotic regime $k/n \to c$ for some $c \in (0, 1)$ as $n \to \infty$ (Ledoit and Peché, 2011; Ledoit and Wolf, 2012). Moreover, simulation results given in Appendix K.2 show that when the marginal density $f_{X}$ of $X$ is very rough, $\hat{\Omega}_{22,k}^{(\Omega^{-1})}$ can blow up even when $k$ is close to $n$. It is an open problem to theoretically explain the dependence of $\hat{\Omega}_{22,k}^{(\Omega^{-1})}$ on the smoothness of $f_{X}$. In fact since $\hat{\Omega}_{22,k}^{(\Omega^{-1})}$ is not consistent for $\Omega_{k}^{-1}$ under the regime $k/n \to c$ for some $c \in (0, 1)$, no theoretical results on the estimation bias of $\hat{\Omega}_{22,k}^{(\Omega^{-1})}$ are available.

Because of the limitations of $\hat{\Omega}_{22,k}^{(\Omega^{-1})}$ discussed here and in Section 3.2, we develop a data-adaptive estimator $\hat{\Omega}_{22,k}^{(\Omega^{-1})}$ in Appendix I to choose for each $k$, which of $\hat{\Omega}_{22,k}^{(\Omega^{-1})}$ or $\hat{\Omega}_{22,k}^{(\Omega^{-1})}$ should be used. Our procedure also selects a particular estimate $\hat{k}_{opt}$ for the true $k$ at which $\hat{\Omega}_{22,k}^{(\Omega^{-1})}$ most reduces the bias of $\hat{\psi}_{1}$. These procedures are motivated by our simulation studies, as the theoretical justifications are not available.

Remark 3.4 (Asymptotic normality when $\Omega_{k}^{-1}$ needs to be estimated). The conditional asymptotic normality of $\sqrt{n/k} \left[ \hat{\Omega}_{22,k}^{(\Omega^{-1})} - E_{\theta} \hat{\Omega}_{22,k}^{(\Omega^{-1})} \right]$ or $\sqrt{n/k} \left[ \hat{\Omega}_{22,k}^{(\Omega^{-1})} - E_{\theta} \hat{\Omega}_{22,k}^{(\Omega^{-1})} \right]$
follows from the same argument as the conditional asymptotic normality of \( \widehat{t}_{22,k}(\Omega_k^{-1}) \) because we can treat \( \widehat{\Omega}_k^{tr} \) or \( \widehat{\Omega}_k^{cov} \) as fixed by conditioning on the training sample or a third independent sample other than training sample/estimation sample. This argument would also imply the asymptotic normality of \( \sqrt{n}(\widehat{\Omega}_k^{shrink})^{-1} - E_\vartheta(\widehat{\Omega}_k^{shrink})^{-1} \) if we could prove that eigenvalues of \( \widehat{\Omega}_k^{shrink} \) are bounded with probability going to one.

We have yet to prove the conditional asymptotic normality of \( \sqrt{n}(\widehat{\Omega}_k^{quasi})^{-1} - E_\vartheta(\widehat{\Omega}_k^{quasi})^{-1} \) as \( k, n \to \infty \). However, based on the qqplots Figure 4 in Appendix L.2, we conjecture that \( \widehat{t}_{22,k}(\Omega_k^{shrink})^{-1} \) and \( \widehat{t}_{22,k}(\Omega_k^{est})^{-1} \) are both conditionally asymptotically normal as \( k, n \to \infty \).

### 3.4. Asymptotics of tests of \( H_{0,k}(\delta) \) (2.5) with unknown \( \Omega_k^{-1} \)

In previous sections we considered the estimation of \( cBias_{\theta,k}(\psi_1) \) when \( \Omega_k^{-1} \) is unknown. In this section we consider if estimating \( \Omega_k^{-1} \) has an effect on the statistical properties of the test of \( H_{0,k}(\delta) \) (2.5) and the upper confidence bound defined in Section 2. Since the results in this section hold for \( \psi = E_\vartheta[\text{var}[A,X]] \) and \( \psi = E_\vartheta[\text{cov}[A,Y,X]] \), we will drop the superscript and use \( \widehat{\chi}_k(\Omega_k^{-1};\zeta_k,\delta) \) (eq. (2.6) and eq. (2.7)) and \( UCB(\Omega_k^{-1};\alpha,\omega) \) (eq. (2.9) and eq. (2.10)) for the oracle test and upper confidence bound.

We now compare the asymptotic properties of the test \( \widehat{\chi}_k(\Omega_k^{-1};\zeta_k,\delta) \) and \( UCB(\Omega_k^{-1};\alpha,\omega) \) to those of \( \widehat{\chi}_k(\Omega_k^{-1};\zeta_k,\delta) \) and \( UCB(\Omega_k^{-1};\alpha,\omega) \). The following Proposition 3.5, proved in Appendix E, gives the sufficient conditions for the asymptotic equivalence between \( \widehat{\chi}_k(\Omega_k^{-1};\zeta_k,\delta) \) and \( \widehat{\chi}_k(\Omega_k^{-1};\zeta_k,\delta) \) for any arbitrary estimator \( \widehat{\Omega}_k^{-1} \).

**Proposition 3.5.** Under the conditions in Proposition 3.1, when \( k \) increases with \( n \) but \( k = o(n) \), for any given \( \delta, \zeta_k > 0 \). Suppose that \( \frac{cBias_{\theta,k}(\psi_1)}{se_\vartheta[\psi_1]} = \gamma \) for some sequence \( \gamma \equiv \gamma(n) \). If

\[
\frac{\widehat{t}_{22,k}(\Omega_k^{-1}) - E_\vartheta[\widehat{t}_{22,k}(\Omega_k^{-1})]}{se_\vartheta[\widehat{t}_{22,k}(\Omega_k^{-1})] \vartheta} \to N(0,1)
\]

and

\[
\frac{|EB_k(\Omega_k^{-1})|}{se_\vartheta[\psi_1]} = o_P\left( |\gamma|\sqrt{\frac{k}{n}} \right)
\]

where \( EB_k(\Omega_k^{-1}) := E_\vartheta[\widehat{t}_{22,k}(\Omega_k^{-1}) - cBias_{\theta,k}(\psi_1)] \), then the rejection probability of \( \widehat{\chi}_k(\Omega_k^{-1};\zeta_k,\delta) \) is asymptotically equal to that of \( \widehat{\chi}_k(\Omega_k^{-1};\zeta_k,\delta) \), as \( n \to \infty \). As a consequence, the asymptotic level and power of \( \widehat{\chi}_k(\Omega_k^{-1};\zeta_k,\delta) \) are the same as those of \( \widehat{\chi}_k(\Omega_k^{-1};\zeta_k,\delta) \) for testing \( H_{0,k}(\delta) \) (2.5).

**Remark 3.6.** In fact, under the alternative to \( H_{0,k}(\delta) \) (2.5), in the regime in Theorem 2.9(2.iii), for which the power of the oracle test converges to 1 as \( n \to \infty \), \( \frac{\widehat{t}_{22,k}(\Omega_k^{-1}) - E_\vartheta[\widehat{t}_{22,k}(\Omega_k^{-1})]}{se_\vartheta[\widehat{t}_{22,k}(\Omega_k^{-1})] \vartheta} = O_P(1) \), rather than convergence in law, suffices for the test \( \widehat{\chi}_k(\Omega_k^{-1};\zeta_k,\delta) \) to also have asymptotic power 1 (shown in Appendix E).

We next give sufficient conditions for the asymptotic equivalence between \( UCB(\Omega_k^{-1};\alpha,\omega) \) and \( UCB(\Omega_k^{-1};\alpha,\omega) \).

**Corollary 3.7.** Under the conditions of Proposition 3.5, suppose that \( \frac{cBias_{\theta,k}(\psi_1)}{se_\vartheta[\psi_1]} = \gamma \) for some fixed \( \gamma \), if

\[
\frac{|EB_k(\Omega_k^{-1})|}{se_\vartheta[\psi_1]} = o_P\left( |\gamma|\sqrt{\frac{k}{n}} \right)
\]

and

\[
\frac{\widehat{t}_{22,k}(\Omega_k^{-1}) - E_\vartheta[\widehat{t}_{22,k}(\Omega_k^{-1})]}{se_\vartheta[\widehat{t}_{22,k}(\Omega_k^{-1})] \vartheta} \to N(0,1)
\]
then, like $\text{UCB}(\hat{\Omega}_k^{-1}; \alpha, \omega)$, $\text{UCB}(\hat{\Omega}_k^{-1}; \alpha, \omega)$ is also a valid asymptotic $(1 - \omega)$ upper confidence bound for the actual asymptotic coverage of the $(1 - \alpha)$ Wald confidence interval centered at $\hat{\psi}_1$ for the parameter $\psi + cT_{\theta,k}(\hat{\psi}_1)$.

Next we consider the implication of Proposition 3.5 and Corollary 3.7 for the statistical properties of the tests and upper confidence bounds based on the estimators $\tilde{\Pi}_{22,k}(\hat{\Omega}_k^{-1})$, $\tilde{\Pi}_{22,k}(\hat{\Omega}_k^{\text{cov}})^{-1}$, $\tilde{\Pi}_{22,k}(\hat{\Omega}_k^{\text{est}})^{-1}$, and $\tilde{\Pi}_{22,k}(\hat{\Omega}_k^{\text{shrink}})^{-1}$ (more details in Appendix E). Our theoretical results are very limited for the following reason:

- The known upper bounds on the estimation bias $EB_{2,k}(\hat{\Omega}_k^{-1})$, such as the result in Proposition 3.1 on $EB_{2,k}(\hat{\Omega}_k^{\text{est}})^{-1}$, Remark 3.2 on $EB_{2,k}(\hat{\Omega}_k^{\text{cov}})^{-1}$, and Mukherjee et al. (2017, Theorem 4) on $EB_{2,k}(\hat{\Omega}_k^{\text{cov}})^{-1}$, are controlled in terms of $||\Pi[\hat{\xi}_b|Z_k]||_{\theta}||\Pi[\hat{\xi}_p|Z_k]||_{\theta}$, which equals $c\text{Bias}_{\theta,k}(\hat{\psi}_1)$ for the expected conditional variance but is greater than or equal to $c\text{Bias}_{\theta,k}(\hat{\psi}_1)$ for the expected conditional covariance. Thus to apply the sufficient conditions in Proposition 3.5 and Corollary 3.7, we need to make the additional restriction that $c\text{Bias}_{\theta,k}(\hat{\psi}_1) \neq o_P \left\{ ||\Pi[\hat{\xi}_b|Z_k]||_{\theta}||\Pi[\hat{\xi}_p|Z_k]||_{\theta} \right\}$

In Appendix E, we prove the following results on whether the conditions related to estimation bias in Proposition 3.5 is true for $\tilde{\Pi}_{22,k}(\hat{\Omega}_k^{\text{est}})^{-1}$, $\tilde{\Pi}_{22,k}(\hat{\Omega}_k^{\text{cov}})^{-1}$ and $\tilde{\Pi}_{22,k}(\hat{\Omega}_k^{\text{shrink}})^{-1}$:

**Lemma 3.8.** Under the conditions in Proposition 3.5, $k\log(k) = o(n)$, and the additional restriction that $c\text{Bias}_{\theta,k}(\hat{\psi}_1) \neq o_P \left\{ ||\Pi[\hat{\xi}_b|Z_k]||_{\theta}||\Pi[\hat{\xi}_p|Z_k]||_{\theta} \right\}$, suppose that $\frac{|c\text{Bias}_{\theta,k}(\hat{\psi}_1)|}{se[\hat{\psi}_1|\theta]} = \gamma$ for some (sequence) $\gamma \equiv \gamma(n)$, then we have

$$|EB_{2,k}(\hat{\Omega}_k^{-1})|_{se[\hat{\psi}_1|\theta]} = O_P \left( |\gamma|\sqrt{\frac{k\log(k)}{n}} \right) \neq o_P \left( |\gamma|\sqrt{\frac{k}{n}} \right)$$

but $|EB_{2,k}(\hat{\Omega}_k^{-1})|_{se[\hat{\psi}_1|\theta]} \neq |EB_{2,k}(\hat{\Omega}_k^{\text{cov}})^{-1})|_{se[\hat{\psi}_1|\theta]}$ and $|EB_{2,k}(\hat{\Omega}_k^{\text{shrink}})^{-1})|_{se[\hat{\psi}_1|\theta]}$ satisfy

$$O_P \left( |\gamma|\sqrt{\frac{k\log(k)}{n}} \right) = o_P \left( |\gamma|\sqrt{\frac{k}{n}} \right).$$

**Remark 3.9.** The restriction $c\text{Bias}_{\theta,k}(\hat{\psi}_1) \neq o_P \left\{ ||\Pi[\hat{\xi}_b|Z_k]||_{\theta}||\Pi[\hat{\xi}_p|Z_k]||_{\theta} \right\}$ will often hold when we estimate both nuisance functions from a single training sample. However when, as with $\hat{\psi}_{1,NR}$ discussed in Remark 2.8, we further split the training sample to estimate the nuisance functions from separate independent samples, it is possible to construct estimators with $c\text{Bias}_{\theta,k}(\hat{\psi}_1) = o_P \left\{ ||\Pi[\hat{\xi}_b|Z_k]||_{\theta}||\Pi[\hat{\xi}_p|Z_k]||_{\theta} \right\}$.

See the discussion in Remark 2.8 and more details can be found in Newey and Robins (2018).

Combining the results in Proposition 3.5, Lemma 3.8, and the asymptotic normalities of $\tilde{\Pi}_{22,k}(\hat{\Omega}_k^{\text{est}})^{-1}$ and $\tilde{\Pi}_{33,k}(\hat{\Omega}_k^{\text{cov}})^{-1}$ (Remark 3.4), we can conclude that, even under the additional restriction that $c\text{Bias}_{\theta,k}(\hat{\psi}_1) \neq o_P \left\{ ||\Pi[\hat{\xi}_b|Z_k]||_{\theta}||\Pi[\hat{\xi}_p|Z_k]||_{\theta} \right\}$ the test $\hat{\chi}_k(\hat{\Omega}_k^{-1}; \zeta_k, \delta)$ with $\tilde{\Pi}_{22,k}(\hat{\Omega}_k^{-1})$ replaced by $\tilde{\Pi}_{22,k}(\hat{\Omega}_k^{\text{est}})^{-1}$ may not have the correct asymptotic level and power as the sufficient condition on $EB_{2,k}(\hat{\Omega}_k^{\text{est}})^{-1}$ in Proposition 3.5 is not satisfied. Nonetheless, the rejection probability of the test $\hat{\chi}_k(\hat{\Omega}_k^{-1}; \zeta_k, \delta)$ with $\tilde{\Pi}_{22,k}(\hat{\Omega}_k^{-1})$ replaced by $\tilde{\Pi}_{22,k}(\hat{\Omega}_k^{\text{est}})^{-1} + \tilde{\Pi}_{33,k}(\hat{\Omega}_k^{\text{cov}})^{-1}$ or $\tilde{\Pi}_{22,k}(\hat{\Omega}_k^{\text{cov}})^{-1}$ is asymptotically equal to that of the oracle test $\hat{\chi}_k(\hat{\Omega}_k^{-1}; \zeta_k, \delta)$. By Corollary 3.7, we can also conclude that $UCB(\hat{\Omega}_k^{-1}; \alpha, \omega)$ and the oracle $UCB(\hat{\Omega}_k^{-1}; \alpha, \omega)$ are also asymptotically equivalent with $\tilde{\Pi}_{22,k}(\hat{\Omega}_k^{-1})$ replaced by $\tilde{\Pi}_{22,k}(\hat{\Omega}_k^{\text{est}})^{-1} + \tilde{\Pi}_{33,k}(\hat{\Omega}_k^{\text{cov}})^{-1}$ or $\tilde{\Pi}_{22,k}(\hat{\Omega}_k^{\text{cov}})^{-1}$.

---

6Recall that $EB_{3,k}(\hat{\Omega}_k^{\text{est}})^{-1}$ is the estimation bias of $\tilde{\Pi}_{22,k}(\hat{\Omega}_k^{\text{est}})^{-1} + \tilde{\Pi}_{33,k}(\hat{\Omega}_k^{\text{cov}})^{-1}$, defined in eq. (3.3).
The asymptotic normality of $\widehat{IF}^{\text{debiased}}_{22,k}(\hat{\Omega}^\text{est}_k)^{-1}$ is unknown, but we still have

$$\frac{\widehat{IF}^{\text{debiased}}_{22,k}(\hat{\Omega}^\text{est}_k)^{-1} - \mathbb{E}_\theta[\widehat{IF}^{\text{debiased}}_{22,k}(\hat{\Omega}^\text{est}_k)^{-1}] - \hat{\theta}}{\text{se}_\theta[\widehat{IF}^{\text{debiased}}_{22,k}(\hat{\Omega}^\text{est}_k)^{-1}] - \hat{\theta}} = O_P(1).$$

It then follows that the test $\hat{\chi}_k(\hat{\Omega}^{-1}_k; \zeta_k, \delta)$ with $\widehat{IF}^{\text{debiased}}_{22,k}(\hat{\Omega}^\text{est}_k)^{-1}$ replaced by $\widehat{IF}^{\text{debiased}}_{22,k}(\hat{\Omega}^\text{est}_k)^{-1}$, by Remark 3.6, still has rejection probability converging to 1 as $n \to \infty$ in the regime Theorem 2.9(2.iii).

For the estimators $\widehat{IF}^{\text{adi}}_{22,k}(\hat{\Omega}^\text{est}_k)^{-1}$ and $\widehat{IF}^{\text{shrink}}_{22,k}(\hat{\Omega}^\text{est}_k)^{-1}$ contributing to the data-adaptive estimator $\widehat{IF}^{\text{adi}}_{22,k}$ reported in the right panel of Table 1 (see Appendix I), we do not as yet obtained satisfactory theoretical results on the estimation bias or asymptotic normality.

3.5. Further results on estimation bias and testing. By Lemma 3.8, to test $H_{0,k}(\delta)$ (2.5) when $\Omega^{-1}_k$ is estimated from the training sample, we need to use $\widehat{IF}^{\text{debiased}}_{22\rightarrow33,k}(\hat{\Omega}^\text{tr}_k)^{-1} := \widehat{IF}^{\text{debiased}}_{22,k}(\hat{\Omega}^\text{tr}_k)^{-1} + \widehat{IF}^{\text{debiased}}_{33,k}(\hat{\Omega}^\text{tr}_k)^{-1}$ instead of $\widehat{IF}^{\text{debiased}}_{22,k}(\hat{\Omega}^\text{est}_k)^{-1}$ to estimate $cBias_{\theta,k}(\hat{\psi}_1)$, in order to ensure that the estimation bias is asymptotically negligible in terms of the level and power of the test $\hat{\chi}_k(\Omega^{-1}_k; \zeta_k, \delta)$.

Now suppose that at some $k = o(n)$, the hypothesis $H_{0,k}(\delta)$ (2.5) is clearly rejected by $\hat{\chi}_k(\hat{\Omega}^{-1}_k; \zeta_k, \delta)$ that uses $\widehat{IF}^{\text{debiased}}_{22\rightarrow33,k}(\hat{\Omega}^\text{tr}_k)^{-1}$. Having thus learned that $\hat{\psi}_1$ undercovers, we consider using the estimator $\hat{\psi}_{2,k}(\hat{\Omega}^\text{tr}_k)^{-1}$. However, before doing so, we would like to evaluate, like we did with $\hat{\psi}_1$, whether the Wald interval $\hat{\psi}_{2,k}(\hat{\Omega}^\text{tr}_k)^{-1} \pm z_{\alpha/2} \Sigma\hat{\psi}_{2,k}(\hat{\Omega}^\text{tr}_k)^{-1}$ undercovers. To do so, similar to Section 2, we can test the null hypothesis that the bias of $\hat{\psi}_{2,k}(\hat{\Omega}^\text{tr}_k)^{-1}$ as an estimator of $\mathbb{E}_\theta[\hat{\psi}_{2,k}(\Omega^{-1}_k)] = \psi + cTB_{\theta,k}(\psi_1)$ is smaller than a fraction $\delta$ of its standard error. To simplify the discussion, assume that $cTB_{\theta,k}(\psi_1) = 0$ so that $\mathbb{E}_\theta[\hat{\psi}_{2,k}(\Omega^{-1}_k)] = \psi$. In that case the bias of $\hat{\psi}_{2,k}(\hat{\Omega}^\text{tr}_k)^{-1}$ for $\psi$ is $EB_{2,k}(\hat{\Omega}^\text{tr}_k)^{-1} = \mathbb{E}_\theta[\hat{\psi}_{2,k}(\hat{\Omega}^\text{tr}_k)^{-1}] - \hat{\psi}_{2,k}(\hat{\Omega}^{-1}_k)$. Formally, we test the following null hypothesis:

$$H_{0,k}(\delta) : \frac{|cBias_{\theta,k}(\hat{\psi}_{2,k}(\hat{\Omega}^\text{tr}_k)^{-1}))|}{\text{se}_\theta[\hat{\psi}_{2,k}]} < \delta.$$  

where $cBias_{\theta,k}(\hat{\psi}_{2,k}(\hat{\Omega}^\text{tr}_k)^{-1})) := \mathbb{E}_\theta[\hat{\psi}_{2,k}(\hat{\Omega}^\text{tr}_k)^{-1}] - \psi = EB_{2,k}(\hat{\Omega}^\text{tr}_k)^{-1}$ when $cTB_{\theta,k}(\psi_1) = 0$. $EB_{2,k}(\hat{\Omega}^\text{tr}_k)^{-1}$ cannot be estimated because it depends on the unknown $cBias_{\theta,k}(\hat{\psi}_1)$. Nonetheless, $EB_{2,k}(\hat{\Omega}^\text{tr}_k)^{-1}$ can be approximated by estimated higher-order influence functions, where the true and estimated $m$-th order influence functions $cTB_{\theta,k}(\hat{\psi}_1) + \psi = \mathbb{E}_\theta[\text{cov}_\theta[A,Y|X]]$ are respectively $\widehat{IF}^{\text{mm,k}}_{22,k}(\hat{\Omega}^{-1}_k)$ and $\widehat{IF}^{\text{mm,k}}_{22,k}(\hat{\Omega}^{-1}_k)$. Here, for any $\Omega_k$

$$\widehat{IF}^{\text{mm,k}}_{22,k}(\hat{\Omega}^{-1}_k) := (-1)^n \frac{(n-m)!}{n!} \sum_{1 \leq i_1 \neq \ldots \neq i_m \leq n} \hat{\varepsilon}_{b,i_1} \hat{Z}_{k,i_1}^\top (\hat{\Omega}^{-1}_k) \prod_{j=3}^m \left\{ \left( \hat{Z}_{k,i_j} \hat{Z}_{k,i_j}^\top - \Omega_k \right) \left( \hat{\Omega}^{-1}_k \right) \right\} \hat{Z}_{k,i_2} \hat{\varepsilon}_{p,i_2}.$$

Define $\widehat{IF}^{\text{mm,k}}_{22\rightarrow33,k}(\hat{\Omega}^{-1}_k) := \sum_{j=1}^m \widehat{IF}^{\text{mm,k}}_{j,j,k}(\hat{\Omega}^{-1}_k)$ and also define the $m$-th order estimation bias to be $EB_{m,k}(\hat{\Omega}^{-1}_k) := \mathbb{E}_\theta[\widehat{IF}^{\text{mm,k}}_{22\rightarrow33,k}(\hat{\Omega}^{-1}_k) - cBias_{\theta,k}(\hat{\psi}_1)]$ for $m \geq 2$. Following Mukherjee et al. (2017), we have

**Lemma 3.10.**

1. $-\widehat{IF}^{\text{mm,k}}_{22\rightarrow33,k}(\hat{\Omega}^{-1}_k)$ is a biased estimator of $EB_{2,k}(\hat{\Omega}^{-1}_k)$, with conditional bias $-EB_{p,k}(\hat{\Omega}^{-1}_k)$;
2. More generally, $-\widehat{IF}^{\text{mm,k}}_{22\rightarrow33,k}(\hat{\Omega}^{-1}_k)$ is a biased estimator of $EB_{m-1,k}(\hat{\Omega}^{-1}_k)$, with conditional bias $-EB_{p,k}(\hat{\Omega}^{-1}_k)$. 
Proof. We only need to prove the second statement, which in turn implies the first statement. By definition the bias of \( \hat{\mathbb{P}}_{t \to m,k}(\hat{\Omega}_{k}^{tr}^{-1}) \) as an estimator of \( EB_{t-1,k}(\hat{\Omega}_{k}^{tr}^{-1}) \) is

\[
\mathbb{E}_{\theta}[\hat{\mathbb{P}}_{t \to m,k}(\hat{\Omega}_{k}^{tr}^{-1}) | \theta] - EB_{t-1,k}(\hat{\Omega}_{k}^{tr}^{-1})
= \mathbb{E}_{\theta}[\hat{\mathbb{P}}_{t \to m,k}(\hat{\Omega}_{k}^{tr}^{-1}) - (\hat{\mathbb{P}}_{22 \to t-1,t-1,k}(\hat{\Omega}_{k}^{tr}^{-1}) - cBias_{\theta,k}(\hat{\psi}_{1}) | \theta)]
= \mathbb{E}_{\theta}[\hat{\mathbb{P}}_{22 \to m,k}(\hat{\Omega}_{k}^{tr}^{-1}) + cBias_{\theta,k}(\hat{\psi}_{1}) | \theta]
\equiv - EB_{m,k}(\hat{\Omega}_{k}^{tr}^{-1})
\]

where we repeatedly use the definition of \( EB_{m,k}(\hat{\Omega}_{k}^{tr}^{-1}) \) for general \( m \geq 2 \).

Under Condition W, \( EB_{m,k}(\hat{\Omega}_{k}^{tr}^{-1}) = O_{p}(\|\Pi[\hat{\xi}_{0} | Z_{k}]\|_{\theta}|\Pi[\hat{\xi}_{0} | Z_{k}]\|_{\theta}(k \log(k) / n)^{m-1/2}) \) (Mukherjee et al., 2017). Hence to test \( \hat{H}_{0,2,k}(\delta) \), we can use \( -\hat{\mathbb{P}}_{33 \to pp,k}(\hat{\Omega}_{k}^{tr}^{-1}) \) to estimate \( EB_{2,k}(\hat{\Omega}_{k}^{tr}^{-1}) \) for some \( p \geq 3 \) and construct the following \( \omega \)-level test

\[
\hat{T}_{2,p,k} := I \left\{ \left| \frac{\hat{\mathbb{P}}_{33 \to pp,k}(\hat{\Omega}_{k}^{tr}^{-1})}{\hat{\mathbb{P}}_{33 \to pp,k}(\hat{\Omega}_{k}^{tr}^{-1})} - z_{\omega/2} \frac{\hat{\mathbb{P}}_{33 \to pp,k}(\hat{\Omega}_{k}^{tr}^{-1})}{\hat{\mathbb{P}}_{33 \to pp,k}(\hat{\Omega}_{k}^{tr}^{-1})} \right| > \delta \right\}.
\]

In particular, further suppose that \( EB_{m,k}(\hat{\Omega}_{k}^{tr}^{-1}) = O_{p}(\|\Pi[\hat{\xi}_{0} | Z_{k}]\|_{\theta}|\Pi[\hat{\xi}_{0} | Z_{k}]\|_{\theta}(k \log(k) / n)^{m-1/2}) \) is tight.

Then we show in Appendix F that the asymptotic level of \( \hat{T}_{2,p,k} \) is guaranteed to be \( \omega \) under \( \hat{H}_{0,2,k}(\delta) \) if we choose \( p \geq 5 \). Specifically, \( p \geq 5 \) ensures that the bias of \( \hat{\mathbb{P}}_{33 \to pp,k}(\hat{\Omega}_{k}^{tr}^{-1}) \) as an estimator of \( EB_{2,k}(\hat{\Omega}_{k}^{tr}^{-1}) \) does not affect the rejection probability of \( \hat{T}_{2,p,k} \) asymptotically. Moreover, increasing \( p \) does not change the asymptotic power of \( \hat{T}_{2,p,k} \) because the larger is \( p \), the smaller the order of the bias of \( \hat{\mathbb{P}}_{33 \to pp,k}(\hat{\Omega}_{k}^{tr}^{-1}) \) as an estimator of \( EB_{2,k}(\hat{\Omega}_{k}^{tr}^{-1}) \); furthermore, increasing \( p \) does not change the asymptotic variance of \( \hat{\mathbb{P}}_{33 \to pp,k}(\hat{\Omega}_{k}^{tr}^{-1}) \), as it is dominated by that of \( \hat{\mathbb{P}}_{33 \to pp,k}(\hat{\Omega}_{k}^{tr}^{-1}) \).

If \( \hat{T}_{2,p,k} \) with \( p \geq 5 \), rejects \( \hat{H}_{0,2,k}(\delta) \), we then consider using the estimator \( \hat{\psi}_{3,k}(\hat{\Omega}_{k}^{tr}^{-1}) := \hat{\psi}_{1} - \hat{\mathbb{P}}_{22 \to 33,k}(\hat{\Omega}_{k}^{tr}^{-1}) \). However, before doing so, we would like to test whether the Wald interval centered at \( \hat{\psi}_{3,k}(\hat{\Omega}_{k}^{tr}^{-1}) \) may undercover by testing whether the conditional bias of \( \hat{\psi}_{3,k}(\hat{\Omega}_{k}^{tr}^{-1}) \) is smaller than a fraction \( \delta \) of its standard error. We can indeed continue this process until the test of this hypothesis first fails to reject for some \( \hat{\psi}_{m,k}(\hat{\Omega}_{k}^{tr}^{-1}) := \hat{\psi}_{1} - \hat{\mathbb{P}}_{22 \to mm,k}(\hat{\Omega}_{k}^{tr}^{-1}) \), \( m \geq 2 \).

Specifically, we sequentially test the following null hypotheses for increasing \( m \)

\[
\hat{H}_{0,m,k}(\delta) : \frac{|cBias_{\theta}(\hat{\psi}_{m,k}(\hat{\Omega}_{k}^{tr}^{-1}))|}{\hat{\mathbb{P}}_{m,k}(\hat{\Omega}_{k}^{tr}^{-1})} = \frac{|EB_{m,k}(\hat{\Omega}_{k}^{tr}^{-1})|}{\hat{\mathbb{P}}_{m,k}(\hat{\Omega}_{k}^{tr}^{-1})} < \delta
\]

using the nominal \( \omega \)-level test

\[
\hat{T}_{m,p,k} := I \left\{ \left| \frac{\hat{\mathbb{P}}_{m+1,m+1 \to pp,k}(\hat{\Omega}_{k}^{tr}^{-1})}{\hat{\mathbb{P}}_{m+1,m+1 \to pp,k}(\hat{\Omega}_{k}^{tr}^{-1})} - z_{\omega/2} \frac{\hat{\mathbb{P}}_{m+1,m+1 \to pp,k}(\hat{\Omega}_{k}^{tr}^{-1})}{\hat{\mathbb{P}}_{m+1,m+1 \to pp,k}(\hat{\Omega}_{k}^{tr}^{-1})} \right| > \delta \right\}.
\]

Similar to \( \hat{T}_{2,p,k} \), \( \hat{T}_{m,p,k} \) is a conservative \( \omega \)-level test of \( \hat{H}_{0,m,k}(\delta) \) if \( p \geq 2m + 1 \) (see Appendix F) when \( EB_{m,k}(\hat{\Omega}_{k}^{tr}^{-1}) = O_{p}(\|\Pi[\hat{\xi}_{0} | Z_{k}]\|_{\theta}|\Pi[\hat{\xi}_{0} | Z_{k}]\|_{\theta}(k \log(k) / n)^{m-1/2}) \) is tight. Again further increasing \( p \) does not change the asymptotic power of \( \hat{T}_{m,p,k} \) as the order of the bias of \( \hat{\mathbb{P}}_{m+1,m+1 \to pp,k}(\hat{\Omega}_{k}^{tr}^{-1}) \) as an estimator of \( EB_{m,k}(\hat{\Omega}_{k}^{tr}^{-1}) \) is decreasing in \( p \) and the asymptotic variance of \( \hat{\mathbb{P}}_{m+1,m+1 \to pp,k}(\hat{\Omega}_{k}^{tr}^{-1}) \) is dominated by \( \hat{\mathbb{P}}_{m+1,m+1 \to pp,k}(\hat{\Omega}_{k}^{tr}^{-1}) \). In fact, by testing \( \hat{H}_{0,m,k}(\delta) \) sequentially, i.e. only testing \( \hat{H}_{0,m,k}(\delta) \) if \( \hat{H}_{0,m-1,k}(\delta) \) is rejected, together with the requirement that \( p \geq 2m + 1 \), we have the following (for proof, see Rosenbaum (2008, Proposition 1)):

**Proposition 3.11.** For \( m = 2, 3, \ldots \), consider the sequential testing procedure in which we sequentially test the hypotheses \( \hat{H}_{0,m,k}(\delta) \) with the nominal \( \omega \)-level test \( \hat{T}_{m,p,k} \) with \( p \geq 2m + 1 \). Let \( \hat{H}_{0,m_{\text{first}},k}(\delta) \) be the first hypothesis for which the test does not reject. We then accept the hypotheses \( \hat{H}_{0,m,k}(\delta) \) for \( m \geq m_{\text{first}} \).
and reject for \( m < m_{\text{first}} \). Then the rejection probability of any true hypothesis \( \hat{H}_{0,m,k}(\delta), m = 2, 3, \ldots \) is less than or equal to \( \omega \) as \( n \to \infty \).

4. Oracle tests when \( k > n \)

Although Theorem 2.5 derived the statistical properties of \( \bar{\mathbb{E}}_{22,k}(\Omega_k^{-1}) \) and \( \bar{\psi}_{2,k}(\Omega_k^{-1}) \) for \( k = o(n^2) \), we have been focused on the case \( k < n \). In this section, our focus is the case \( k > n \). We assume \( \Omega_k^{-1} \) is known and omit the dependence on \( \Omega_k^{-1} \) in the notation. We leave the unknown \( \Omega_k^{-1} \) case to a separate paper, because estimation of \( \Omega_k^{-1} \) with \( k > n \) requires additional assumptions that may not hold. Further, we restrict attention to the expected conditional variance \( \psi = \mathbb{E}_\theta[\text{var}_\theta[A|X]] \) as we will use the fact that the truncation bias \( cTB_{\theta,k}(\hat{\psi}_1) \) is non-increasing in \( k \) (see Lemma 2.2). Because of the close relationship between the statistical properties of tests and estimators of \( \psi = \mathbb{E}_\theta[\text{var}_\theta[A|X]] \) and of \( \text{CSBias}_\theta(\hat{\psi}_1) \) (see eq. (2.11)), results in this section apply essentially unchanged to the latter.

For some \( k_0 = o(n) \) with \( k_0 \) close to \( n \), e.g. \( k_0 = n/\log(n) \), we might quite generally prefer a Wald CI centered at \( \hat{\psi}_{2,k_0} \) rather than \( \hat{\psi}_1 \) because the bias of \( \hat{\psi}_{2,k_0} \) is no greater than that of \( \hat{\psi}_1 \) and their variances are close. If so we would like to test whether \( \hat{\psi}_{2,k} = \pm \frac{z}{2\sqrt{2}} \hat{\sigma} \hat{\psi}_{2,k} \) at \( k = k_0 \) covers \( \psi = \mathbb{E}_\theta[\text{var}_\theta[A|X]] \) at its nominal level by testing the null hypothesis that the ratio of the bias of \( \hat{\psi}_{2,k} \) to its standard error is smaller than a fraction \( \delta \). In Section 4.1 we construct such a test for any \( k = o(n^2) \). In Remark 4.1, we explain heuristically why we restrict \( k \) to be no greater in order than \( n^2 \) and why no consistent test can be constructed. Remark 1.1 and Remark 2.7 consider these same issues but less heuristically. In addition, we develop a sequential multiple testing procedure that tests the above null hypothesis at level \( \omega \) for each of \( J \) different \( k \) with \( k_0 < k_1 < \cdots < k_{j-1} < n^2 \) with \( k_{j-1} < k_j = o(n^2) \) for some \( k_j \). Our procedure tests this null hypothesis sequentially beginning with \( k_0 \) and stops at the first \( k_j \) for which the test does not reject and the accepts the null for \( k_{j+1}, \cdots, k_{j-1} \). In Section 4.3, we also show that the sequential procedure protects the \( \omega \)-level of all \( J \) tests.

**Remark 4.1** (Why the truncation bias cannot be estimated). Recall \( \hat{\psi}_{2,k} \) has mean \( \psi + cTB_{\theta,k}(\hat{\psi}_1) \), upward bias \( cTB_{\theta,k}(\hat{\psi}_1)(\geq 0) \) and standard error of order \( \max\{k, n\}/n^2 \). Figure 2 will perform double duty in this section. Specifically the same vertical bars will represent different statistics - depending on which y-axis label is under discussion. Figure 2, when associated with the right y-axis label “\( \hat{\psi}_{2,k} \)”, shows nominal 95% Wald intervals \( \hat{\psi}_{2,k} = 1.96 \hat{\sigma} \hat{\psi}_{2,k} \) for various \( k \). The x-axis is the unknown \( \psi = \mathbb{E}_\theta[\text{var}_\theta[A|X]] \), which, of course, does not vary with \( k \). The centers \( \hat{\psi}_{2,k} \) of the intervals are decreasing in \( k \) but, in another sample from the same distribution, this need not be the case. However, the unobserved means \( \psi + cTB_{\theta,k}(\hat{\psi}_1) \) and bias \( cTB_{\theta,k}(\hat{\psi}_1) \) must be non-increasing in \( k \). As expected, the lengths of the intervals are increasing with \( k \) and the length of the final interval at \( k \) of order \( n^2 \) has standard error of order 1. Further the bias \( cTB_{k=n^2,\theta}(\hat{\psi}_1) \) at \( k = n^2 \) must also be order 1, as indicated by the length of the distance from the x-axis to the bottom of the confidence interval. Consider an alternative scenario in which the graph is unchanged but \( \psi \) is moved below the x axis by a distance equal to 4 times the distance from the dotted line to the current x-axis. Then the unknown truncation bias \( cTB_{\theta,k}(\hat{\psi}_1) \) in the second scenario is over 4 times that in the first, yet the data and its analysis would remain unchanged. Trying larger \( k \), \( k \gg n^2 \), would not help because the standard error of \( \hat{\psi}_{2,k} \) is of order much greater than a truncation bias of order 1 (see Remark 2.7).

To formalize the proposed testing procedure, we start with testing the following null hypothesis, at \( k = k_0 \):

\[
H_{0,2,k}(\delta) : \frac{cBias_\theta(\hat{\psi}_{2,k})}{\text{se}_\theta(\hat{\psi}_{2,k}|\theta)} < \delta,
\]

(4.1)
where \( cBias_\theta(\hat{\psi}_{2,k}) \) is the conditional bias of \( \hat{\psi}_{2,k} \) and equals the truncation bias of \( \hat{\psi}_1 \) at \( k^8 \):

\[
(4.2) \quad cBias_\theta(\hat{\psi}_{2,k}) := \mathbb{E}_\theta[\hat{\psi}_{2,k} - \psi|\theta] = \mathbb{E}_\theta[\hat{\psi}_1 - \psi - \hat{\mu}_{22,k}] = cBias_\theta(\hat{\psi}_1) - cBias_\theta, k(\hat{\psi}_1) = cTB_\theta, k(\hat{\psi}_1).
\]

\( H_{0,2,k}(\delta) \) (4.1) is analogous to \( H_0(\delta) \) (2.4) considered in Section 2.

We will use Figure 2, now associated with the left y-axis label, to illustrate how to test \( H_{0,2,k}(\delta) \) (4.1). We assume that even \( \hat{\psi}_{2,n^2} \) (not shown in the figure) has a bias of order 1. We are given an ordered set \( K = \{k_0 < k_1 < k_2 < \cdots < k_j : k_0 = o(n), k_j < k_{j+1}, j = 0, \cdots, J - 1, k_j = o(n^2)\}^9 \) of candidate \( k \)’s. Thus there are \( J + 1 \) candidate \( k \)’s in Figure 2.

• We now use the upper panel of Figure 2 to explain our test of \( H_{0,2,k_0}(\delta) \) (4.1). The y-value of the point at the bottom of the vertical ‘error’ bar corresponding to \( k_j \) is \( y_j = \frac{\hat{\psi}_{2,k_j}}{se[\hat{\psi}_{2,k_j}]} - \delta/2 \) for \( j = 0, 1, \cdots, J \). The length of the error bar of \( k_j \) is \( z_{\omega/J} \frac{se[\hat{\mu}_{22,k_j} - \mu_{22,k_j}]}{se[\hat{\psi}_{2,k_j}]} \) (and thus 0 for \( k_0 \)) with the upper most point being \( \frac{\hat{\psi}_{2,k_j}}{se[\hat{\psi}_{2,k_j}]} - \delta/2 + z_{\omega/(J-1)} \frac{se[\hat{\psi}_{2,k_j} - \psi]}{se[\hat{\psi}_{2,k_j}]} \). The point \( y_0 \) associated with \( k_0 \) is the blue point at the left. If it lies outside at least one of the error bars to its right, we reject \( H_{0,2,k_0}(\delta) \) (4.1). We use \( z_{\omega/J} \) in the definition of error bars to adjust for the \( J \) multiple comparisons. As shown in the figure, we reject \( H_{0,2,k_0}(\delta) \) (4.1) because the blue point \( \frac{\hat{\psi}_{2,k_0}}{se[\hat{\psi}_{2,k_0}]} - \delta/2 \) is outside the error bar at \( k_{J-1} \) (purple).

• After \( H_{0,2,k_0}(\delta) \) (4.1) is rejected, naturally we would like to test \( H_{0,2,k_1}(\delta) \) (4.1), as shown in the middle panel of Figure 2: Since the bias of \( \hat{\psi}_{2,k_1} \) is no greater than \( \hat{\psi}_{2,k_0} \) and the variance of \( \hat{\psi}_{2,k_1} \) is in general greater than that of \( \hat{\psi}_{2,k_0} \), we would expect that the actual asymptotic coverage of the Wald CI associated with \( \hat{\psi}_{2,k_1} \) should be closer to its nominal coverage. To test \( H_{0,2,k_1}(\delta) \) (4.1), we follow exactly the same procedure as for \( H_{0,2,k_0}(\delta) \) (4.1). In the middle panel, the upper end of the error bars for different \( k_j \) are defined to be \( \frac{\hat{\psi}_{2,k_j}}{se[\hat{\psi}_{2,k_j}]} - \delta/2 + z_{\omega/(J-1)} \frac{se[\hat{\mu}_{22,k_j} - \mu_{22,k_j}]}{se[\hat{\psi}_{2,k_j}]} \) for \( j = 2, \cdots, J \). When \( \frac{\hat{\psi}_{2,k_1}}{se[\hat{\psi}_{2,k_1}]} - \delta/2 \) (the leftmost green point) lies outside at least one of the error bars to its right, we reject \( H_{0,2,k_1}(\delta) \) (4.1). We use \( z_{\omega/(J-1)} \) in the definition of error bars to adjust for the \( J - 1 \) total comparisons. We find \( H_{0,2,k_1}(\delta) \) (4.1) is rejected because the green point \( \frac{\hat{\psi}_{2,k_1}}{se[\hat{\psi}_{2,k_1}]} - \delta/2 \) lies outside the error bar at \( k_{J-1} \) (purple).

• As \( H_{0,2,k_1}(\delta) \) (4.1) is rejected, we next test \( H_{0,2,k_2}(\delta) \) (4.1) as in the lower panel. The upper end of the error bars for different estimators are defined as \( \frac{\hat{\psi}_{2,k_j}}{se[\hat{\psi}_{2,k_j}]} - \frac{\hat{\psi}_{2,k_j}}{se[\hat{\psi}_{2,k_j}]} - \delta/2 + z_{\omega/(J-2)} \frac{se[\hat{\psi}_{2,k_j} - \psi]}{se[\hat{\psi}_{2,k_j}]} \) for \( j = 3, \cdots, J \). We fail to reject \( H_{0,2,k_2}(\delta) \) (4.1) because \( \frac{\hat{\psi}_{2,k_2}}{se[\hat{\psi}_{2,k_2}]} - \delta/2 \) (the leftmost black point) is covered by all the error bars to its right.

• We then terminate the testing procedure and accept \( H_{0,2,k}(\delta) \) (4.1) at \( k_2 \) for \( k \geq k_2 \).

The above sequential bias testing algorithm has the following features.

1. The algorithm tests the hypotheses \( H_{0,2,k_1}(\delta) \) (4.1) sequentially at the first \( k_j \) for which \( H_{0,2,k_j}(\delta) \) is not rejected and accepts \( H_{0,2,k_1}(\delta) \) and the remaining untested hypotheses.

2. For each null hypothesis of interest \( H_{0,2,k}(\delta) \) (4.1) for \( k \in K \), we perform \( J - k \) tests and reject \( H_{0,2,k}(\delta) \) (4.1) if at least one of the tests reject.

---

8One can thus rephrase \( H_{0,2,k}(\delta) \) (2.5) as testing if the truncation bias of \( \hat{\psi}_1 \) at \( k \) is smaller than a fraction \( \delta \) of its standard error.

9For example, if the range of candidate \( k \)’s is between \( n \) and \( n^2 \), when \( J \) is a fixed integer, then in general \( k_j \) is not only less than \( k_{j+1} \), but also \( k_j = o(k_{j+1}) \); when \( J = \log_2(n) \), then we can choose \( k_j = \frac{1}{2} k_{j+1} \) and \( k_j \) is the closest integer greater than \( n \) and a power of 2. We still require \( k = o(n^2) \) because the asymptotic normality of \( \hat{\mu}_{22,k} \) does not hold if \( k \) is of order \( n^2 \) or greater.
Below we analyze the *sequential bias testing algorithm* in detail. In particular, we address the following three questions:

(1) In Section 4.1 we answer: Why does the *sequential bias testing algorithm* test each $H_{0,2,k}(\delta)$ (4.1) using the particular test it does?

(2) In Section 4.2 we answer: Why does *sequential bias testing algorithm* perform multiple comparisons when testing each $H_{0,2,k}(\delta)$ (4.1)?

(3) In Section 4.3 we answer: Why does the *sequential bias testing algorithm* test $H_{0,2,k_j+1}(\delta)$ (4.1) only if it rejects $H_{0,2,k_j}(\delta)$ (4.1)?

4.1. **Answer to question (1).** The answer to the first question mirrors the development of the test $\hat{\chi}_k^{(1)}(\Omega_k^{-1}; \zeta_k, \delta)$ (eq. (2.6)) in Section 2. As $H_0(\delta)$ (2.4) cannot be tested consistently, neither can $H_{0,2,k}(\delta)$ (4.1) be tested consistently because $cBias_\theta(\hat{\psi}_{2,k})$ cannot be estimated from data. Instead we can test

\[(4.3) \quad H_{0,2,k \rightarrow k'}(\delta) : \frac{cBias_\theta(k', \hat{\psi}_{2,k})}{\text{se}[\hat{\psi}_{2,k}]|} < \delta \]

where $cBias_\theta(k', \hat{\psi}_{2,k}) := cBias_\theta(\hat{\psi}_{2,k}) - cBias_\theta(\hat{\psi}_{2,k'}) = E_\theta[\hat{\Pi}_{22,k'} - \hat{\Pi}_{22,k}]$ is the estimable part of the bias of $\hat{\psi}_{2,k}$. Similar to Lemma 2.2, we have the following (logical) orderings on $cBias_\theta(k', \hat{\psi}_{2,k})$ and $H_{0,2,k \rightarrow k'}(\delta)$ (4.3) for $\psi = E_\theta[\text{var}_I[A'X]]$:

**Lemma 4.2.** For the parameter $\psi = E_\theta[\text{var}_I[A'X]]$:

1. Given $k < k_1 < k_2$, $cBias_{\theta,k_1}(\hat{\psi}_{2,k}) \leq cBias_{\theta,k_2}(\hat{\psi}_{2,k}) \leq cBias_\theta(\hat{\psi}_{2,k})$;
2. Given $k_1 < k_2 < k'$, $cBias_{\theta,k_1}(\hat{\psi}_{2,k_2}) \leq cBias_{\theta,k_2}(\hat{\psi}_{2,k_2})$;
3. Given $k < k'$, if $H_{0,2,k}(\delta)$ (4.1) is true, then $H_{0,2,k \rightarrow k'}(\delta)$ (4.3) is true; if $H_{0,2,k \rightarrow k'}(\delta)$ (4.3) is false, then $H_{0,2,k}(\delta)$ (4.1) is false.

**Proof.** (1) - (3) directly follows from $cBias_{\theta,k_1}(\hat{\psi}_{2,k}) = E_\theta[\hat{\Pi}_{22,k'} - \hat{\Pi}_{22,k}]$ and the larger $k'$ (or the smaller $k$), the larger the difference $E_\theta[\hat{\Pi}_{22,k'} - \hat{\Pi}_{22,k}]$. \hfill \Box

By Lemma 4.2(2), $H_{0,2,k \rightarrow k'}(\delta)$ (4.3) plays the same role for $H_{0,2,k}(\delta)$ (4.1) as $H_0(\delta)$ (2.5) for $H_0(\delta)$ (2.4) of Section 2.

The comparison between the leftmost point and every error bar to its right in each panel of Figure 2 is equivalent to test $H_{0,2,k_j \rightarrow k'}(\delta)$ (4.3) by the following test with $\zeta_{k \rightarrow k'} = z_{\omega/(J-j)}$ when $k = k_j$ and $k' > k_j$, $k_j, k' \in K$:

\[(4.4) \quad \hat{\chi}_{2,k \rightarrow k'}(\zeta_{k \rightarrow k'}, \delta) \equiv \hat{\chi}_{2,k \rightarrow k'}(\Omega_k^{-1}, \Omega_k^{-1}; \zeta_{k \rightarrow k'}, \delta) = I \left\{ \frac{\hat{\Pi}_{22,k'} - \hat{\Pi}_{22,k}}{\text{se}[\hat{\psi}_{2,k}]|} - \zeta_{k \rightarrow k'} \frac{\text{se}[\hat{\Pi}_{22,k'} - \hat{\Pi}_{22,k}]}{\text{se}[\hat{\psi}_{2,k}]|} > \delta \right\} \]

where $\hat{\psi}_{2,k} - \hat{\psi}_{2,k'} \equiv \hat{\Pi}_{22,k'} - \hat{\Pi}_{22,k}$. Similar to Theorem 2.9 of Section 2, $\hat{\chi}_{2,k \rightarrow k'}(\zeta_{k \rightarrow k'}, \delta)$ enjoys asymptotic properties summarized in Proposition 4.3 below, as a test of $H_{0,2,k \rightarrow k'}(\delta)$ (4.3) for $k, k' \in K$ and $k < k'$:

**Proposition 4.3.** Under *Condition W*, when $k', k$ increases with $n$, $k' > k$, and $k'^2$, for any given $\delta, \zeta_{k \rightarrow k'} > 0$, suppose that $\frac{cBias_{\theta,k}(\hat{\psi}_{2,k})}{\text{se}[\psi[k] \psi]} = \gamma$ for some (sequence) $\gamma = \gamma(n)$, then the rejection probability of $\hat{\chi}_{2,k \rightarrow k'}(\zeta_{k \rightarrow k'}, \delta)$ converges to $1 - \Phi(\zeta_{k \rightarrow k'} - \sqrt{\max[k(n)] \frac{k'}{k}} \vartheta(\gamma - \delta))$ where $\vartheta > 0$ and $\sqrt{\max[k(n)] \frac{k'}{k}} \text{se}[\hat{\Pi}_{22,k'} - \hat{\Pi}_{22,k}] \rightarrow \vartheta$, as $n \rightarrow \infty$. In particular,

1. under $H_{0,2,k \rightarrow k'}(\delta)$ (4.3), $\hat{\chi}_{2,k \rightarrow k'}(\zeta_{k \rightarrow k'}, \delta)$ is an asymptotically conservative level $1 - \Phi(\zeta_{k \rightarrow k'})$ one-sided test, as $n \rightarrow \infty$;
2. under the alternative to $H_{0,2,k \rightarrow k'}(\delta)$ (4.3), i.e. $\gamma \equiv \gamma(n) > \delta$, \hfill \Box
(i) if $\gamma - \delta = C \sqrt{\frac{k'}{\max(k,n)}}$ for some $C > 0$, then $\hat{\chi}_{2,k \rightarrow k'}(\zeta_{k \rightarrow k'}, \delta)$ has rejection probability asymptotically equivalent to $1 - \Phi(\zeta_{k \rightarrow k'} - C\vartheta)$, as $n \rightarrow \infty$;

(ii) if $\gamma - \delta = o(\sqrt{\frac{k'}{\max(k,n)}})$, then $\hat{\chi}_{2,k \rightarrow k'}(\zeta_{k \rightarrow k'}, \delta)$ has rejection probability asymptotically equivalent to the level of the test $1 - \Phi(\zeta_{k \rightarrow k'})$, as $n \rightarrow \infty$;

(iii) if otherwise, then $\hat{\chi}_{2,k \rightarrow k'}(\zeta_{k \rightarrow k'}, \delta)$ has rejection probability converging to 1, as $n \rightarrow \infty$.

**Remark 4.4.** When $k' > k > n$, the local alternative mentioned in Proposition 4.3(2) becomes increasing with $k'$. This is also made clear in Figure 2: in each panel, the error bar (of order $\sqrt{\frac{k'}{k}}$) becomes wider as $k' = k_j$ increases with $j$. Thus to reject a comparison with probability approaching 1 with a larger $k'$, we need a larger decrease in bias from $\hat{\varphi}_{2,k}$ to $\hat{\varphi}_{2,k'}$. This can lead to non-monotonic rejection probabilities with increasing $k'$: even though $H_{0,2,k \rightarrow k_j}(\delta)$ (4.3) being logically implies $H_{0,2,k \rightarrow k_{j+1}}(\delta)$ (4.3) being false, it is possible to reject $H_{0,2,k \rightarrow k_j}(\delta)$ (4.3) with probability approaching 1 but fail to reject $H_{0,2,k \rightarrow k_{j+1}}(\delta)$ (4.3) with probability strictly less than 1, simply because the local alternative also increases from $k_j$ to $k_{j+1}$. For instance, in the upper pane of Figure 2, we reject $H_{0,2,k_0 \rightarrow k_{j-1}}(\delta)$ (4.3) (the purple error bar) but fail to reject $H_{0,2,k_0 \rightarrow k_j}(\delta)$ (4.3) (the rightmost red error bar).

We have addressed the first question – the visual comparison described in Figure 2 corresponds to the test $\hat{\chi}_{2,k \rightarrow k'}(\zeta_{k \rightarrow k'}, \delta)$ of $H_{0,2,k \rightarrow k'}(\delta)$ (4.3), whose asymptotic properties are summarized in Proposition 4.3.

**4.2. Answer to question (2).** As explained in Remark 4.4, for $k_j > k_{j-1} > k$, when the hypothesis of interest $H_{0,2,k}(\delta)$ (4.1) and the actually tested hypothesis $H_{0,2,k \rightarrow k_{j-1}}(\delta)$ (4.3) are false, it is possible to reject $H_{0,2,k \rightarrow k_{j-1}}(\delta)$ (4.3) with probability approaching 1, but fail to reject $H_{0,2,k \rightarrow k_j}(\delta)$ (4.3) with positive probability even though the latter being false is logically implied by the former, because of the increasing local alternatives. Such incoherence is portrayed in the upper panel of Figure 2, where we succeed in rejecting $H_{0,2,k_0}(\delta)$ at $k_{j-1}$ but fail to do so at $k_j$, because the increase in variance from $k_{j-1}$ to $k_j$ overshadows the decrease in bias.

For the above reason, we need to conduct multiple comparisons, i.e. test hypotheses $H_{0,2,k \rightarrow k'}(\delta)$ (4.3) for all $k' \in \mathcal{K}$ and $k' > k$ in order to improve the chance of rejecting $H_{0,2,k}(\delta)$ (4.1) when it is indeed false.

Consequently we should adjust for multiple comparisons when testing $H_{0,2,k}(\delta)$ (4.1) given $k \in \mathcal{K}$. When $J$ is bounded, by the joint normality of $\{Z_{22,k_j}, j = 0, 1, \cdots, J\}$, Proposition 4.3(1) implies that we can simply choose $\zeta_{k_j \rightarrow k'} = \omega/(J-j)$ to insure that each test $\hat{\chi}_{k_j \rightarrow k'}(\zeta_{k_j \rightarrow k'}, \delta)$ is an asymptotic $\omega/(J-j)$ level test of $H_{0,2,k_j \rightarrow k'}(\delta)$ (4.3) for $k' \in \mathcal{K}$ and $k' > k_j$. Then under $H_{0,2,k_j}(\delta)$ (4.1) (which implies $H_{0,2,k_j \rightarrow k'}(\delta)$ (4.3) for all $k' \in \mathcal{K}$ and $k' > k_j$ by Lemma 4.2(2)), the level of the joint tests is guaranteed to be $\omega$. We summarize the above arguments in a more general form in Corollary 4.5(1) below:

**Corollary 4.5.** Suppose that $J$ is bounded.

1. Given $k_j \in \mathcal{K}$, under $H_{0,2,k_j}(\delta)$ (4.1), the joint tests $\hat{\chi}_{2,k_j \rightarrow k'}(\omega/(J-j), \delta)$ for all $k' \in \mathcal{K}$ and $k' > k_j$ have asymptotic level $\omega$ as $n \rightarrow \infty$;

2. Given $k_j \in \mathcal{K}$, if there is a subset $\mathcal{K}'$ of $\mathcal{K} \setminus \{k_0, \ldots, k_j\}$ such that for $k' \in \mathcal{K}'$, $H_{0,2,k_j \rightarrow k'}(\delta)$ (4.3) is false and $\frac{cBias_{\theta, k}(\hat{\varphi}_{2,k_j})}{st\hat{\varphi}_{2,k_j}} = \gamma$ for some (sequence) $\gamma \equiv \gamma(n) > \delta$, then the rejection probability of the joint tests $\hat{\chi}_{2,k_j \rightarrow k'}(\omega/(J-j), \delta)$ is greater than or equal to $\max\{1 - \Phi(\omega/(J-j) - \sqrt{\frac{\max(k,n)}{k}}\vartheta(\gamma - \delta)), k' \in \mathcal{K}'\}$ as $n \rightarrow \infty$, where $\vartheta$ is defined in Proposition 4.3(2). In particular, if there is any $k' \in \mathcal{K}'$ such that $\gamma - \delta$ is of order greater than $\sqrt{\frac{k'}{\max(k,n)}}$, then the rejection probability of the joint tests converges to 1 as $n \rightarrow \infty$.

---

10The local alternative regime $\gamma - \delta \asymp \sqrt{\frac{k'}{\max(k,n)}}$ is a consequence of the variance of the statistic $\frac{\hat{\varphi}_{2,k} - \bar{\varphi}_{2,k}}{\hat{\varphi}_{2,k}}$ is of order $k'/\max(k,n)$. 
Corollary 4.5(2) summarizes the asymptotic power of the joint tests \( \hat{\chi}_{2,k_j} \to k' (z_{\omega/(J-j)}, \delta) \) for all \( k' \in \mathcal{K} \) and \( k' > k_j \). When any one of the hypotheses \( H_{0,2,k_j \to k'} (\delta) \) is false and outside the regimes given in Proposition 4.3(2.i, 2.ii), then the joint tests have probability approaching 1 to reject \( H_{0,2,k_j} (\delta) \) for any \( k_j \in \mathcal{K} \).

4.3 Answer to question (3). Finally, we explain why the hypotheses \( H_{0,2,k_j} (\delta) \) for \( k_j \in \mathcal{K} \) are sequentially tested (or tested in order). To simplify our argument, we focus on \( H_{0,2,k_0} (\delta) \) and \( H_{0,2,k_1} (\delta) \) only. To test \( H_{0,2,k_0} (\delta) \), we are actually testing \( H_{0,2,k_0 \to k_1} (\delta) : \frac{cBias_{\theta,k_1} (\hat{\psi}_{2,k_0})}{se[\hat{\psi}_{2,k_0}]} < \delta \) (4.3) using \( \hat{\chi}_{2,k_0 \to k_1} (z_{0 \to k_1}, \delta) \) for \( j = 1, \ldots, J \). Suppose all the actually tested hypotheses associated with \( H_{0,2,k_0} (\delta) \), \( H_{0,2,k_0 \to k_1} (\delta) \) for \( j = 1, \ldots, J \), are true, then logically so are all the actually tested hypotheses \( H_{0,2,k_1 \to k_2 \to k_1} (\delta) : \frac{cBias_{\theta,k_1} (\hat{\psi}_{2,k_1})}{se[\hat{\psi}_{2,k_1}]} < \delta \) associated with \( H_{0,2,k_1} (\delta) \). To see this, for the same \( k_j \), \( cBias_{\theta,k_j} (\hat{\psi}_{2,k_0}) \geq cBias_{\theta,k_j} (\hat{\psi}_{2,k_1}) \). This follows because for any \( k' > k_j \), \( cBias_{\theta,k'} (\hat{\psi}_{2,k}) \) is the expected difference between \( \hat{\chi}_{2,k_0 \to k_j} (z_{0 \to k_j}, \delta) \) for \( j = 1, \ldots, J \). Suppose all the actually tested hypotheses \( H_{0,2,k_0} (\delta) \), \( H_{0,2,k_0 \to k_1} (\delta) \) true because \( \frac{cBias_{\theta,k_1} (\hat{\psi}_{2,k_1})}{se[\hat{\psi}_{2,k_1}]} < \delta \) under \( H_{0,2,k_0 \to k_1} (\delta) \) (4.3). It is straightforward to generalize to any \( k_{\ell-1} < k_\ell \).

Lemma 4.6. For every \( j = \ell, \ldots, J \), suppose that \( se[\hat{\psi}_{2,k_{\ell-1}}] \leq se[\hat{\psi}_{2,k_\ell}] \). If \( H_{0,2,k_{\ell-1} \to k_\ell} (\delta) \) is true, then \( H_{0,2,k_{\ell-1} \to k_\ell} (\delta) \) is also true for every \( j = \ell + 1, \ldots, J \).

Such (indirect) logical ordering between \( H_{0,2,k_0} (\delta) \) and \( H_{0,2,k_1} (\delta) \) or more generally between \( H_{0,2,k_1} (\delta) \) and \( H_{0,2,k_2} (\delta) \) for any \( k_{\ell-1} < k_\ell \) gives rise to the strategy of testing \( H_{0,2,k_\ell} (\delta) \) only if \( H_{0,2,k_{\ell-1}} (\delta) \) is rejected. We now describe the sequential bias testing algorithm abstractly: given (1) the ordered set of candidate k’s \( \mathcal{K} = \{ k_0 < k_1 < k_2 < \cdots < k_j : k_0 = o(n), k_j < k_{j+1}, j = 0, \ldots, J-1, j \text{ or } k_{\ell-1} < k_\ell \} \), (2) the corresponding set of estimators \( \{ \hat{\chi}[1], \hat{\chi} [2,k], \hat{\chi} [2,k] \} : k, k' \in \mathcal{K}, k' > k \} \) as described in Theorem 1.3, Theorem 2.5 and Appendix H, (3) the corresponding set of cutoffs \( \{ \zeta_{0,k} : k' > k, k' \in \mathcal{K} \} \) and (4) the corresponding set of desired levels of the tests \( \{ 0 < \omega_k \leq 1/2 : k \in \mathcal{K} \} \):

- For \( j = 0, \ldots, J-1 \):
  - At \( k = k_j \), test the null hypothesis of interest \( H_{0,2,k_j} (\delta) \) by jointly testing multiple hypotheses \( H_{0,2,k_j \to k'} (\delta) \) using the test \( \hat{\chi}_{k_j \to k'} (z_{\omega/(J-j)}, \delta) \), for all \( k' \in \mathcal{K} \) such that \( k' > k_j \).
    * For \( k' \in \mathcal{K} \) such that \( k' > k_j \), if any one of \( H_{0,2,k_j \to k'} (\delta) \) is rejected, then \( k = k_{j+1} \) and continue the iteration.
    * Otherwise, we report that we cannot reject \( H_{0,2,k_j} (\delta) \) based on the available information and stop the iteration at \( k = k_j \).

Because the hypotheses \( H_{0,2,k_j} (\delta) \) for \( k_j \in \mathcal{K} \) are sequentially tested, we immediately have the following result on the level of the sequential bias testing algorithm for each null hypothesis of interest \( H_{0,2,k_j} (\delta) \) (4.1). The proof is straightforward and can be found in Rosenbaum (2008, Proposition 1).

Proposition 4.7. For every \( k \in \mathcal{K} \), the sequential bias testing algorithm is an asymptotic \( \omega_k \) level test of the null hypothesis of interest \( H_{0,2,k} (\delta) \) (4.1).

Remark 4.8. We now make some further comments on the sequential bias testing algorithm:

1. The power of the sequential bias testing algorithm is more involved because of the local alternatives. For any \( k \in \mathcal{K} \) such that \( H_{0,2,k} (\delta) \) is false, and at least one of its associated multiple hypotheses \( H_{0,2,k \to k'} (\delta) \) for \( k' > k \), is false and not in or below the corresponding local alternatives (Proposition 4.3(2.ii)), then the sequential bias testing algorithm rejects \( H_{0,2,k} (\delta) \) with probability converging to 1 as \( n \to \infty \). We leave the asymptotic power under more complicated scenario to future work.
(2) Crucial to Corollary 4.5 is the assumption that $J$ is bounded. When $J$ grows with $n$, the joint asymptotic normality of $\{\hat{\Pi}_{22,k}, k \in K\}$ is still open. In principle, regardless of whether $J$ is bounded or not, we could also use recent developments on the exponential tail bound with explicit constants on second-order degenerate $U$-statistics in Houdré and Reynaud-Bouret (2003, Theorem 3.1) or in Giné and Nickl (2016, Theorem 3.4.8) for bounded kernel or in Chakrabortty and Kuchibhotla (2018, Theorem 1) for unbounded sub-Weibull kernel (Kuchibhotla and Chakrabortty, 2018). However, because $\hat{\Pi}_{22,k'} - \hat{\Pi}_{22,k}$ is not an exact second-order degenerate $U$-statistic, to use such exponential tail bounds for hypothesis testing purposes requires a more careful analysis to obtain constants that can be estimated from data. Results along this line can be found in Mukherjee et al. (2016); Mukherjee and Sen (2018) but more careful analysis is needed to obtain explicit constants.

(3) It is possible to refine the strategy for multiple testing adjustment considered above, in which we distribute the desired overall type-I error uniformly to every $J - j$ actually tested hypotheses associated with the hypothesis of actual interest $H_{0,2,k_j}(\delta)$ (4.1). Instead we can prioritize certain actually tested hypotheses over others. Such non-uniform strategy also appeared in Spokoiny and Vial (2009) in the context of adaptive estimation in Gaussian sequence model.

In this section, we sketched how we can generalize the testing idea developed in Section 2 on the DR-ML estimator $\hat{\psi}_1$ to $\hat{\psi}_{2,k}$ for $k = o(n^2)$. One open problem is how to construct confidence bound on the actual asymptotic coverage from the sequential bias testing algorithm. Furthermore, in the case where we do not have $X$-semisupervised data, to make such strategy applicable, we need to estimate the density $f_X$ of $X$ in moderate/high dimensions or estimate $\Omega^{-1}_k$ with $k > n$ when the sample covariance matrix is singular. Recent literature (Liang, 2018) on using generative adversarial networks (GAN) to estimate probability density functions could be useful for estimating high-dimensional density $f_X$ when $k > n$.

5. Open problems

We conclude by mentioning some important open questions.

- Will the good finite-sample performance of $\hat{\Pi}^{\text{quasi}}_{22,k}([\hat{\Omega}^\text{est}_k]^{-1})$, $\hat{\Pi}^{\text{quasi}}_{22,k}([\hat{\Omega}^{\text{shrink}}_k]^{-1})$ and, especially, $\hat{\Pi}^{\text{adapt}}_{22,k}$ continue to be seen under diverse simulation settings yet to be studied? If yes, what are the theoretical reasons for their success?
- Can we find estimators of $c_{\text{Bias},k}(\hat{\psi}_1)$ that outperform $\hat{\Pi}^{\text{quasi}}_{22,k}([\hat{\Omega}^\text{est}_k]^{-1})$, $\hat{\Pi}^{\text{quasi}}_{22,k}([\hat{\Omega}^{\text{shrink}}_k]^{-1})$ and $\hat{\Pi}^{\text{adapt}}_{22,k}$? In fact, since our actual interest is in a statistical functional of the inverse covariance matrix $\Omega^{-1}_k$ and not $\Omega^{-1}_k$ itself, could we find improved estimators of $c_{\text{Bias},k}(\hat{\psi}_1)$ that do not begin by estimating $\Omega^{-1}_k$ but rather attempt to directly estimate the functional of interest?
- Can we improve on our current inference strategy for parameters for which the truncation bias is not a monotone in $k$?
- Can we find near optimal methods to select the basis functions $Z_k$ out of a dictionary $V$?
Fig. 2. An illustration of the sequential bias testing algorithm.

Depicted is a hypothetical data generating mechanism in which sequential bias testing algorithm rejects both $H_{0,2,k_0}(\delta)$ (4.1) and $H_{0,2,k_1}(\delta)$ (4.1) but fails to reject $H_{0,2,k_2}(\delta)$ (4.1). In each panel (upper/middle/lower), $H_{0,2,k_j}(\delta)$ (4.1) is tested by checking whether $\hat{\psi}_{2,k_j}/\hat{se}[\hat{\psi}_{2,k_j}] - \delta/2$ (the leftmost point in each panel) is outside any of the error bars to its right ($j = 0/1/2$). The error bars are defined in the main text.

Acknowledgements

We would like to thank Dr. Thomas M. Kolokotrones, Dr. Linbo Wang, Professor Michael Wolf for valuable discussions.
REFERENCES

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: a system for large-scale machine learning. In *OSDI*, volume 16, pages 265–283, 2016.

Miguel A Arcones and Evarist Giné. On the bootstrap of U and V statistics. *The Annals of Statistics*, pages 655–674, 1992.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *arXiv preprint arXiv:1610.01271*, 2016.

Rajeev Ayyagari. Applications of influence functions to semiparametric regression models. PhD thesis, Harvard University, 2010.

Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

Alexandre Belloni, Daniel Chen, Victor Chernozhukov, and Christian Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429, 2012.

Alexandre Belloni, Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, 186(2):345–366, 2015.

Rabi N Bhattacharya and Jayanta K Ghosh. A class of U-statistics and asymptotic normality of the number of k-clusters. *Journal of Multivariate Analysis*, 43(2):300–330, 1992.

Peter J Bickel and David A Freedman. Some asymptotic theory for the bootstrap. *The Annals of Statistics*, pages 1196–1217, 1981.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Leo Breiman. *Classification and regression trees*. Routledge, 2017.

Abhishek Chakrabortty and Tianxi Cai. Efficient and adaptive linear regression in semi-supervised settings. *The Annals of Statistics*, 46(4):1541–1572, 2018.

Abhishek Chakrabortty and Arun Kumar Kuchibhotla. Tail bounds for canonical U-statistics and U-processes with unbounded kernels. 2018.

O. Chapelle, B. Schölkopf, and A. Zien. *Semi-Supervised Learning*. Adaptive computation and machine learning. MIT Press, 2010.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018a.

Victor Chernozhukov, Whitney K Newey, and Rahul Singh. Learning L2 continuous regression functionals via regularized riesz representers. *arXiv preprint arXiv:1809.05224*, 2018b.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

Max H Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference: Application to causal effects and other semiparametric estimands. *arXiv preprint arXiv:1809.09953*, 2018.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*, volume 40. Cambridge University Press, 2016.

Wolfgang Härdle, Gerard Kerkyacharian, Dominique Picard, and Alexander Tsybakov. *Wavelets, approximation, and statistical applications*, volume 129. Springer Science & Business Media, 1998.

Christian Houdré and Patricia Reynaud-Bouret. Exponential inequalities, with constants, for U-statistics of order two. In *Stochastic Inequalities and Applications*, pages 55–69. Springer, 2003.

Marie Huskova and Paul Janssen. Consistency of the generalized bootstrap for degenerate U-statistics. *The Annals of Statistics*, pages 1811–1823, 1993.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

Arun Kumar Kuchibhotla and Abhishek Chakrabortty. Moving beyond sub-Gaussianity in high-dimensional statistics: Applications in covariance estimation and linear regression. *arXiv preprint*
Olivier Ledoit and Sandrine Péché. Eigenvectors of some large sample covariance matrix ensembles. *Probability Theory and Related Fields*, 151(1-2):233–264, 2011.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

Olivier Ledoit and Michael Wolf. Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics*, 40(2):1024–1060, 2012.

Olivier Ledoit and Michael Wolf. Numerical implementation of the QuEST function. *Computational Statistics & Data Analysis*, 115:199–223, 2017.

Lingling Li, Eric Tchetgen Tchetgen, Aad van der Vaart, and James Robins. Robust inference with higher order influence functions: Part II. In *Joint Statistical Meetings, Minneapolis, Minnesota*, 2005.

Tengyuan Liang. On how well generative adversarial networks learn densities: Nonparametric and parametric results. *arXiv preprint arXiv:1811.03179*, 2018.

Stéphane Mallat. *A wavelet tour of signal processing*. Elsevier, 1999.

Rajarshi Mukherjee and Subhabrata Sen. Optimal adaptive inference in random design binary regression. *Bernoulli*, 24(1):699–739, 2018.

Rajarshi Mukherjee, Eric Tchetgen Tchetgen, and James Robins. On adaptive estimation of nonparametric functionals. *arXiv preprint arXiv:1608.01364*, 2016.

Rajarshi Mukherjee, Whitney K. Newey, and James M. Robins. Semiparametric efficient empirical higher order influence function estimators. *arXiv preprint arXiv:1705.07577*, 2017.

Whitney K Newey. Semiparametric efficiency bounds. *Journal of Applied Econometrics*, 5(2):99–135, 1990.

Whitney K Newey and James M Robins. Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv preprint arXiv:1801.09138*, 2018.

Victor H Peña, Tze Leung Lai, and Qi-Man Shao. *Self-normalized processes: Limit theory and statistical applications*. Springer Science & Business Media, 2008.

Ya’acov Ritov, Peter J Bickel, Anthony C Gamst, and Bastiaan Jan Korneel Kleijn. The bayesian analysis of complex, high-dimensional models: Can it be coda? *Statistical Science*, 29(4):619–639, 2014.

James Robins, Lingling Li, Eric Tchetgen, and Aad van der Vaart. Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and Statistics: Essays in Honor of David A. Freedman*, pages 335–421. Institute of Mathematical Statistics, 2008.

James Robins, Eric Tchetgen Tchetgen, Lingling Li, and Aad van der Vaart. Semiparametric minimax rates. *Electronic Journal of Statistics*, 3:1305, 2009.

James M Robins and Andrea Rotnitzky. Comments on “Inference for semiparametric models: some questions and an answer”. *Statistica Sinica*, pages 920–936, 2001.

James M Robins, Peng Zhang, Rajeev Ayyagari, Roger Logan, Eric Tchetgen Tchetgen, Lingling Li, Thomas Lumley, Aad van der Vaart, and HEI Health Review Committee. New statistical approaches to semiparametric regression with application to air pollution research. *Research report (Health Effects Institute)*, (175):3, 2013.

James M. Robins, Lingling Li, Eric Tchetgen Tchetgen, and Aad van der Vaart. Asymptotic normality of quadratic estimators. *Stochastic Processes and Their Applications*, 126(12):3733–3759, 2016.

James M Robins, Lingling Li, Rajarshi Mukherjee, Eric Tchetgen Tchetgen, and Aad van der Vaart. Minimax estimation of a functional on a structured high-dimensional model. *The Annals of Statistics*, 45(5):1951–1987, 2017.

Paul R Rosenbaum. Testing hypotheses in order. *Biometrika*, 95(1):248–252, 2008.

Andrea Rotnitzky, Ezequiel Smucler, and James M Robins. Characterization of parameters with a mixed bias property. Technical report, 2019a.

Andrea Rotnitzky, Ezequiel Smucler, and James M Robins. A unifying approach for doubly-robust ℓ1 regularized estimation of causal contrasts. Technical report, 2019b.

Herman Rubin and RA Vitale. Asymptotic distribution of symmetric statistics. *The Annals of Statistics*, pages 165–170, 1980.
Mark Rudelson. Random vectors in the isotropic position. *Journal of Functional Analysis*, 164(1):60–72, 1999.

Daniel O Scharfstein, Andrea Rotnitzky, and James M Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120, 1999a.

Daniel O Scharfstein, Andrea Rotnitzky, and James M Robins. Rejoinder. *Journal of the American Statistical Association*, 94(448):1135–1146, 1999b.

Anton Schick. On asymptotically efficient estimation in semiparametric models. *The Annals of Statistics*, 14(3):1139–1151, 1986.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *arXiv preprint arXiv:1708.06633*, 2017.

Vladimir Spokoiny and Céline Vial. Parameter tuning in pointwise adaptation using a propagation approach. *The Annals of Statistics*, 37(5B):2783–2807, 2009.

Charles J Stone. Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, pages 1348–1360, 1980.

Charles J Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, pages 1040–1053, 1982.

Taiji Suzuki. Adaptivity of deep reLU network for learning in besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=H1ebTsActm.

Alexandre B Tsybakov. *Introduction to nonparametric estimation*. Springer Science & Business Media, 2009.

Aad van der Vaart. Part III: Semiparametric statistics. *Lectures on Probability Theory and Statistics*, pages 331–457, 2002.

Aad W van der Vaart. *Asymptotic statistics*, volume 3. Cambridge University Press, 1998.

AW van der Vaart and Jon Wellner. *Weak Convergence and Empirical Processes: with Applications to Statistics*. Springer Science & Business Media, 1996.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Wenjing Zheng and Mark J van der Laan. Cross-validated targeted minimum-loss-based estimation. In *Targeted Learning*, pages 459–474. Springer, 2011.
Appendix A. Proof of Theorem 2.9

Proof. We denote the oracle test as $\hat{\chi}_k(Ω_k^{-1}; ζ_k, δ)$ and focus on the two-sided test only. We only need to compute the rejection probability:

$$\lim_{n \to \infty} P_\theta \left( \frac{|\hat{\Pi}_{22,k}(Ω_k^{-1})|}{se[\hat{\Pi}_{22,k}(Ω_k^{-1})]} - ζ_k \frac{se[\hat{\Pi}_{22,k}(Ω_k^{-1})]}{se[\hat{\psi}_1]} > δ \right)$$

$$= \lim_{n \to \infty} \left\{ P_\theta \left( \frac{\hat{\Pi}_{22,k}(Ω_k^{-1})}{se[\hat{\Pi}_{22,k}(Ω_k^{-1})]} > ζ_k + \delta \frac{se[\hat{\psi}_1]}{se[\hat{\Pi}_{22,k}(Ω_k^{-1})]} \right) + P_\theta \left( \frac{\hat{\Pi}_{22,k}(Ω_k^{-1})}{se[\hat{\Pi}_{22,k}(Ω_k^{-1})]} < -ζ_k - \delta \frac{se[\hat{\psi}_1]}{se[\hat{\Pi}_{22,k}(Ω_k^{-1})]} \right) \right\}$$

$$= \lim_{n \to \infty} P_\theta \left( \frac{\hat{\Pi}_{22,k}(Ω_k^{-1})}{se[\hat{\Pi}_{22,k}(Ω_k^{-1})]} - cBias_\theta(\hat{\psi}_1) > ζ_k - cBias_\theta(\hat{\psi}_1) + \delta \frac{se[\hat{\psi}_1]}{se[\hat{\Pi}_{22,k}(Ω_k^{-1})]} \right)$$

$$+ \lim_{n \to \infty} P_\theta \left( \frac{\hat{\Pi}_{22,k}(Ω_k^{-1})}{se[\hat{\Pi}_{22,k}(Ω_k^{-1})]} - cBias_\theta(\hat{\psi}_1) < -ζ_k - cBias_\theta(\hat{\psi}_1) - \delta \frac{se[\hat{\psi}_1]}{se[\hat{\Pi}_{22,k}(Ω_k^{-1})]} \right)$$

$$= \lim_{n \to \infty} P_\theta \left( \frac{\hat{\Pi}_{22,k}(Ω_k^{-1})}{se[\hat{\Pi}_{22,k}(Ω_k^{-1})]} (1 + o_P(1)) > ζ_k - (γ - δ) \frac{se[\hat{\psi}_1]}{se[\hat{\Pi}_{22,k}(Ω_k^{-1})]} (1 + o_P(1)) \right)$$

$$+ \lim_{n \to \infty} P_\theta \left( \frac{\hat{\Pi}_{22,k}(Ω_k^{-1})}{se[\hat{\Pi}_{22,k}(Ω_k^{-1})]} (1 + o_P(1)) < -ζ_k - (γ + δ) \frac{se[\hat{\psi}_1]}{se[\hat{\Pi}_{22,k}(Ω_k^{-1})]} (1 + o_P(1)) \right)$$

$$= 1 - \Phi \left( ζ_k - (γ - δ) \frac{se[\hat{\psi}_1]}{se[\hat{\Pi}_{22,k}(Ω_k^{-1})]} \right) + \Phi \left( -ζ_k - (γ + δ) \frac{se[\hat{\psi}_1]}{se[\hat{\Pi}_{22,k}(Ω_k^{-1})]} \right)$$

$$= 2 - \Phi \left( ζ_k - (γ - δ) \frac{se[\hat{\psi}_1]}{se[\hat{\Pi}_{22,k}(Ω_k^{-1})]} \right) - \Phi \left( ζ_k + (γ + δ) \frac{se[\hat{\psi}_1]}{se[\hat{\Pi}_{22,k}(Ω_k^{-1})]} \right).$$

\[ \square \]

Appendix B. A Possible Explanation of Stability of $\hat{\Pi}_{22,k}(Ω_k^{\text{est}})^{-1}$ and $\hat{\Pi}_{22,k}(Ω_k^{\text{quasi}})^{-1}$

In this section, as promised, we discuss why $\hat{\Pi}_{22,k}(Ω_k^{\text{est}})^{-1}$ and $\hat{\Pi}_{22,k}(Ω_k^{\text{quasi}})^{-1}$ are more stable than $\hat{\Pi}_{22,k}(Ω_k^{\text{true}})^{-1}$ in finite sample. We conjecture that such stability could be due to the “cancellation of eigenvalues” by self-normalization (Peña et al., 2008). To see this, we consider the matrix form of $\hat{\Pi}_{22,k}(Ω_k^{\text{est}})^{-1}$:

$$\hat{\Pi}_{22,k}(Ω_k^{\text{est}})^{-1} = \frac{1}{n - 1} \tilde{\varepsilon}_\theta^T \left( \mathbf{Z}_k^{\text{est}} (\mathbf{Z}_k^{\text{est}}^T \mathbf{Z}_k^{\text{est}})^{-1} \mathbf{Z}_k^{\text{est}}^T - \text{Diag}(\mathbf{Z}_k^{\text{est}} (\mathbf{Z}_k^{\text{est}}^T \mathbf{Z}_k^{\text{est}})^{-1} \mathbf{Z}_k^{\text{est}}^T) \right) \tilde{\varepsilon}_p$$

where $\tilde{\varepsilon}_\theta = (\tilde{\varepsilon}_{\theta,1}, \ldots, \tilde{\varepsilon}_{\theta,n})^T$, $\tilde{\varepsilon}_p = (\tilde{\varepsilon}_{p,1}, \ldots, \tilde{\varepsilon}_{p,n})^T$, and Diag (M) denotes the diagonal matrix with the diagonal elements of matrix M. Consider the singular value decomposition (SVD) of $\mathbf{Z}_k^{\text{est}} = \tilde{\mathbf{U}}_k \mathbf{D}_k \hat{\mathbf{V}}_k^T$ from the estimation sample, where $\mathbf{D}_k^2$ is the eigenvalues
of the sample covariance matrix estimator \( \hat{\Omega}_k^{\text{est}} \) up to constant. Then it is easy to see that

\[
\hat{F}_{22,k}(\hat{\Omega}_k^{\text{est}})^{-1}
\]

\[
= \frac{1}{n-1} \varepsilon_b^T \left\{ \begin{array}{c}
\hat{U}_k \hat{D}_k \hat{V}_k^T \left( \hat{V}_k \hat{D}_k^2 \hat{V}_k^T \right)^{-1} \hat{V}_k \hat{D}_k \hat{U}_k^T \\
- \text{Diag} \left( \hat{U}_k \hat{D}_k \hat{V}_k^T \left( \hat{V}_k \hat{D}_k^2 \hat{V}_k^T \right)^{-1} \hat{V}_k \hat{D}_k \hat{U}_k^T \right)
\end{array} \right\} \cdot \hat{\varepsilon}_p
\]

\[
= \frac{1}{n-1} \varepsilon_b^T \left\{ \hat{U}_k \hat{D}_k^2 \hat{U}_k^T - \text{Diag} \left( \hat{U}_k \hat{D}_k^2 \hat{D}_k \hat{U}_k^T \right) \right\} \cdot \hat{\varepsilon}_p
\]

by which we can explicitly see how the eigenvalues \( \hat{D}_k \) got cancelled from the second equality to the third equality.

Similarly, for \( \hat{F}_{22,k}(\hat{\Omega}_k^{\text{est}})^{-1} \), we have

\[
\hat{F}_{22,k}(\hat{\Omega}_k^{\text{est}})^{-1}
\]

\[
= \hat{F}_{22,k}(\hat{\Omega}_k^{\text{est}})^{-1} + \frac{1}{n-1} \varepsilon_b^T \left[ \begin{array}{c}
\text{Diag} \left( \hat{U}_k \cdot \hat{U}_k^T \right) \left\{ \hat{U}_k \cdot \hat{U}_k^T - \text{Diag} \left( \hat{U}_k \cdot \hat{U}_k^T \right) \right\}
\end{array} \right] \cdot \hat{\varepsilon}_p,
\]

again without involving the eigenvalues of \( \hat{\Omega}_k^{\text{est}} \). Moreover, with the SVD formulation, one can interpret \( \hat{F}_{22,k}(\hat{\Omega}_k^{\text{est}})^{-1} \) as follows: given the basis matrix \( \hat{Z}_k \), one first obtains its left singular vector \( \hat{U}_k \), then replaces \( \hat{Z}_k \) by \( \hat{U}_k \) and replaces \( \Omega_k \) by the identity matrix in \( \hat{F}_{22,k}(\Omega_k^{-1}) \) to get \( \hat{F}_{22,k}(\hat{\Omega}_k^{\text{est}})^{-1} \), and finally adds the correction terms to get \( \hat{F}_{22,k}(\hat{\Omega}_k^{\text{est}})^{-1} \).

Such “cancellation of eigenvalues” does not happen in \( \hat{F}_{22,k}(\hat{\Omega}_k^{\text{tr}})^{-1} \). Consider the SVD of \( \hat{Z}_k^{\text{tr}} = \hat{U}_k \hat{D}_k \hat{V}_k^T \) from the training sample. Similar to \( \hat{F}_{22,k}(\hat{\Omega}_k^{\text{est}})^{-1} \), the matrix form of \( \hat{F}_{22,k}(\hat{\Omega}_k^{\text{tr}})^{-1} \) is

\[
\hat{F}_{22,k}(\hat{\Omega}_k^{\text{tr}})^{-1}
\]

\[
= \frac{1}{n-1} \varepsilon_b^T \left\{ \begin{array}{c}
\hat{U}_k \hat{D}_k \hat{V}_k^T \cdot \left( \hat{V}_k \hat{D}_k^2 \hat{V}_k^T \right)^{-1} \cdot \hat{V}_k \hat{D}_k \hat{U}_k^T \\
- \text{Diag} \left( \hat{U}_k \hat{D}_k \hat{V}_k^T \cdot \left( \hat{V}_k \hat{D}_k^2 \hat{V}_k^T \right)^{-1} \cdot \hat{V}_k \hat{D}_k \hat{U}_k^T \right)
\end{array} \right\} \cdot \hat{\varepsilon}_p,
\]

in which case \( \hat{V}_k^T \hat{V}_k \neq \text{Id} \), where \( \text{Id} \) is the identity matrix and hence there is no cancellation in the eigenvalues as in \( \hat{F}_{22,k}(\hat{\Omega}_k^{\text{est}})^{-1} \) or \( \hat{F}_{22,k}(\hat{\Omega}_k^{\text{tr}})^{-1} \).

**Appendix C. Derivation of Eq. (3.1)**

\[
EB_{2,k}(\hat{\Omega}_k^{\text{est}})^{-1} \equiv \mathbb{E}_\theta \left[ \hat{F}_{22,k}(\hat{\Omega}_k^{\text{est}})^{-1} - \hat{F}_{22,k}(\Omega_k^{-1}) \right] \hat{\theta}
\]

\[
= \mathbb{E}_\theta \left[ \hat{\xi}_{b,1} z_k(X_1)^T \cdot \left\{ \frac{1}{n} \sum_{i=1}^n z_k(X_i) z_k(X_i)^T \right\}^{-1} - \Omega_k^{-1} \right] \cdot z_k(X_2) \hat{\xi}_{p,2} | \hat{\theta}
\]

\[
= \mathbb{E}_\theta \left[ \hat{\xi}_{b,1} z_k(X_1)^T \cdot \left( \hat{\Omega}_k^{\text{est},-1,-2} - \Omega_k^{-1} \right) \cdot z_k(X_2) \hat{\xi}_{p,2} | \hat{\theta} \right]
\]

\[
= \mathbb{E}_\theta \left[ \hat{\xi}_{b,1} z_k(X_1)^T \cdot \left( \hat{\Omega}_k^{\text{est},-1,-2} - \Omega_k^{-1} \right) \cdot z_k(X_2) \hat{\xi}_{p,2} | \hat{\theta} \right]
\]
where the second equality follows from the fact that
\[ \frac{1}{n} \sum_{i=1}^{n} \xi_k(X_i) \xi_k(X_i) \cdot \hat{\Omega}_{k,-1,-2}^{-1} \cdot \xi_k(X_2) \hat{\xi}_p \]
where the second equality follows from the definition
\[ \hat{\Omega}_{k,-i_1,-i_2} := \frac{1}{n} \sum_{i \in \mathbb{I} \neq i_1, i_2} \xi_k(X_i) \xi_k(X_i)^\top \]
for any \( 1 \leq i_1 \neq i_2 \leq n \), and the third equality is due to the exact expansion of the matrix inverse
\[ [\hat{\Omega}_{k,-1,-2}]^{-1} = \left( \hat{\Omega}_{k,-1,-2} + \frac{1}{n} \sum_{i=1,2} \xi_k(X_i) \xi_k(X_i)^\top \right)^{-1} \]
(C.1) := (I) + (II).

C.1. Upper bound on (I) of eq. (3.1).
\[
|\langle I \rangle | = | \mathbb{E}_\theta \left[ \hat{\xi}_b \xi_k(X)^\top \theta \right] \cdot \mathbb{E}_\theta \left[ \hat{\Omega}_{k,-1,-2}^{-1} \cdot \Omega_k \right] \cdot \mathbb{E}_\theta \left[ \xi_k(X) \hat{\xi}_p \theta \right] | \\
\leq | \mathbb{E}_\theta \left[ \hat{\xi}_b \xi_k(X)^\top \theta \right] \cdot \mathbb{E}_\theta \left[ \Omega_k^{-1} \left( \hat{\Omega}_{k,-1,-2} - \Omega_k \right) \Omega_k^{-1} \right] \cdot \mathbb{E}_\theta \left[ \xi_k(X) \hat{\xi}_p \theta \right] | \\
+ | \mathbb{E}_\theta \left[ \hat{\xi}_b \xi_k(X)^\top \theta \right] \cdot \mathbb{E}_\theta \left[ \left( \Omega_k^{-1} \left( \hat{\Omega}_{k,-1,-2} - \Omega_k \right) \Omega_k^{-1 -2} \right)^2 \right] \cdot \mathbb{E}_\theta \left[ \xi_k(X) \hat{\xi}_p \theta \right] | \\
= \frac{2}{n-2} | \mathbb{E}_\theta \left[ \hat{\xi}_b \xi_k(X)^\top \theta \right] \cdot \Omega_k^{-1} \cdot \mathbb{E}_\theta \left[ \xi_k(X) \hat{\xi}_p \theta \right] | \\
+ | \mathbb{E}_\theta \left[ \hat{\xi}_b \xi_k(X)^\top \theta \right] \cdot \mathbb{E}_\theta \left[ \left( \Omega_k^{-1} \left( \hat{\Omega}_{k,-1,-2} - \Omega_k \right) \Omega_k^{-1 -2} \right)^2 \right] \cdot \mathbb{E}_\theta \left[ \xi_k(X) \hat{\xi}_p \theta \right] | \\
\leq \left\| \mathbb{P} \left[ \hat{\xi}_b \xi_k(X)^\top \theta \right] \right\| \cdot \left\| \mathbb{P} \left[ \hat{\xi}_p \xi_k(X)^\top \theta \right] \right\| \cdot \left( \frac{1}{n-2} + \mathbb{E}_\theta \left[ \left\| \hat{\Omega}_{k,-1,-2} - \Omega_k \right\| \right] \right) \\
\leq \left\| \hat{\xi}_b \right\| \cdot \left\| \hat{\xi}_p \right\| \cdot \frac{2}{n} \cdot \mathbb{E}_\theta \left[ \left\| \hat{\Omega}_{k,-1,-2} - \Omega_k \right\| \right] \\
\leq \left\| \hat{\xi}_b \right\| \cdot \left\| \hat{\xi}_p \right\| \cdot \mathbb{E}_\theta \left[ \left\| \hat{\Omega}_{k,-1,-2} - \Omega_k \right\| \right],
\]
where the second line inequality follows from the following exact expansion of matrix inverse
\[ [\hat{\Omega}_{k,-1,-2}]^{-1} = \Omega_k^{-1} - \frac{1}{n} \left( \hat{\Omega}_{k,-1,-2} - \Omega_k \right) \Omega_k^{-1} + \left( \frac{1}{n} \left( \hat{\Omega}_{k,-1,-2} - \Omega_k \right) \Omega_k^{-1} \right)^2 [\hat{\Omega}_{k,-1,-2}]^{-1}
\]
and triangle inequality, the third line equality follows from the fact that \( \hat{\Omega}_{k,-1,-2} \) unbiasedly estimates \( \frac{1}{n} \Omega_k \), the fourth line inequality applies Cauchy-Schwarz inequality, the definition of operator norm, together with the identity
\[ \left\| \mathbb{E}_\theta \left[ \hat{\xi}_b \xi_k(X)^\top \theta \right] \right\| \Omega_k^{-1/2} = \mathbb{E}_\theta \left[ \left\| \mathbb{P} \left[ \hat{\xi}_b \xi_k(X)^\top \theta \right] \right\| \right],
\]
and the last line inequality follows from the contraction norm property of linear projections and the assumption that \( \Omega_k \) and \( \hat{\Omega}_{k,-1,-2} \) both have bounded eigenvalues.

Appendix D. Proof of Proposition 3.1

Proof. Recall eq. (3.1) in the main text, we have computed the order of the dominating terms (I), which are upper bounded by \( \left\| \mathbb{P} \left[ \hat{\xi}_b \xi_k(Z)^\top \theta \right] \right\| \cdot \left\| \mathbb{P} \left[ \hat{\xi}_p \xi_k(Z)^\top \theta \right] \right\| \cdot \frac{k_{\log(k)}}{n} \). We also corrected the contribution to the bias by (II) in eq. (3.1) by the extra de-bias term in \( \mathbb{P}^{\text{debiased}} \left( \hat{\Omega}_{k,-1,-2} \right) \) compared to \( \mathbb{P}^{\text{debiased}} \left( \Omega_k^{-1} \right) \). So the claimed bias bound in Appendix D has been proved.
To show the variance bound, we first show that \( \text{var}_\theta \left[ \widehat{\Omega}_{22,k}^{-1} \right] \) can also be bound by the claimed variance upper bound for \( \widehat{\Omega}_{22,k}^{-1} \). To show this, we use the following facts:

1. \( \widehat{\Omega}_{k}^{-1} \equiv \left( \widehat{\Omega}_{k}^{-1} \right)^{-1} - \frac{1}{n} \widehat{\Omega}_{k}^{-1} \overline{z}_k(X_2) \overline{z}_k(X_2)^\top \left( \widehat{\Omega}_{k}^{-1} \right)^{-1} \);
2. \( \sup_x |\overline{z}_k(x)^\top \Omega^{-1} \overline{z}_k(x)| \lesssim k \) for any \( \Omega^{-1} \) with bounded eigenvalues;
3. \( \overline{z}_b \) and \( \overline{z}_p \) are bounded.

(1) holds because \( \widehat{\Omega}_{k}^{-1} = \widehat{\Omega}_{k}^{-1} + \frac{1}{n} \overline{z}_k(X_2) \overline{z}_k(X_2)^\top \). (2) and (3) hold by the assumptions in Proposition 3.1.

Then following a standard change of measure argument as in the proof of Theorem 3 in Mukherjee et al. (2017):

\[
\text{var}_\theta \left[ \widehat{\Omega}_{22,k}^{-1} \right] \\
\leq \text{E}_\theta \left[ \left( \widehat{\Omega}_{22,k}^{-1} \right)^2 \right] \\
\leq \left| \frac{dP_\theta}{dP_\theta} \right|_\infty \text{E}_\theta \left[ \left( \widehat{\Omega}_{22,k}^{-1} \right)^2 \right] \\
\leq \frac{1}{n^2} \text{E}_\theta \left[ \left( \overline{z}_k(X_1)^\top \widehat{\Omega}_{k}^{-1} \overline{z}_k(X_2) \overline{z}_k(X_2)^\top \widehat{\Omega}_{k}^{-1} \overline{z}_k(X_1) \right)^2 \right] \\
= \frac{2 \| \overline{z}_b \|_2^2 \| \overline{z}_p \|_2^2}{n^2} \text{E}_\theta \left[ \left( \overline{z}_k(X_1)^\top \left( \widehat{\Omega}_{k}^{-1} \right)^{-1} \overline{z}_k(X_2) \overline{z}_k(X_2)^\top \left( \widehat{\Omega}_{k}^{-1} \right)^{-1} \overline{z}_k(X_1) \right)^2 \right] \\
\leq \frac{1}{n^2} \left\{ k + \frac{k^3}{n^2} \right\} \lesssim \frac{k}{n^2}.
\]

where the third line follows from \( \text{E}_\theta[Y - \hat{b}(X)|X] = \text{E}_\theta[A - \hat{p}(X)|X] = 0 \), the fourth line follows from fact (1) mentioned above, the fifth line follows from the trivial inequality \( (a + b)^2 \leq 2a^2 + 2b^2 \), and the sixth line follows from facts (2) and (3) mentioned above. For the variance bound of the de-bias term, we can simply follow the same arguments established below in Appendix G for the variance bound of the de-bias term in \( \widehat{\Omega}_{22,k}^{-1} \) and show that it is upper bounded by \( k/n^2 \).

\[
\text{APPENDIX E. PROOF OF PROPOSITION 3.5 AND LEMMA 3.8}
\]

In this section, we first sketch the proof of Proposition 3.5.

As discussed in Section 3.4, since \( \widehat{\Omega}_{22,k}(\hat{\Omega}^{-1}_k) \) is generally unknown, it needs to be replaced by an estimator such as \( \widehat{\Omega}_{22,k}(\hat{\Omega}_k^{-1}) \), \( \widehat{\Omega}_{22,k}(\hat{\Omega}^{-1}_k) \), \( \widehat{\Omega}_{debiased}^{\text{quasi}}(\hat{\Omega}^{-1}_k) \), \( \widehat{\Omega}_{22,k}^{\text{quasi}}(\hat{\Omega}^{-1}_k) \), or \( \widehat{\Omega}_{22,k}(\hat{\Omega}^{-1}_k) \).

For ease of presentation, in this section we use \( \widehat{\Omega}_{22,k}(\hat{\Omega}^{-1}_k) \) to denote an estimator of \( c\text{Bias}_{\theta,k}(\hat{\psi}_1) \) with \( \hat{\Omega}^{-1}_k \) replaced by some generic estimator. For example, If we use \( \widehat{\Omega}_{22,k}(\hat{\Omega}^{-1}_k) \), then \( \widehat{\Omega}_{22,k}(\hat{\Omega}^{-1}_k) = \widehat{\Omega}_{22,k}(\hat{\Omega}^{-1}_k) \). We consider the following standardized statistic in \( \hat{\chi}_k(\hat{\Omega}^{-1}_k; \zeta_k, \delta) \) when \( \hat{\Omega}^{-1}_k \) is replaced
by $\hat{\Omega}_k^{-1}$:

(E.1) \[
\frac{\text{se}[\hat{\psi}_1]}{\text{se}[\hat{\psi}_1]} \left( \frac{\text{IF}_{22,k}(\hat{\Omega}_k^{-1})}{\text{se}[\hat{\psi}_1]} - \delta \right) \]

\[
= \left( \frac{\text{IF}_{22,k}(\hat{\Omega}_k^{-1})}{\text{se}[\hat{\psi}_1]} - c\text{Bias}_{\theta,k}(\hat{\psi}_1) \right) + \frac{\text{se}[\hat{\psi}_1]}{\text{se}[\hat{\psi}_1]} \left( c\text{Bias}_{\theta,k}(\hat{\psi}_1) - \delta \right) \]

\[
= \left( \frac{\text{IF}_{22,k}(\hat{\Omega}_k^{-1})}{\text{se}[\hat{\psi}_1]} - \text{IF}_{\theta}(\hat{\Omega}_k^{-1}) \right) + \frac{\text{se}[\hat{\psi}_1]}{\text{se}[\hat{\psi}_1]} \left( \gamma + \text{IF}_{\theta}(\hat{\Omega}_k^{-1}) - \delta \right) \frac{\text{se}[\hat{\psi}_1]}{\text{se}[\hat{\psi}_1]} (1 + o_P(1)) \]

\[
:= (A + B) (1 + o_P(1)). \]

The effect of estimating $\Omega_k^{-1}$ on the asymptotic validity of the test $\hat{x}_k(\Theta_k^{-1}; \zeta_k, \delta)$ of $H_{0,k}(\delta)$ thus depends on the orders of terms $A$ and $B$, where $A$ is a zero mean and unit variance term, $B$ depends on the estimation bias due to estimating $\Omega_k^{-1}$ by $\hat{\Omega}_k^{-1}$, the fraction $\delta$ in the null hypothesis, and the truth $c\text{Bias}_{\theta,k}(\hat{\psi}_1)$. Thus the asymptotic validity of the test $\hat{x}_k(\hat{\Omega}_k^{-1}; \zeta_k, \delta)$ depends on whether the asymptotic normality of $A$ holds and whether the contribution of estimation bias term to $B$ is asymptotically negligible. If $\frac{\text{IF}_{\theta}(\hat{\Omega}_k^{-1})}{\text{se}[\hat{\psi}_1]} = o_P(\gamma \sqrt{\frac{k}{n}})$, then $B$ is not affected as $n \to \infty$. Then the exact asymptotic level and power follows once $A$ is asymptotic normal. Hence we proved Proposition 3.5. When above the local alternative to $H_{0,k}(\delta)$, as long as $A$ is $O_P(1)$, $B$ is of greater order than $A$ so the rejection probability still converges to 1 regardless of the normality of $A$.

In terms of Lemma 3.8, we first study $\text{IF}_{22,k}(\hat{\Omega}_k^{-1})$. For $\text{IF}_{22,k}(\hat{\Omega}_k^{-1})$, $A$ is asymptotic normally distributed conditioning on training sample for $\text{IF}_{22,k}(\hat{\Omega}_k^{-1})$ (see Remark 3.4). However, Mukherjee et al. (2017, Theorem 4) implies that, under Condition W, $\text{IF}_{22,k}(\hat{\Omega}_k^{-1}) = O_P(\sqrt{k \log(k)})$. Therefore $\frac{\text{IF}_{22,k}(\hat{\Omega}_k^{-1})}{\text{se}[\hat{\psi}_1]} = O_P(\gamma \sqrt{\frac{k \log(k)}{n}})$. If using $\text{IF}_{22,k}(\hat{\Omega}_k^{-1})$ instead, Mukherjee et al. (2017, Theorem 4) now implies that $\frac{\text{IF}_{22,k}(\hat{\Omega}_k^{-1})}{\text{se}[\hat{\psi}_1]} = O_P(\gamma \sqrt{\frac{k \log(k)}{n}})$. $\text{IF}_{22,k}(\hat{\Omega}_k^{-1})$ and $\text{IF}_{22,k}(\hat{\Omega}_k^{-1})$ have an improved order of estimation bias, which is the same as the order of the estimation bias of $\text{IF}_{22,k}(\hat{\Omega}_k^{-1}) + \text{IF}_{22,k}(\hat{\Omega}_k^{-1})$. So we reach the same conclusion as $\text{IF}_{22,k}(\hat{\Omega}_k^{-1}) + \text{IF}_{22,k}(\hat{\Omega}_k^{-1})$. Hence we proved Lemma 3.8.
APPENDIX F. Derivation of why $p \geq 2m + 1$ in Section 3.5

We consider when \( \left( \frac{n}{k} \right)^{m/2} \frac{EB_{p,k}(\hat{\Omega}_k^{\text{res}})^{-1}}{\text{se}_M(\hat{\psi}_{m,k}|\theta)} \) is $o_P(1)$ by the following calculation:

\[
\left( \frac{n}{k} \right)^{m/2} \frac{EB_{p,k}(\hat{\Omega}_k^{\text{res}})^{-1}}{\text{se}_M(\hat{\psi}_{m,k}|\theta)} = \left( \frac{n}{k} \right)^{m/2} \frac{\log(k)}{n} \frac{\|\Pi[\hat{\zeta}_k | Z_k]\|_{\theta}^2 \|\Pi[\hat{\zeta}_p | Z_k]\|_{\theta} (\frac{k\log(k)}{n})^{\frac{m-1}{2}}}{\text{se}_M(\hat{\psi}_{m,k}|\theta)} \\
= \left( \frac{k}{n} \right)^{p/2-m} \log(p) \frac{n^{m-p}}{n^{m-p}}
\]

where the third line is due to supposition that the upper bound in Mukherjee et al. (2017) on \( EB_{m,k}(\hat{\Omega}_k^{\text{res}})^{-1} \) is tight. Thus when $k = o(n)$, the RHS of eq. (F.1) is converging to zero if $p/2 - m > 0$ i.e. $p \geq 2m + 1$.

APPENDIX G. Proof of the variance bound for \( \hat{\Omega}_k^{\text{quasi}}(\hat{\Omega}_k^{\text{est}})^{-1} \)

Recall that the difference between \( \hat{\Omega}_k^{\text{quasi}}(\hat{\Omega}_k^{\text{est}})^{-1} \) and \( \hat{\Omega}_k^{\text{quasi}}(\hat{\Omega}_k^{\text{est}})^{-1} \) is the following extra term:

\[
\frac{1}{n^2(n-1)} \sum_{1 \leq i_1 \neq i_2 \leq n} \left\{ \hat{\zeta}_{b_{i_1}i_2} \hat{\zeta}_k(X_{i_1})^\top \left( \hat{\Omega}_k^{\text{est}} \right)^{-1} \left( \hat{\zeta}_k(X_{i_1}) \hat{\zeta}_k(X_{i_1})^\top + \hat{\zeta}_k(X_{i_2}) \hat{\zeta}_k(X_{i_2})^\top \right) \left( \hat{\Omega}_k^{\text{est}} \right)^{-1} \hat{\zeta}_k(X_{i_2}) \hat{\zeta}_{p_{i_1}i_2} \right\}.
\]

Thus for variance contribution for this extra term, by symmetry, we only need to consider the variance under the law $\hat{\theta}$ of the following term:

\[
\text{var}_{\hat{\theta}} \left[ \frac{1}{n^2(n-1)} \sum_{1 \leq i_1 \neq i_2 \leq n} \left\{ \hat{\zeta}_{b_{i_1}i_2} \hat{\zeta}_k(X_{i_1})^\top \left( \hat{\Omega}_k^{\text{est}} \right)^{-1} \left( \hat{\zeta}_k(X_{i_1}) \hat{\zeta}_k(X_{i_1})^\top + \hat{\zeta}_k(X_{i_2}) \hat{\zeta}_k(X_{i_2})^\top \right) \left( \hat{\Omega}_k^{\text{est}} \right)^{-1} \hat{\zeta}_k(X_{i_2}) \hat{\zeta}_{p_{i_1}i_2} \right\} \right]
\]

\[
= \frac{1}{n^3(n-1)} \text{E}_{\hat{\theta}} \left[ \hat{\zeta}_{b_{i_1}i_2}^2 \hat{\zeta}_k(X_{i_1})^\top \left( \hat{\Omega}_k^{\text{est}} \right)^{-1} \left( \hat{\zeta}_k(X_{i_1}) \hat{\zeta}_k(X_{i_1})^\top + \hat{\zeta}_k(X_{i_2}) \hat{\zeta}_k(X_{i_2})^\top \right) \left( \hat{\Omega}_k^{\text{est}} \right)^{-1} \hat{\zeta}_k(X_{i_2}) \hat{\zeta}_{p_{i_1}i_2} \right] \]

\[
\leq \frac{k^2}{n^3(n-1)} \left\{ k + \frac{k^3}{n^2} \right\} \leq \frac{k^3}{n^4}
\]

where the second inequality follows from the exact same calculation when we upper bound eq. (D.1). In summary,

\[
\begin{align*}
\text{var}_{\hat{\theta}} \left[ \hat{\Omega}_k^{\text{quasi}}(\hat{\Omega}_k^{\text{est}})^{-1} \right] \\
\lesssim \frac{k}{n^2} + \frac{k^3}{n^4} \lesssim \frac{k}{n^2}.
\end{align*}
\]
APPENDIX H. ON THE ESTIMATOR OF \( \text{var}_{\Theta} \left[ \hat{P}_{22,k} \right] \).

Up to now, we focused our discussion only on obtaining point estimates of \( c\text{Bias}_k(\theta; \hat{\theta}) \), but not on how to estimate the variance of \( \hat{P}_{22,k} \) in practice. In this section, to make our exposition simpler, we choose to focus on estimating \( \text{var}_{\Theta} \left[ \hat{P}_{22,k}(\Omega_k^{-1}) | \hat{\theta} \right] \). Under Condition W(3) \( \left\| \frac{dP_{\theta}}{dP_{\hat{\theta}}} \right\|_{\infty} < \infty \) (Robins et al., 2008, 2017; Mukherjee et al., 2017), from asymptotic consideration, the expectation of the following statistic conditioning on the training sample is of the same order as \( \text{var}_{\Theta} \left[ \hat{P}_{22,k}(\Omega_k^{-1}) | \hat{\theta} \right] \):

\[
\text{var}_{\Theta} \left[ \hat{P}_{22,k}(\Omega_k^{-1}) \right] := \frac{1}{n^2(n-1)^2} \sum_{1\leq i_1 \neq i_2 \leq n} \left\{ \hat{\varepsilon}_{b,i_1} \hat{\varepsilon}_{k}(X_{i_1})^\top \Omega_k^{-1} \hat{\varepsilon}_{k}(X_{i_2}) \hat{\varepsilon}_{p,i_2} \hat{\varepsilon}_{p,i_2} \hat{\varepsilon}_{k}(X_{i_2})^\top \Omega_k^{-1} \hat{\varepsilon}_{k}(X_{i_1}) \hat{\varepsilon}_{b,i_1} + \hat{\varepsilon}_{b,i_1} \hat{\varepsilon}_{k}(X_{i_1})^\top \Omega_k^{-1} \hat{\varepsilon}_{k}(X_{i_2}) \hat{\varepsilon}_{p,i_2} \hat{\varepsilon}_{p,i_2} \hat{\varepsilon}_{k}(X_{i_2})^\top \Omega_k^{-1} \hat{\varepsilon}_{k}(X_{i_1}) \hat{\varepsilon}_{b,i_1} \right\}.
\]

To see this, we can change the measure from \( P_{\theta} \) to \( P_{\hat{\theta}} \) under Condition W(2), the boundedness of \( \left\| \frac{dP_{\theta}}{dP_{\hat{\theta}}} \right\|_{\infty} \). Then under the probability measure \( P_{\hat{\theta}} \), \( \hat{P}_{22,k}(\Omega_k^{-1}) \) has mean zero. So we have

\[
\text{var}_{\Theta} \left[ \hat{P}_{22,k}(\Omega_k^{-1}) | \hat{\theta} \right] \equiv \mathbb{E}_{\Theta} \left[ \left( \hat{P}_{22,k}(\Omega_k^{-1}) \right)^2 | \hat{\theta} \right] = \frac{1}{n(n-1)} \mathbb{E}_{\Theta} \left\{ \hat{\varepsilon}_{b,i_1} \hat{\varepsilon}_{k}(X_{i_1})^\top \Omega_k^{-1} \hat{\varepsilon}_{k}(X_{i_2}) \hat{\varepsilon}_{p,i_2} \hat{\varepsilon}_{p,i_2} \hat{\varepsilon}_{k}(X_{i_2})^\top \Omega_k^{-1} \hat{\varepsilon}_{k}(X_{i_1}) \hat{\varepsilon}_{b,i_1} + \hat{\varepsilon}_{b,i_1} \hat{\varepsilon}_{k}(X_{i_1})^\top \Omega_k^{-1} \hat{\varepsilon}_{k}(X_{i_2}) \hat{\varepsilon}_{p,i_2} \hat{\varepsilon}_{p,i_2} \hat{\varepsilon}_{k}(X_{i_2})^\top \Omega_k^{-1} \hat{\varepsilon}_{k}(X_{i_1}) \hat{\varepsilon}_{b,i_1} \right\},
\]

where the last equality follows from the fact that the expected value of all the cross product terms zero under \( P_{\theta} \). Then one can see that the expectation of \( \text{var}_{\Theta} \left[ \hat{P}_{22,k}(\Omega_k^{-1}) \right] \) conditioning on the training sample under the probability measure \( P_{\theta} \) is exactly \( \text{var}_{\Theta} \left[ \hat{P}_{22,k}(\Omega_k^{-1}) | \Theta \right] \) and consequently the expectation of \( \text{var}_{\Theta} \left[ \hat{P}_{22,k}(\Omega_k^{-1}) \right] \) conditioning on the training sample is equal to \( \text{var}_{\Theta} \left[ \hat{P}_{22,k}(\Omega_k^{-1}) | \hat{\theta} \right] \) up to constant.

Unlike \( \text{var}_{\Theta} \left[ \hat{P}_{22,k}(\Omega_k^{-1}) | \Theta \right] \), however, \( \text{var}_{\Theta} \left[ \hat{P}_{22,k}(\Omega_k^{-1}) | \hat{\theta} \right] \) has other non-zero terms, the exact forms of which are given in eq. (H.3) in Appendix H.2. In fact, \( \text{var}_{\Theta} \left[ \hat{P}_{22,k}(\Omega_k^{-1}) \right] \) is only unbiased estimating the term (A) in eq. (H.3). As a result, \( \text{var}_{\Theta} \left[ \hat{P}_{22,k}(\Omega_k^{-1}) \right] \) is not necessarily a good finite-sample estimator of \( \text{var}_{\Theta} \left[ \hat{P}_{22,k}(\Omega_k^{-1}) | \hat{\theta} \right] \). In Table 5, the uncorrected variance estimators \( \text{var} \) for \( \hat{P}_{22,k}(\Omega_k^{-1}) \), \( \hat{P}_{22,k}(\Omega_k^{-1}) \), \( \hat{P}_{22,k}(\Omega_k^{-1}) \), \( \hat{P}_{22,k}(\Omega_k^{-1}) \) and \( \hat{P}_{22,k}(\Omega_k^{-1}) \) are generally underestimating the corresponding MC variance, even when their finite-sample performance is stable, i.e. not blowing up. For example, when \( k = 512 \), the MCav of \( \text{var} \left[ \hat{P}_{22,k}(\Omega_k^{-1}) \right] \) is only \( 2.46 \times 10^{-4} \), lower than the MC variance of \( \hat{P}_{22,k}(\Omega_k^{-1}) \) (4.67 \times 10^{-4}). Such underestimation of the variance can lead to under-coverage. To resolve this issue, we explore the possibility of making a finite-sample correction to \( \text{var} \) to obtain the estimated variance used in the simulation studies in this paper. Specifically, we define

\[
\text{var} \left[ \hat{P}_{22,k}(\Omega_k^{-1}) \right] := \text{var} \left[ \hat{P}_{22,k}(\Omega_k^{-1}) \right] + \frac{1}{n^2(n-1)^2} \sum_{i_1 \neq i_2 \neq i_3} \left\{ \hat{\varepsilon}_{b,i_1} \hat{\varepsilon}_{k}(X_{i_1})^\top \Omega_k^{-1} \hat{\varepsilon}_{k}(X_{i_2}) \hat{\varepsilon}_{p,i_2} \hat{\varepsilon}_{p,i_2} \hat{\varepsilon}_{k}(X_{i_2})^\top \Omega_k^{-1} \hat{\varepsilon}_{k}(X_{i_1}) \hat{\varepsilon}_{b,i_3} + \hat{\varepsilon}_{b,i_1} \hat{\varepsilon}_{k}(X_{i_1})^\top \Omega_k^{-1} \hat{\varepsilon}_{k}(X_{i_2}) \hat{\varepsilon}_{p,i_2} \hat{\varepsilon}_{p,i_2} \hat{\varepsilon}_{k}(X_{i_2})^\top \Omega_k^{-1} \hat{\varepsilon}_{k}(X_{i_1}) \hat{\varepsilon}_{b,i_3} + \hat{\varepsilon}_{b,i_1} \hat{\varepsilon}_{k}(X_{i_1})^\top \Omega_k^{-1} \hat{\varepsilon}_{k}(X_{i_2}) \hat{\varepsilon}_{p,i_3} \hat{\varepsilon}_{p,i_3} \hat{\varepsilon}_{k}(X_{i_3})^\top \Omega_k^{-1} \hat{\varepsilon}_{k}(X_{i_1}) \hat{\varepsilon}_{b,i_3} + \hat{\varepsilon}_{b,i_1} \hat{\varepsilon}_{k}(X_{i_1})^\top \Omega_k^{-1} \hat{\varepsilon}_{k}(X_{i_2}) \hat{\varepsilon}_{p,i_3} \hat{\varepsilon}_{p,i_3} \hat{\varepsilon}_{k}(X_{i_3})^\top \Omega_k^{-1} \hat{\varepsilon}_{k}(X_{i_1}) \hat{\varepsilon}_{b,i_3} + \hat{\varepsilon}_{b,i_1} \hat{\varepsilon}_{k}(X_{i_1})^\top \Omega_k^{-1} \hat{\varepsilon}_{k}(X_{i_2}) \hat{\varepsilon}_{p,i_3} \hat{\varepsilon}_{p,i_3} \hat{\varepsilon}_{k}(X_{i_3})^\top \Omega_k^{-1} \hat{\varepsilon}_{k}(X_{i_1}) \hat{\varepsilon}_{b,i_3} \right\},
\]

where the correction term eq. (H.2) is an unbiased estimator of the term (B) in eq. (H.3). As shown in Table 5, again when \( k = 512 \), the MCav of \( \text{var} \left[ \hat{P}_{22,k}(\Omega_k^{-1}) \right] \) is \( 5.27 \times 10^{-4} \), which is greater than the MC variance (4.67 \times 10^{-4}). Unlike \( \text{var} \left[ \hat{P}_{22,k}(\Omega_k^{-1}) \right] \), we observe empirically that \( \text{var} \left[ \hat{P}_{22,k}(\Omega_k^{-1}) \right] \) often
Table 5. Simulation result for $\mathbb{E}_\theta \left[(A - p(X))^2\right] \equiv 1$, with $\beta_{fx} = 0.4$

| $k$ | $\hat{\text{IF}}_{22,k}(\Omega_k^{-1})$ | $\hat{\text{IF}}_{22,k}(\Omega_{\text{est}}^{-1})$ | $\hat{\text{IF}}_{22,k}(\Omega_{\text{shrink}}^{-1})$ |
|-----|---------------------------------|---------------------------------|---------------------------------|
| 256 | 2.33 0.63 1.69                 | 2.41 0.78 1.54                 | 2.21 0.59 1.48                 |
| 512 | 4.67 2.46 5.27                 | 273.97 217.55 382.39          | 5.31 1.76 4.29                 |
| 1024| 8.59 5.00 13.65               | Blow up Blow up Blow up       | 7.80 2.04 9.96                 |
| 2048| 15.33 10.08 38.97             | Blow up Blow up Blow up       | 4.55 1.07 24.94                |

A comparison of the Monte Carlo variance and the MCav of the estimated variance $\hat{\text{IF}}_{22,k}(\Omega_k^{-1})$, $\hat{\text{IF}}_{22,k}(\Omega_{\text{est}}^{-1})$, and $\hat{\text{IF}}_{22,k}(\Omega_{\text{shrink}}^{-1})$ based on $\hat{\text{var}}$ or $\hat{\text{var}}$. All the numbers in the table should be multiplied by $10^{-4}$. For more details on the data generating mechanism, see Appendix K.1.

overestimates the MC variance of $\hat{\text{IF}}_{22,k}(\Omega_k^{-1})$. Therefore, we choose to use $\hat{\text{var}}\left[\hat{\text{IF}}_{22,k}(\Omega_k^{-1})\right]$ because it gives more conservative inference by slightly overestimating the variance.

In practice, we do not find it necessary to further correct the remaining term in $\hat{\text{var}}\left[\hat{\text{IF}}_{22,k}(\Omega_k^{-1})\right]$, which are term (C) and term (D) in eq. (H.3) as $\hat{\text{var}}\left[\hat{\text{IF}}_{22,k}(\Omega_k^{-1})\right]$ has already been overestimating the MC variance sometimes. Furthermore, for term (C), one can simplify it as follows:

$$(C) := \frac{(n-2)(n-3)}{n(n-1)} \mathbb{E}_\theta \left[\mathbb{E}_{b,1}z_k(X_1)^\top \Omega_k^{-1}z_k(X_2)\mathbb{E}_{p,2}X_3\mathbb{E}_{b,4}^\top \Omega_k^{-1}z_k(X_4)\mathbb{E}_{b,4}\right]$$

So the difference between (C) and (D) is only of order $1/n^2$.

**Remark H.1** (Variance estimator of $\hat{\text{IF}}_{22,k}(\Omega_{\text{est}}^{-1})$). For $\hat{\text{IF}}_{22,k}(\Omega_{\text{est}}^{-1})$ and $\hat{\text{IF}}_{22,k}(\Omega_{\text{shrink}}^{-1})$, since we estimate the covariance matrix estimator from the training sample, the above analysis on their variance and variance estimators immediately applies. However, because we estimate the covariance matrix from the estimation sample in $\hat{\text{IF}}_{22,k}(\Omega_{\text{est}}^{-1})$, we cannot directly argue that (C) = (D) asymptotically. In this paper, we use $\hat{\text{var}}[\hat{\text{IF}}_{22,k}(\Omega_{\text{est}}^{-1})]$ as the variance estimator of $\hat{\text{IF}}_{22,k}(\Omega_{\text{est}}^{-1})$, as we found in the third column of Table 5 that $\hat{\text{var}}[\hat{\text{IF}}_{22,k}(\Omega_{\text{est}}^{-1})]$ is either quite close to or overestimates the MC variance of $\hat{\text{IF}}_{22,k}(\Omega_{\text{est}}^{-1})$. Indeed, as shown in the third column of Table 5, when $k = 1024$, $\hat{\text{var}}[\hat{\text{IF}}_{22,k}(\Omega_{\text{est}}^{-1})] = 9.96 \times 10^{-4}$, slightly overestimates its MC variance $7.80 \times 10^{-4}$. Interestingly, when $k = 2048$, $\hat{\text{var}}[\hat{\text{IF}}_{22,k}(\Omega_{\text{est}}^{-1})] = 24.94 \times 10^{-4}$ is almost 6 times of its MC variance $4.55 \times 10^{-4}$, but we know from Table 3 that we should not use $\hat{\text{IF}}_{22,k}(\Omega_{\text{est}}^{-1})$ at $k = 2048$.

**H.1. Bootstrapping higher-order influence functions.** The variance estimator described above is not entirely satisfying as it often overestimates the actual variance of $\hat{\text{IF}}_{22,k}$ in finite sample. Another strategy of constructing estimator of the variance or standard deviation of $\hat{\text{IF}}_{22,k}$ is through bootstrap resampling strategies (Bickel and Freedman (1981)[Section 3]; Arcones and Giné (1992); Huskova and Janssen (1993)). We only consider $\hat{\text{IF}}_{22,k}(\Omega_k^{-1})$ because of its simple structure and known theoretical properties. We want to remark that as long as the inverse covariance matrix $\Omega^{-1}$ is estimated from a sample other than the estimation sample, similar result should hold for its variant such as $\hat{\text{IF}}_{22,k}(\Omega_{\text{H}}^{-1})$ as well. We consider the following simple nonparametric bootstrap resampling of $\hat{\text{IF}}_{22,k}(\Omega_k^{-1})$ with $B$ bootstrap samples:

- For $b = 1, \ldots, B$:
  - Draw independent $i_1^b, \ldots, i_n^b$ from $\{1, \ldots, n\}$ with replacement;
Table 6. Simulation result for $\mathbb{E}_{\theta} \left[ (A - p(X))^2 \right] \equiv 1$, with $\beta_{f_X} = 0.4$

| $k$  | $\text{MCvar} \left[ \widehat{\text{IF}}_{22,k}(\Omega_k^{-1}) \right]$ | $\text{MCavg} \left[ \text{var} \left[ \widehat{\text{IF}}_{22,k}(\Omega_k^{-1}) \right] \right]$ | $\text{MCavg} \left[ \text{var}^B \left[ \widehat{\text{IF}}_{22,k}(\Omega_k^{-1}) \right] \right]$ |
|------|-------------------------------------------------|-------------------------------------------------|-------------------------------------------------|
| 256  | 2.33                                           | 1.69                                            | 2.54                                            |
| 512  | 4.67                                           | 5.27                                            | 4.50                                            |
| 1024 | 8.59                                           | 13.65                                           | 8.66                                            |
| 2048 | 15.33                                          | 38.97                                           | 16.51                                           |

All the numbers in the table should be multiplied by $10^{-4}$. For more details on the data generating mechanism, see Appendix K.1.

- Compute $\widehat{\text{IF}}_{22,k}(\Omega_k^{-1})^b$ from the resampled dataset $\{O_i^j, j = 1, \ldots, n\}$;
- Then evaluate $\text{var}^B \left[ \widehat{\text{IF}}_{22,k}(\Omega_k^{-1}) \right] = \frac{1}{B-1} \sum_{b=1}^B \left( \widehat{\text{IF}}_{22,k}(\Omega_k^{-1})^b - \frac{1}{B} \sum_{b=1}^B \widehat{\text{IF}}_{22,k}(\Omega_k^{-1})^b \right)^2$.

Table 6 displays the finite-sample performance of the variance estimators $\text{var}^B \left[ \widehat{\text{IF}}_{22,k}(\Omega_k^{-1}) \right]$ of $\widehat{\text{IF}}_{22,k}(\Omega_k^{-1})$ by bootstrap resampling, which are closer to the MC variance than $\text{var} \left[ \widehat{\text{IF}}_{22,k}(\Omega_k^{-1}) \right]$ proposed in the previous section across different $k$’s.

H.2. Exact formula of var$_{\theta} \left[ \widehat{\text{IF}}_{22,k}(\Omega_k^{-1}) \right]$.

\[
\begin{align*}
\text{var}_{\theta} \left[ \widehat{\text{IF}}_{22,k}(\Omega_k^{-1}) \right] &= \text{E}_{\theta} \left[ \text{var} \left[ \widehat{\text{IF}}_{22,k}(\Omega_k^{-1}) \right] \right] \\
&= \text{E}_{\theta} \left[ \text{var} \left[ \widehat{\text{IF}}_{22,k}(\Omega_k^{-1}) \right] \right] \\
&= nP^{-2} \cdot nP \cdot \text{E}_{\theta} \left[ \left\{ \begin{array}{c} \hat{e}_{b,1} \hat{z}_k(X_1)^\top \Omega_k^{-1} \hat{e}_{b,2} \hat{z}_k(X_2) \hat{e}_{p,2} \hat{z}_k(X_2)^\top \Omega_k^{-1} \hat{e}_{b,3} \\
+ \hat{e}_{b,1} \hat{z}_k(X_1)^\top \Omega_k^{-1} \hat{e}_{b,2} \hat{z}_k(X_2) \hat{e}_{p,2} \hat{z}_k(X_2)^\top \Omega_k^{-1} \hat{e}_{b,3} \end{array} \right\} \hat{\theta} \right] \\
&+ nP^{-2} \cdot nP \cdot n^{-2} \cdot \text{E}_{\theta} \left[ \left\{ \begin{array}{c} \hat{e}_{b,1} \hat{z}_k(X_1)^\top \Omega_k^{-1} \hat{e}_{b,2} \hat{z}_k(X_2) \hat{e}_{p,2} \hat{z}_k(X_2)^\top \Omega_k^{-1} \hat{e}_{b,3} \\
+ \hat{e}_{b,1} \hat{z}_k(X_1)^\top \Omega_k^{-1} \hat{e}_{b,2} \hat{z}_k(X_2) \hat{e}_{p,2} \hat{z}_k(X_2)^\top \Omega_k^{-1} \hat{e}_{b,3} \end{array} \right\} \hat{\theta} \right] \\
&+ nP^{-2} \cdot nP \cdot n^{-2} \cdot \text{E}_{\theta} \left[ \left\{ \begin{array}{c} \hat{e}_{b,1} \hat{z}_k(X_1)^\top \Omega_k^{-1} \hat{e}_{b,2} \hat{z}_k(X_2) \hat{e}_{p,2} \hat{z}_k(X_2)^\top \Omega_k^{-1} \hat{e}_{b,3} \\
+ \hat{e}_{b,1} \hat{z}_k(X_1)^\top \Omega_k^{-1} \hat{e}_{b,2} \hat{z}_k(X_2) \hat{e}_{p,2} \hat{z}_k(X_2)^\top \Omega_k^{-1} \hat{e}_{b,3} \end{array} \right\} \hat{\theta} \right] \\
&- \left\{ \text{E}_{\theta} \left[ \hat{e}_{b,1} \hat{z}_k(X_1)^\top \Omega_k^{-1} \hat{e}_{b,2} \hat{z}_k(X_2) \hat{e}_{p,2} \hat{z}_k(X_2)^\top \Omega_k^{-1} \hat{e}_{b,3} \right] \right\}^2 \\
&= \frac{1}{n(n-1)} \text{E}_{\theta} \left[ \left\{ \begin{array}{c} \hat{e}_{b,1} \hat{z}_k(X_1)^\top \Omega_k^{-1} \hat{e}_{b,2} \hat{z}_k(X_2) \hat{e}_{p,2} \hat{z}_k(X_2)^\top \Omega_k^{-1} \hat{e}_{b,3} \\
+ \hat{e}_{b,1} \hat{z}_k(X_1)^\top \Omega_k^{-1} \hat{e}_{b,2} \hat{z}_k(X_2) \hat{e}_{p,2} \hat{z}_k(X_2)^\top \Omega_k^{-1} \hat{e}_{b,3} \\
+ \hat{e}_{b,1} \hat{z}_k(X_1)^\top \Omega_k^{-1} \hat{e}_{b,2} \hat{z}_k(X_2) \hat{e}_{p,2} \hat{z}_k(X_2)^\top \Omega_k^{-1} \hat{e}_{b,3} \end{array} \right\} \hat{\theta} \right] \\
&+ \frac{n-2}{n(n-1)} \text{E}_{\theta} \left[ \left\{ \begin{array}{c} \hat{e}_{b,1} \hat{z}_k(X_1)^\top \Omega_k^{-1} \hat{e}_{b,2} \hat{z}_k(X_2) \hat{e}_{p,2} \hat{z}_k(X_2)^\top \Omega_k^{-1} \hat{e}_{b,3} \\
+ \hat{e}_{b,1} \hat{z}_k(X_1)^\top \Omega_k^{-1} \hat{e}_{b,2} \hat{z}_k(X_2) \hat{e}_{p,2} \hat{z}_k(X_2)^\top \Omega_k^{-1} \hat{e}_{b,3} \\
+ \hat{e}_{b,1} \hat{z}_k(X_1)^\top \Omega_k^{-1} \hat{e}_{b,2} \hat{z}_k(X_2) \hat{e}_{p,2} \hat{z}_k(X_2)^\top \Omega_k^{-1} \hat{e}_{b,3} \end{array} \right\} \hat{\theta} \right] \\
&+ \frac{(n-2)(n-3)}{n(n-1)} \text{E}_{\theta} \left[ \left\{ \begin{array}{c} \hat{e}_{b,1} \hat{z}_k(X_1)^\top \Omega_k^{-1} \hat{e}_{b,2} \hat{z}_k(X_2) \hat{e}_{p,2} \hat{z}_k(X_2)^\top \Omega_k^{-1} \hat{e}_{b,3} \\
+ \hat{e}_{b,1} \hat{z}_k(X_1)^\top \Omega_k^{-1} \hat{e}_{b,2} \hat{z}_k(X_2) \hat{e}_{p,2} \hat{z}_k(X_2)^\top \Omega_k^{-1} \hat{e}_{b,3} \\
+ \hat{e}_{b,1} \hat{z}_k(X_1)^\top \Omega_k^{-1} \hat{e}_{b,2} \hat{z}_k(X_2) \hat{e}_{p,2} \hat{z}_k(X_2)^\top \Omega_k^{-1} \hat{e}_{b,3} \end{array} \right\} \hat{\theta} \right] \\
&= (A) + (B) + (C) - (D).
\end{align*}
\]

where $nP_r$ is the number of possible permutations of $r$ objects from a set of $n$ objects.

**Remark H.2.** Term (B) has some redundancy: the expectations of the first term and the fourth term are equal because of symmetry. This symmetry also applies to the estimator of term (B) (i.e. eq. (H.2)).
APPENDIX I. ON THE DATA-ADAPTIVE ESTIMATOR \( \hat{\theta}_{22,k}^{\text{adapt}} \) AND “OPTIMAL” \( k^* < n \)

We consider the expected conditional variance \( \psi(\theta) = \mathbb{E}_\theta [\text{var}_\theta (A | X)] \) as we know \( c_{\text{Bias}}(\hat{\psi}_1) \) increases with \( k \). Simulation results show that

1. the estimation bias of \( \hat{\theta}_{22,k}^{\text{quasi}} \left( \hat{\Omega}_k^{\text{est}} \right)^{-1} \) increases with \( k \), reflected by that the MCav of \( \hat{\theta}_{22,k}^{\text{quasi}} \left( \hat{\Omega}_k^{\text{est}} \right)^{-1} \) can decrease when \( k \) is near \( n \),

2. and \( \hat{\theta}_{22,k}^{\text{adapt}} \left( \hat{\Omega}_k^{\text{shrink}} \right)^{-1} \) blows up either when \( k \) is small compared to \( n \) or over all \( k \) when the density \( f_X \) of \( X \) is very rough.

Thus we design a data-adaptive algorithm to decide at each \( k \), whether \( \hat{\theta}_{22,k}^{\text{quasi}} \left( \hat{\Omega}_k^{\text{est}} \right)^{-1} \) can be used as \( \hat{\theta}_{22,k}^{\text{adapt}} \) and if not whether \( \hat{\theta}_{22,k}^{\text{adapt}} \left( \hat{\Omega}_k^{\text{shrink}} \right)^{-1} \) can be used. As a byproduct, the proposed data-adaptive algorithm also outputs a “optimal” \( k^* < n \) and the corresponding \( \hat{\theta}_{22,k^*}^{\text{adapt}} \), as the estimator of \( c_{\text{Bias}}(\hat{\psi}_1) \) that can best approximates \( c_{\text{Bias}}(\hat{\psi}_1) \) when \( \Omega^{-1} \) needs to be estimated. Ideally, a part of the estimation sample or a totally independent sample should be reserved for the implementation of the data-adaptive algorithm.

Since we do not have theoretical results on the estimation bias of \( \hat{\theta}_{22,k} \left( \hat{\Omega}_k^{\text{est}} \right)^{-1} \) and \( \hat{\theta}_{22,k} \left( \hat{\Omega}_k^{\text{shrink}} \right)^{-1} \), the algorithm is developed based on the empirical observations from the simulation studies. The statistical properties of this strategy are still under investigation.

In contrast to \( c_{\text{Bias}}(\hat{\psi}_1) \) monotonically increasing with \( k \), the MCav of \( \hat{\theta}_{22,k} \left( \hat{\Omega}_k^{\text{est}} \right)^{-1} \) starts to decrease when \( k \) is close to \( n \) as the estimation bias of \( \hat{\theta}_{22,k} \left( \hat{\Omega}_k^{\text{est}} \right)^{-1} \) increases with \( k \). Therefore, the first step of the data adaptive algorithm is to identify the point \( k^\text{quasi} \) at which \( \hat{\theta}_{22,k} \left( \hat{\Omega}_k^{\text{est}} \right)^{-1} \) stops increasing. For \( k \leq k^\text{quasi} \), we choose \( \hat{\theta}_{22,k}^{\text{adapt}} = \hat{\theta}_{22,k} \left( \hat{\Omega}_k^{\text{est}} \right)^{-1} \); for \( k > k^\text{quasi} \), we decide whether \( \hat{\theta}_{22,k} \left( \hat{\Omega}_k^{\text{shrink}} \right)^{-1} \) can be used as \( \hat{\theta}_{22,k}^{\text{adapt}} \) as when \( k \) is small relative to \( n \) or the density \( f_X \) is very rough, \( \hat{\theta}_{22,k} \left( \hat{\Omega}_k^{\text{shrink}} \right)^{-1} \) blows up.

We now describe the algorithm step by step. Suppose that we are given following ordered set \( \{k_1 < k_2 < \cdots < k_J\} \) of all candidate \( k \)'s. For \( j = 1, \ldots, J \), when \( j = j \), for some user-specified parameter \( c_j > 0 \):\(^{11}\)

- If \( \hat{\theta}_{22,k_{j+1}}^{\text{quasi}} \left( \hat{\Omega}_{k_{j+1}}^{\text{est}} \right)^{-1} < \hat{\theta}_{22,k_{j}}^{\text{quasi}} \left( \hat{\Omega}_{k_{j}}^{\text{est}} \right)^{-1} - \frac{c_j \text{var} \left( \hat{\theta}_{22,k_{j}}^{\text{quasi}} \left( \hat{\Omega}_{k_{j}}^{\text{est}} \right)^{-1} \right)}{1/2} \), the iteration terminates and outputs \( j^\text{quasi} = j \) (and \( k^\text{quasi} = k_j \)). For \( k \leq k^\text{quasi} \), the algorithm outputs \( \hat{\theta}_{22,k}^{\text{adapt}} = \hat{\theta}_{22,k}^{\text{quasi}} \left( \hat{\Omega}_k^{\text{est}} \right)^{-1} \).
- Otherwise, \( \hat{j} = j + 1 \) and repeat the above procedure.
  - If \( \hat{j} = J + 1 \), the entire data-adaptive algorithm terminates and outputs \( j^* = j^\text{quasi} = J \) (and \( k^* = k_J \)). For all \( k \leq K_J \), the algorithm outputs \( \hat{\theta}_{22,k}^{\text{adapt}} = \hat{\theta}_{22,k} \left( \hat{\Omega}_k^{\text{est}} \right)^{-1} \).
  - Finally, the algorithm also outputs \( \hat{\theta}_{22,k^*}^{\text{adapt}} = \hat{\theta}_{22,k_J}^{\text{adapt}} \).

Otherwise, we need to decide the lowest \( k \) such that \( \hat{\theta}_{22,k} \left( \hat{\Omega}_k^{\text{shrink}} \right)^{-1} \) can be used as \( \hat{\theta}_{22,k}^{\text{adapt}} \) for \( j > j^\text{quasi} \). For \( j^* = j^\text{quasi} + 1, j^\text{quasi} + 2, \ldots, J \), when \( \hat{j} = j^* \), for some user-specified parameter \( v_{j^*} > 0 \) (see Remark I.1):

* If \( \frac{\text{var} \left( \hat{\theta}_{22,k}^{\text{quasi}} \left( \hat{\Omega}_k^{\text{shrink}} \right)^{-1} \right)}{\text{var} \left( \hat{\theta}_{22,k} \left( \hat{\Omega}_k^{\text{quasi}} \left( \hat{\Omega}_k^{\text{shrink}} \right)^{-1} \right) \right)} \leq v_{j^*} \), the algorithm outputs \( j_{\text{shrink}} = \hat{j} \) (and \( k_{\text{shrink}} = \hat{k}_j \)). Then we need to decide the largest \( k \) such that \( \hat{\theta}_{22,k} \left( \hat{\Omega}_k^{\text{shrink}} \right)^{-1} \) can be used as \( \hat{\theta}_{22,k}^{\text{adapt}} \) for \( \hat{j} = j_{\text{shrink}} \).

\(^{11}\)Here one could choose \( c_j \) as 1 as a preliminary default setting.
· If \( \frac{\text{var}[\hat{\varphi}_{22,k}^\text{adapt}(\hat{\Omega}_k^{\text{est}})^{-1}]}{\text{var}[\varphi_{22,k}^\text{adapt}(\hat{\Omega}_k^{\text{est}})^{-1}]} > w_j \), the entire data-adaptive algorithm terminates and outputs \( j^* = j^\text{shrink} = j'' \) (and \( k^* \equiv k_j^\text{shrink} = k_j^\text{shrink} = k_j^\text{adapt} \)). For \( k_j^\text{shrink} \leq k \leq k_j^\text{shrink} \), the algorithm outputs \( \hat{\varphi}_{22,k}^\text{adapt} = \hat{\varphi}_{22,k}(\hat{\Omega}_k^{\text{est}})^{-1} \). Finally the algorithm outputs \( \hat{\varphi}_{22,k}^\text{adapt} = \hat{\varphi}_{22,k}^\text{adapt} \).

· Otherwise, \( j'' = j'' + 1 \) and repeat the above procedure. The entire algorithm terminates when \( j'' = J + 1 \).

* Otherwise, \( j' = j' + 1 \) and repeat the above procedure.

· If \( j' = J + 1 \), the entire data-adaptive algorithm terminates and outputs \( k^* = k_j^\text{quasi} \).

For any \( k > k_j^\text{quasi} \), the algorithm outputs \( \hat{\varphi}_{22,k} = \text{NA} \). Finally the algorithm also outputs \( \hat{\varphi}_{22,k}^\text{adapt} = \hat{\varphi}_{22,k}^\text{adapt} \).

In the end, the algorithm outputs \( k_j^\text{quasi} \), the largest \( k \) such that \( \hat{\varphi}_{22,k} = \hat{\varphi}_{22,k}(\hat{\Omega}_k^{\text{est}})^{-1} \), \( k_j^\text{shrink} \) and \( k_j^\text{adapt} \), the smallest and the largest \( k \) such that \( \hat{\varphi}_{22,k} = \hat{\varphi}_{22,k}(\hat{\Omega}_k^{\text{est}})^{-1} \), and \( k_j^\text{opt} \), the optimal \( k \) such that \( c\text{Bias}_{\theta,k}(\hat{\psi}_1) \) can be estimated by \( \hat{\varphi}_{22,k} = \hat{\varphi}_{22,k}(\hat{\Omega}_k^{-1}) \) when \( \hat{\Omega}_k^{-1} \) is estimated from data. In addition: for \( k \leq k_j^\text{quasi} \), the algorithm assigns \( \hat{\varphi}_{22,k} = \hat{\varphi}_{22,k}(\hat{\Omega}_k^{\text{est}})^{-1} \); for \( k_j^\text{shrink} \leq k \leq k_j^\text{quasi} \), the algorithm assigns \( \hat{\varphi}_{22,k} = \hat{\varphi}_{22,k}(\hat{\Omega}_k^{\text{est}})^{-1} \); for \( k_j^\text{quasi} < k < k_j^\text{shrink} \) or \( k > k_j^\text{shrink} \), the algorithm assigns \( \hat{\varphi}_{22,k} = \text{NA} \).

The algorithm also outputs \( \hat{\varphi}_{22,k}^\text{adapt} \) as the optimal data-adaptive estimator.

**Remark 1.** To decide \( k_j^\text{shrink} \) and \( k_j^\text{quasi} \), we need to specify the cutoff \( v_j > 0 \) and \( w_j > 0 \). Since the variance of \( \hat{\varphi}_{22,k}^{\text{adapt}}(\hat{\Omega}_k^{-1}) \) is of order \( k/n^2 \), one would expect the variance of \( \hat{\varphi}_{22,k}^{\text{adapt}}(\hat{\Omega}_k^{-1}) \) to grow linearly with \( k \). When choosing \( v_j \) to decide \( k_j^\text{shrink} \), we compare if \( \frac{\text{var}[\hat{\varphi}_{22,k_j}^{\text{adapt}}(\hat{\Omega}_k^{\text{est}})^{-1}]}{\text{var}[\varphi_{22,k_j}^{\text{adapt}}(\hat{\Omega}_k^{\text{est}})^{-1}]} \leq v_j \). Thus a reasonable choice is to set \( v_j \) proportional to \( C(k_j/k_j^\text{quasi}) \) for some \( C > 0 \). When choosing \( w_j \) to decide \( k_j^\text{quasi} \), we compare if \( \frac{\text{var}[\hat{\varphi}_{22,k_j+1}^{\text{adapt}}(\hat{\Omega}_k^{\text{est}})^{-1}]}{\text{var}[\varphi_{22,k_j}^{\text{adapt}}(\hat{\Omega}_k^{\text{est}})^{-1}]} \leq w_j \), again a reasonable choice is to choose \( w_j \) to be \( C(k_j+1/k_j) \) for some \( C > 0 \). In terms of the constant \( C > 0 \), as a heuristic, one can plot the ratios \( \frac{\text{var}[\hat{\varphi}_{22,k_j+1}^{\text{adapt}}(\hat{\Omega}_k^{\text{est}})^{-1}]}{\text{var}[\varphi_{22,k_j}^{\text{adapt}}(\hat{\Omega}_k^{\text{est}})^{-1}]} \cdot k_j \) for all \( k_j < k_j^\text{quasi} \) against \( k_j \) as information on reasonable range of the constant \( C \).

As for the expected conditional covariance \( \mathbb{E}_\theta[\text{cov}_\theta[A,Y|X]] \), however, \( c\text{Bias}_{\theta,k}(\hat{\psi}_1) \) is not guaranteed to increase with \( k \) even under Condition B. To circumvent such non-monotonicity, we first find \( k_j^\text{quasi;\theta} \) for the functional \( \mathbb{E}_\theta[\text{var}_\theta[Y|X]] \) and \( k_j^\text{quasi;p} \) for the functional \( \mathbb{E}_\theta[\text{var}_\theta[A|X]] \) respectively using the strategy described above for the expected conditional variance. Then we choose \( k_j^\text{quasi} = \min \{k_j^\text{quasi;\theta}, k_j^\text{quasi;p} \} \). For any \( k \leq k_j^\text{quasi} \), we choose \( \hat{\varphi}_{22,k} = \hat{\varphi}_{22,k}(\hat{\Omega}_k^{\text{est}})^{-1} \).

Then for \( k > k_j^\text{quasi} \), we use the same variance comparison strategy to determine \( k_j^\text{shrink} \) and \( k_j^\text{quasi} \). Eventually we use the same rule to determine \( k_j^\text{opt} \) and the “optimal” estimator \( \hat{\varphi}_{22,k}^\text{adapt} \).

**Appendix J. Generating functions from Hölder spaces in simulation studies**

A more detailed exposition can be found in Li et al. (2005). We first denote \( \omega(x) \) and \( \mu(x) \) as the Daubechies wavelet function with certain number of vanishing moments and its corresponding scaling function. We further denote \( \omega_{j,\ell}(x) := 2^{j/2} \omega(2^j x - \ell) \) and \( \mu_{j,\ell}(x) := 2^{j/2} \mu(2^j x - \ell) \) as the dilated (by \( 2^j \)) and translated (by \( \ell \)) of \( \omega(x) \) and \( \mu(x) \) respectively. By multi-resolution analysis (Mallat, 1999), \( \{\mu_{0,\ell}(x), \omega_{j,\ell}(x) : j \in \mathbb{Z}^+, \ell \in \mathbb{Z} \} \) is a complete orthonormal system in \( L_2 \) space. Thus for any function
$f \in L_2$, it permits a wavelet expansion of the following form

$$f(x) = \sum_{\ell \in \mathbb{Z}} \alpha_\ell \mu_{0,\ell}(x) + \sum_{j \in \mathbb{Z}^+} \sum_{\ell \in \mathbb{Z}} \gamma_{j,\ell} \omega_{j,\ell}(x)$$

where $\alpha_\ell, \gamma_{j,\ell} \in \mathbb{R}$ are the corresponding coefficients of the wavelet expansion.

It follows from Härdle et al. (1998, Theorem 9.6) that for any function $f(x)$ represented as wavelets expansions of the form in eq. (J.1): if $\mu(x)$ is $q$ times weakly differentiable and $\Delta^q \mu(x)$ is bounded, and for any desired smoothness index $\beta_f$, if $0 < \beta_f < q$ and there exists a constant $L > 0$ such that $\sup_\ell |\alpha_\ell| \leq L$ and $\sup_{j,\ell} 2^{j(\beta_f + 1/2)} |\gamma_{j,\ell}| \leq L$, then $f \in \text{Hölder}(\beta_f)$. Since we are interested in Hölder($\beta$) with $\beta < 1$, we can choose Daubechies wavelet function $\omega$ with more than one vanishing moments. In the simulations conducted in this paper, we fix Daubechies wavelet function $\omega$ with 3 vanishing moments (D6 mother/father wavelets) and generate the nuisance functions in the following manner:

- Define $J := \{0, 3, 6, 9, 10, 16\}$. For any desired smoothness index $0 < \beta_b < 1$ and $0 < \beta_p < 1$,

$$b(x) = \sum_{j \in J, \ell \in \mathbb{Z}} 2^{-j(\beta_b + 0.5)} \omega_{j,\ell}(x)$$

and

$$p(x) = \sum_{j \in J, \ell \in \mathbb{Z}} 2^{-j(\beta_p + 0.5)} \omega_{j,\ell}(x);$$

- For the marginal density function $f_X$ of the covariates $x$, given desired smoothness index $0 < \beta_{f_X} < 1$,

$$f_X(X) \propto 1 + \exp \left\{ \frac{1}{2} \sum_{j \in J, \ell \in \mathbb{Z}} 2^{-j(\beta_{f_X} + 0.5)} \omega_{j,\ell}(X) \right\}.$$ 

Then in simulation, $X$ will be sampled proportional to the value of $f_X(X)$.

### Appendix K. Simulation experiments

In this section, we first describe the details of our simulation studies, the results of which are reported in Table 1, Table 3, Table 4, Table 5, and Table 2. Afterwards, we will report two other sets of simulation studies to further support the finite sample performance of the estimators discussed in Section 3. Before describing the specific data generating mechanisms, we would like to mention several commonalities in these sets of simulation experiments:

- In the simulation studies, the nuisance functions (i.e. outcome regression $b$ and propensity score $p$) are generated to belong to Hölder classes with certain smoothness index $0 < \beta_b, \beta_p < 1$. In Appendix J, we discuss in detail how to generate such functions. In particular, we choose $\beta_b$ and $\beta_p$ such that $(\beta_b + \beta_p)/2 = 0.5$ for the average conditional covariance functional and $\beta_b = 0.25$ for the average conditional variance functional. These smoothness conditions have been shown to be at the boundary between $\sqrt{n}$-estimable and non-$\sqrt{n}$-estimable regimes in the minimax sense (Robins et al., 2009). We also restrict $X \in [0, 1]$.
- In most of the simulations, the nuisance functions are estimated by nonparametric kernel regression with cross validation. In Appendix K.3, we will examine whether estimating nuisance functions by convolutional neural networks still lead to similar qualitative conclusions to further support that our methodology is indeed agnostic to the use of any particular machine learning algorithm.
- For each $k$, the basis used to construct $\widehat{\mathcal{P}}_{22,k}$ is the D6 father wavelets at level $\log_2(k)$. The chosen basis functions satisfy Condition B. We plan to compare the performance of using different basis and even using data-driven algorithm to select basis in future work.
- In Appendix K.1 and Appendix K.2, we fix one training sample of size $n = 2500$ across all the simulations so all the simulation results in these two sections are conditional in nature. The estimation samples are different from simulation to simulation. In Appendix K.3, for each simulation, we have a total sample of size $N = 5000$ and divide the total sample into two equal-sized parts. Following the notations in Section 1 and Theorem 1.3 about cross-fit estimators, we first
fix one part as the training sample to estimate nuisance functions $b$ and $p$ and use the remaining estimation sample to estimate $\hat{\psi}_1, \widehat{IF}_{22,k}, \hat{\psi}_2, k$ and their estimated variances; then we obtain $\hat{\psi}_1, \widehat{IF}_{22,k}, \hat{\psi}_2, k$ and their estimated variances with the training and estimation samples reversed. To obtain the final point estimates and estimated variances, we simply take the average over the two separately constructed estimators:

\[
\begin{align*}
\widehat{\psi}_{\text{cross-fit},1} &= \frac{1}{2} \left( \hat{\psi}_1 + \overline{\psi}_1 \right), \\
\widehat{IF}_{\text{cross-fit},22,k} &= \frac{1}{2} \left( \widehat{IF}_{22,k} + \overline{IF}_{22,k} \right), \\
\widehat{\psi}_{\text{cross-fit},2,k} &= \frac{1}{2} \left( \hat{\psi}_2 + \overline{\psi}_2 \right), \\
\var\left( \widehat{\psi}_{\text{cross-fit},1} \right) &= \frac{1}{4} \left\{ \var\left( \hat{\psi}_1 \right) + \var\left( \overline{\psi}_1 \right) \right\}, \\
\var\left( \widehat{IF}_{\text{cross-fit},22,k} \right) &= \frac{1}{4} \left\{ \var\left( \widehat{IF}_{22,k} \right) + \var\left( \overline{IF}_{22,k} \right) \right\}, \\
\var\left( \widehat{\psi}_{\text{cross-fit},2,k} \right) &= \var\left( \widehat{\psi}_{\text{cross-fit},1} \right) + \var\left( \widehat{IF}_{\text{cross-fit},22,k} \right).
\end{align*}
\]

- To compute $\widehat{IF}_{22,k}(\Omega_k^{-1})$, we need to know the true $\Omega_k$ for each data generating mechanism. Since the analytical form of $\Omega_k$ is difficult to derive, we estimate $\Omega_k$ with sample covariance matrix from an independent sample of extremely large size of $8 \times 10^6$ (or X-semisupervised dataset) and use this estimator of $\Omega_k$ as the “oracle” $\Omega_k$.
- In Appendix K.1, we generated 1000 random datasets, whereas in the other simulation designs, we generate 200 random datasets in the simulation.

**Remark K.1** (On the methods of estimating nuisance functions). In most of the simulation studies, we use kernel nonparametric regression with bandwidth selected to minimize the cross-validated mean squared errors. We would like to emphasize that the only information that we need from nuisance function estimation is the residuals $\hat{\epsilon}_b$ and $\hat{\epsilon}_p$ for each subject. Therefore in order to construct $\widehat{IF}_{22,k}(\Omega_k^{-1})$ and its estimators, we are agnostic to any particular method of nuisance function estimation. Even if we use more complex methods such as deep convolutional neural networks (Farrell et al., 2018) and random forests (Wager and Athey, 2018), we can still construct $\widehat{IF}_{22,k}(\Omega_k^{-1})$ and its estimators from the data in the same manner as if the nuisance functions were estimated using kernel nonparametric regression. As a proof-of-concept, we also explored to use convolutional neural networks to estimate the nuisance functions $b$ and $p$ to study if second-order influence functions can still detect the bias in the DR-ML estimator in Appendix K.3.

**K.1. Simulation design for results from Table 1 to Table 2.** For the results in Table 1, Table 3, Table 4, Table 5, and Table 2, we choose the average conditional variance functional $\psi = E_0 \left[ (A - \mathbb{P}(X))^2 \right]$ as the target parameter of interest. As mentioned before, we generate $b(x) = \text{H"older}(b_0 = 0.251)$ where the covariates $X$ is one-dimensional drawn from a probability density function $f_X \in \text{H"older}(\beta_{fX} = 0.4)$, both following the recipe given in Appendix J. Then we generate $A \overset{iid}{\sim} N(p(X), 1)$ so $\psi = 1$.

**K.2. Non-smooth marginal density function $f_X(\cdot)$.** In this section, we compare the performance of the proposed estimators in Section 3 when the marginal density $f_X(\cdot)$ of the covariates $X$ belongs to Hölder class with smoothness index $\beta_{fX} = 0.125$ (Table 7 and Table 8) and $\beta_{fX} = 0.01$ (Table 9 and Table 10), whereas $\beta_b = \beta_p = 0.25$. We generate $A \overset{iid}{\sim} N(p(X), 1)$ so again $\psi = 1$. Similar to Table 1 and Table 3, we report the MCav of the point estimates and estimated standard errors of $\widehat{IF}_{22,k}(\Omega_k^{-1}), \widehat{IF}_{22,k}(\hat{\Omega}_k^{\text{emp}})^{-1}, \widehat{IF}_{22,k}(\hat{\Omega}_k^{\text{est}})^{-1}$, and $\widehat{IF}_{22,k}(\hat{\Omega}_k^{\text{shrink}})^{-1}$. The results can be found in Table 7 and Table 8 for $\beta_{fX} = 0.125$ and in Table 9 and Table 10 for $\beta_{fX} = 0.01$.

For $\beta_{fX} = 0.125$, the main message is similar to what we have discussed before. As in Table 7, $\widehat{IF}_{22,k}(\hat{\Omega}_k^{\text{shrink}})^{-1}$ starts to break down as early as $k = 512$, where its MCav of estimated standard error
Table 7. Simulation result for $E_{\theta}[(A - p(X))^2] \equiv 1$, with $\beta_{fx} = 0.125$, Bias($\psi_1$) $\approx 0.171$ (1)

| $k$ | $\hat{\Omega}_{22,k}^{-1}$ | $\hat{\Omega}_{tr,k}^{-1}$ | $\hat{\Omega}_{est,k}^{-1}$ | $\hat{\Omega}_{shrink,k}^{-1}$ |
|-----|---------------------------|--------------------------|--------------------------|--------------------------|
| 0   | (0)                       | (0)                      | (0)                      | (0)                      |
| 256 | 0.060 (0.015)             | 0.081 (0.020)            | 0.065 (0.014)            | Blow up (Blow up)        |
| 512 | 0.061 (0.021)             | 0.185 (0.130)            | 0.067 (0.020)            | Blow up (Blow up)        |
| 1024| 0.125 (0.036)             | Blow up (Blow up)        | 0.119 (0.032)            | 0.140 (0.036)            |
| 2048| 0.170 (0.062)             | Blow up (Blow up)        | 0.059 (0.050)            | 0.206 (0.062)            |

A comparison between $\hat{\Omega}_{22,k}$, $\hat{\Omega}_{tr,k}$, $\hat{\Omega}_{est,k}$ and $\hat{\Omega}_{shrink,k}$. The numbers in the parentheses are Monte Carlo average of estimated standard errors of the corresponding estimators. For more details on the data generating mechanism, see Appendix K.2.

Table 8. Simulation result for $E_{\theta}[(A - p(X))^2] \equiv 1$, with $\beta_{fx} = 0.125$, Bias($\psi_1$) $\approx 0.171$ (2)

| $k$ | $\hat{\Omega}_{22,k}^{-1}$ | $\hat{\psi}_{2,k}(\hat{\Omega}_{k}^{-1})$ | $\hat{\Omega}_{est,k}^{-1}$ | $\hat{\psi}_{2,k}(\hat{\Omega}_{k}^{-1})$ | $\hat{\Omega}_{shrink,k}^{-1}$ | $\hat{\psi}_{2,k}(\hat{\Omega}_{k}^{-1})$ |
|-----|---------------------------|------------------------------------------|--------------------------|------------------------------------------|--------------------------|------------------------------------------|
| 0   | (0)                       | (0)                                      | (0)                      | (0)                                      | (0)                      | (0)                                      |
| 256 | 0.060 (0.015)             | 0.040                                    | 0.111 (0.038)            | 0.065 (0.014)                           | 0.148                    | 0.106 (0.036)                            |
| 512 | 0.061 (0.021)             | 0.080                                    | 0.110 (0.041)            | 0.067 (0.020)                           | 0.250                    | 0.104 (0.039)                            |
| 1024| 0.125 (0.036)             | 0.820                                    | 0.046 (0.050)            | 0.119 (0.032)                           | 0.780                    | 0.051 (0.046)                            |
| 2048| 0.170 (0.062)             | 0.980                                    | 0.001 (0.071)            | 0.206 (0.062)                           | 0.960                    | -0.036 (0.071)                           |

We reported the MCav of point estimates and standard errors (first column in each panel) of $\hat{\Omega}_{22,k}$, together with the coverage probability of 90% confidence intervals (second column in each panel) of $\hat{\psi}_{2,k}$, the MCav of the bias and standard errors (third column in each panel) of $\hat{\psi}_{2,k}$ and $\hat{\psi}_{2,k}$. For more details, see Appendix K.2.

is 0.130, greater than that (0.021) of $\hat{\Omega}_{22,k}$ in magnitude. At this point, $\hat{\Omega}_{22,k}$ is the best estimator of $cBias_{\theta,k}(\hat{\psi}_1)$ and $\hat{\Omega}_{22,k}$ blows up when $k$ is not that large. At $k = 1024$, $\hat{\Omega}_{22,k}$ completely blows up. Interestingly, when the marginal density function $f_X$ is not very smooth, one can observe some finite-sample MC bias of $\hat{\Omega}_{22,k}$: the MCav of $\hat{\Omega}_{22,k}$ is 0.140 whereas the MCav of $\hat{\Omega}_{22,k}$ is 0.125 at $k = 1024$. Similarly, At $k = 2048$, the difference between the MCav of $\hat{\Omega}_{22,k}$ and $\hat{\Omega}_{22,k}$ is 0.206 - 0.170 = 0.036. The difference between $\hat{\Omega}_{22,k}$ and $\hat{\Omega}_{22,k}$ is unlikely attributed to the variability of the estimators, because the standard error of the MCav should be only about 0.036/$\sqrt{200}$ $\approx$ 0.0025 at $k = 1024$ and 0.062/$\sqrt{200}$ $\approx$ 0.00438 at $k = 2048$. In Table 10, we use the data-adaptive strategy in Appendix I to decide whether $\hat{\Omega}_{22,k}$ or $\hat{\Omega}_{22,k}$ should be reported at each $k$. Similarly, when $k < 2048$, $\hat{\Omega}_{22,k}$ is chosen as $\hat{\Omega}_{22,k}$ at the corresponding $k$ whereas when $k = 2048$, $\hat{\Omega}_{22,k}$ is chosen as $\hat{\Omega}_{22,k}$ because its estimated standard error does not exceed that of $\hat{\Omega}_{22,k}$. For $\beta_{fx} = 0.01$ (Table 9 and Table 10), there is a key difference: $\hat{\Omega}_{22,k}$ does not work at all over the whole range of $k$. This is an interesting phenomenon. As discussed in Mukherjee et al. (2017), one advantage of using sample covariance matrix estimator in higher-order influence function over using density estimation is its independence of the smoothness of the density of the covariates. However, from the numerical evidence in Table 9, this might not be true for some general covariance matrix estimators such as the nonlinear shrinkage estimator (Ledoit and Wolf, 2012). At $k = 2048$, $\hat{\Omega}_{22,k}$ breaks down when $\beta_{fx} = 0.01$. Following the data-adaptive strategy proposed in Appendix I, we will choose $k^* = 1024$ and use the corresponding $\hat{\Omega}_{22,k}$ as the best approximation of $cBias_{\theta}(\hat{\psi}_1)$.
Table 9. Simulation result for $\mathbb{E}_\theta \left[ (A - p(X))^2 \right] \equiv 1$, with $\beta_{fX} = 0.01$, $\text{Bias}(\hat{\psi}_1) \approx 0.197$ (1)

| $k$ | $\hat{\Pi}_{22,k}(\Omega_k^{-1})$ | $\hat{\Pi}_{22,k}(\Omega_k^{\mathrm{tr}})^{-1}$ | $\hat{\Pi}^\text{quasi}_{22,k}(\Omega_k^\text{est})^{-1}$ | $\hat{\Pi}_{22,k}(\Omega_k^\text{shrink})^{-1}$ |
|-----|----------------------------------|---------------------------------------------|---------------------------------------------|---------------------------------------------|
| 0   | 0 (0)                            | 0 (0)                                       | 0 (0)                                       | 0 (0)                                       |
| 256 | 0.083 (0.018)                    | 0.097 (0.025)                               | 0.079 (0.016)                               | Blow up (Blow up)                           |
| 512 | 0.083 (0.021)                    | 1.45 (3.58)                                 | 0.082 (0.020)                               | Blow up (Blow up)                           |
| 1024| 0.151 (0.038)                    | Blow up (Blow up)                           | 0.136 (0.033)                               | Blow up (Blow up)                           |
| 2048| 0.170 (0.067)                    | Blow up (Blow up)                           | 0.063 (0.051)                               | Blow up (Blow up)                           |

A comparison between $\hat{\Pi}_{22,k}(\Omega_k^{-1})$, $\hat{\Pi}_{22,k}(\Omega_k^{\mathrm{tr}})^{-1}$, $\hat{\Pi}^\text{quasi}_{22,k}(\Omega_k^{\text{est}})^{-1}$ and $\hat{\Pi}_{22,k}(\Omega_k^{\text{shrink}})^{-1}$. The numbers in the parentheses are Monte Carlo average of estimated standard errors of the corresponding estimators. For more details on the data generating mechanism, see Appendix K.2.

Table 10. Simulation result for $\mathbb{E}_\theta \left[ (A - p(X))^2 \right] \equiv 1$, with $\beta_{fX} = 0.01$, $\text{Bias}(\hat{\psi}_1) \approx 0.197$ (2)

| $k$ | $\hat{\Pi}_{22,k}(\Omega_k^{-1})$ | $\hat{\psi}_{2,k}(\Omega_k^{-1})$ | $\text{MC Coverage}$ (90% Wald CI) | $\text{Bias}(\hat{\psi}_{2,k}(\Omega_k^{-1}))$ | $\hat{\Pi}_{22,k}(\Omega_k^{-1})$ | $\hat{\psi}_{2,k}(\Omega_k^{-1})$ | $\text{MC Coverage}$ (90% Wald CI) | $\text{Bias}(\hat{\psi}_{2,k}(\Omega_k^{-1}))$ |
|-----|----------------------------------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|
| 0   | 0 (0)                            | 0 (0)                            | 0.197 (0.034)                    | 0.069 (0.034)                    |
| 256 | 0.083 (0.018)                    | 0.06                             | 0.115 (0.038)                    | 0.079 (0.016)                    | 0.06                             | 0.118 (0.037)                    |
| 512 | 0.083 (0.021)                    | 0.08                             | 0.114 (0.041)                    | 0.082 (0.020)                    | 0.06                             | 0.116 (0.039)                    |
| 1024| 0.151 (0.038)                    | 0.84                             | 0.047 (0.051)                    | 0.136 (0.033)                    | 0.69                             | 0.062 (0.047)                    |
| 2048| 0.170 (0.067)                    | 0.99                             | 0.005 (0.075)                    | 0.063 (0.051)                    | 0.14                             | 0.135 (0.061)                    |

We reported the MCav of point estimates and standard errors (first column in each panel) of $\hat{\Pi}_{22,k}(\Omega_k^{-1})$ and $\hat{\Pi}_{22,k}(\Omega_k^{-1})$, together with the coverage probability of 90% confidence intervals (second column in each panel) of $\hat{\psi}_{2,k}(\Omega_k^{-1})$ and $\hat{\psi}_{2,k}(\Omega_k^{-1})$, the MCav of the bias and standard errors (third column in each panel) of $\hat{\psi}_{2,k}(\Omega_k^{-1})$ and $\hat{\psi}_{2,k}(\Omega_k^{-1})$. For more details, see Appendix K.2.

K.3. Average conditional covariance functional. For the average conditional covariance functional, we still choose $\beta_b$ and $\beta_p$ such that $(\beta_b + \beta_p)/2 = 0.5$. In particular, we choose $\beta_b = 0.4$ and $\beta_p = 0.1$. $b$ and $p$ are again generated according to Appendix J. For simplicity, we choose uniformly distributed one-dimensional covariates $X$ in this section. We generate $Y \overset{iid}{\sim} N(b(X), 1)$ and $A \overset{iid}{\sim} N(p(X), 1)$ independently so $\psi \equiv 0$. For this simulation experiment, we estimated $b$ and $p$ in two different ways: (1) in the numerical results reported in Table 11 and Table 12, we estimated $b$ and $p$ by nonparametric kernel regression with cross validation as all the other simulations; (2) in the numerical results reported in Table 13 and Table 14, we estimated $b$ and $p$ by convolutional neural networks. The data- adaptive procedure in Appendix I is employed here to decide at each $k$, whether $\hat{\Pi}_{22,k}(\Omega_k^{\text{est}})^{-1}$ or $\hat{\Pi}_{22,k}(\Omega_k^{\text{shrink}})^{-1}$ should be used as the data-adaptive estimator reported in Table 12 and Table 14.

When estimating $b$ and $p$ by nonparametric kernel regression, the results can be found in Table 11 and Table 12. The basic message of this simulation experiment is very similar to Table 1, Table 3 and Table 4. Again, we found that $\hat{\Pi}_{22,k}(\Omega_k^{\text{shrink}})^{-1}$ is a very good estimator of $c\text{Bias}_b\theta_k(\hat{\psi}_1)$ when $k = 1024, 2048$ and this is likely because we generated $X$’s from uniform distribution which has a very smooth density function.

Remark K.2 (Architecture of convolutional neural networks). We estimated $b$ and $p$ by convolutional neural networks with 20 and 30 layers respectively ($\text{depth} = 20$ for $b$ and $30$ for $p$), each layer with 100 neurons (width = 100 for all layers). We choose deeper networks for $p$ ($\beta_p = 0.1$) because it is less smooth than $b$ ($\beta_b = 0.4$). The default rectified linear unit (ReLU) activation function was used in every intermediate layer and we set the learning rate parameter to be $2 \times 10^{-4}$. Recently, there are some very interesting theoretical analyses on the convergence rates of neural network estimators of functions in Hölder (Schmidt-Hieber, 2017) or Besov (Suzuki, 2019) type of function spaces. However, it is still quite difficult to choose the “right” architecture based on these theoretical results because the corresponding width and depth parameters are only optimal up to constants. In the simulation, we did not try to
Table 11. Simulation result for $\mathbb{E}_\theta [(Y - b(X))(A - p(X))] \equiv 0$, Bias($\psi_1) \approx 0.139 (1)$

| $k$ | $\widehat{F}_{22,k}(\Omega_k^{-1})$ | $\widehat{F}_{22,k}(\Omega_k^{-1})$ | $\widehat{F}_{22,k}(\Omega_k^{-1})$ | $\widehat{F}_{22,k}(\Omega_k^{-1})$ |
|-----|---------------------------------|---------------------------------|---------------------------------|---------------------------------|
| 0   | 0 (0)                           | 0 (0)                           | 0 (0)                           | 0 (0)                           |
| 256 | 0.033 (0.009)                   | 0.043 (0.012)                   | 0.041 (0.0085)                  | Blow up (Blow up)               |
| 512 | 0.039 (0.013)                   | 0.071 (0.061)                   | 0.043 (0.011)                   | 0.041 (0.014) (⋆)               |
| 1024| 0.094 (0.021)                   | Blow up (Blow up)               | 0.092 (0.017)                   | 0.095 (0.021)                   |
| 2048| 0.133 (0.033)                   | Blow up (Blow up)               | 0.135 (0.034)                   | 0.135 (0.034)                   |

A comparison between $\widehat{F}_{22,k}(\Omega_k^{-1})$, $\widehat{F}_{22,k}(\Omega_k^{-1})$, $\widehat{F}_{22,k}(\Omega_k^{-1})$, and $\widehat{F}_{22,k}(\Omega_k^{-1})$. The numbers in the parentheses are Monte Carlo average of estimated standard errors of the corresponding estimators. In (⋆), we reported MC median instead of MCav of the point estimates and their corresponding estimated standard error because of the existence of a few outliers. $b$ and $p$ are estimated by nonparametric kernel regression with cross validation. For more details on the data generating mechanism, see Appendix K.3.

Table 12. Simulation result for $\mathbb{E}_\theta [(Y - b(X))(A - p(X))] \equiv 0$, Bias($\psi_1) \approx 0.139 (2)$

| $k$ | $\widehat{F}_{22,k}(\Omega_k^{-1})$ | $\widehat{F}_{22,k}(\Omega_k^{-1})$ | $\widehat{F}_{22,k}(\Omega_k^{-1})$ | $\widehat{F}_{22,k}(\Omega_k^{-1})$ |
|-----|---------------------------------|---------------------------------|---------------------------------|---------------------------------|
| 0   | 0 (0)                           | 0 (0)                           | 0 (0)                           | 0 (0)                           |
| 256 | 0.033 (0.009)                   | 0.101 (0.023)                   | 0.041 (0.0085)                  | 0.098 (0.022)                   |
| 512 | 0.039 (0.013)                   | 0.100 (0.024)                   | 0.043 (0.011)                   | 0.096 (0.024)                   |
| 1024| 0.094 (0.021)                   | 0.045 (0.029)                   | 0.092 (0.017)                   | 0.047 (0.027)                   |
| 2048| 0.133 (0.034)                   | 0.0058 (0.039)                  | 0.135 (0.034)                   | 0.004 (0.040)                   |

We reported the MCav of point estimates and standard errors (first column in each panel) of $\widehat{F}_{22,k}(\Omega_k^{-1})$ and $\widehat{F}_{22,k}(\Omega_k^{-1})$, together with the coverage probability of 90% confidence intervals (second column in each panel) of $\widehat{\psi}_{2,k}(\Omega_k^{-1})$ and $\widehat{\psi}_{2,k}(\Omega_k^{-1})$, the MCav of the bias and standard errors (third column in each panel) of $\widehat{\psi}_{2,k}(\Omega_k^{-1})$ and $\widehat{\psi}_{2,k}(\Omega_k^{-1})$. $b$ and $p$ are estimated by nonparametric kernel regression with cross validation. For more details, see Appendix K.3.

To optimize the network architectures in order to obtain "optimal" prediction of $b$ and $p$. As expected, with the current architecture setup, the DR-ML estimator based on convolutional neural network nuisance estimators indeed has a slightly larger bias (MCav of bias = 0.243) than that based on nonparametric kernel regression nuisance estimators (MCav of bias = 0.139). All implementation was done using the R interface to Tensorflow (Abadi et al., 2016).

Even when $b$ and $p$ are estimated by convolutional neural networks, we find very similar qualitative results in Table 13 and Table 14. Without optimally tuning the network architectures, the DR-ML estimator $\hat{\psi}_1$ has higher bias than that based on nonparametric kernel regression with cross validation. As a consequence, the corresponding $\widehat{F}_{22,k}(\Omega_k^{-1})$ and $\widehat{F}_{22,k}(\Omega_k^{-1})$ (see Table 14) are also greater in magnitude than those based on nonparametric kernel regression in Table 12. This evidence further supports the applicability of the second-order influence function to test and correct bias in the DR-ML estimator.

Appendix L. Supplementary Figures

L.1. Histograms of the upper confidence bound. In Figure 3, we display the histograms of $UCB^{(1)}(\Omega_k^{-1}; \alpha = 0.10, \omega = 0.10)$ and $UCB^{(1)}(\Omega_k^{-1}; \alpha = 0.10, \omega = 0.10)$ at $k = 2048$ in simulation experiment described in Appendix K.1.

L.2. Qqplots of $\widehat{F}_{22,k}(\Omega_k^{-1})$, $\widehat{F}_{22,k}(\Omega_k^{-1})$, and $\widehat{F}_{22,k}(\Omega_k^{-1})$ in Appendix K.1. In Figure 4, we display the qqplots of $\widehat{F}_{22,k}(\Omega_k^{-1})$, $\widehat{F}_{22,k}(\Omega_k^{-1})$, and $\widehat{F}_{22,k}(\Omega_k^{-1})$ in simulation experiment described in Appendix K.1 over $k = 512, 1024, 2048$. 
Table 13. Simulation result for $\mathbb{E}_{\theta} [(Y - b(X))(A - p(X))] \equiv 0$, Bias(\hat{\psi}_1) \approx 0.243 (1)

| $k$  | \(\widehat{IF}_{22,k}(\Omega^{-1}_{k})\) | \(\widehat{IF}_{22,k}(\hat{\Omega}^{tr}_{k}^{-1})\) | \(\widehat{IF}_{22,k}(\hat{\Omega}^{est}_{k}^{-1})\) | \(\widehat{IF}_{22,k}(\hat{\Omega}^{shrink}_{k}^{-1})\) |
|------|---------------------------------|---------------------------------|---------------------------------|---------------------------------|
| 0    | 0 (0)                           | 0 (0)                           | 0 (0)                           | 0 (0)                           |
| 256  | 0.143 (0.015)                   | 0.159 (0.020) (*)               | 0.154 (0.014)                   | Blown up (Blown up)             |
| 512  | 0.143 (0.018)                   | 0.204 (0.073) (*)               | 0.158 (0.016)                   | 0.137 (0.019) (*)               |
| 1024 | 0.198 (0.027)                   | Blow up (Blow up)              | 0.194 (0.022)                   | 0.200 (0.027)                   |
| 2048 | 0.237 (0.042)                   | Blow up (Blow up)              | 0.084 (0.031)                   | 0.241 (0.042)                   |

A comparison between \(\widehat{IF}_{22,k}(\Omega^{-1}_{k})\), \(\widehat{IF}_{22,k}(\hat{\Omega}^{tr}_{k}^{-1})\), \(\widehat{IF}_{22,k}(\hat{\Omega}^{est}_{k}^{-1})\), and \(\widehat{IF}_{22,k}(\hat{\Omega}^{shrink}_{k}^{-1})\). The numbers in the parentheses are Monte Carlo average of estimated standard errors of the corresponding estimators. In (*), we reported MC median instead of MCav of the point estimates and their corresponding estimated standard error because of the existence of a few outliers. $b$ and $p$ are estimated by convolutional neural networks. For more details on the data generating mechanism, see Appendix K.3.

Table 14. Simulation result for $\mathbb{E}_{\theta} [(Y - b(X))(A - p(X))] \equiv 0$, Bias(\hat{\psi}_1) \approx 0.243 (2)

| $k$  | MC Coverage | Bias(\hat{\psi}_{2,k}(\Omega^{-1}_{k})) | MC Coverage | Bias(\hat{\psi}_{2,k}(\hat{\Omega}^{-1}_{k})) |
|------|-------------|---------------------------------|-------------|---------------------------------|
| 0    | 0 (0)       | 0.233 (0.024)                   | 0 (0)       | 0.243 (0.024)                   |
| 256  | 0.143 (0.015) | 0.099 (0.028)                   | 0.154 (0.014) | 0.015                           |
| 512  | 0.143 (0.018) | 0.099 (0.031)                   | 0.158 (0.016) | 0.030                           |
| 1024 | 0.198 (0.027) | 0.045 (0.036)                   | 0.194 (0.022) | 0.570                           |
| 2048 | 0.237 (0.042) | 0.0052 (0.048)                 | 0.241 (0.042) | 1.000                           |

We reported the MCav of point estimates and standard errors (first column in each panel) of \(\widehat{IF}_{22,k}(\Omega^{-1}_{k})\) and \(\widehat{IF}_{22,k}(\hat{\Omega}^{-1}_{k})\), together with the coverage probability of 90% confidence intervals (second column in each panel) of \(\hat{\psi}_{2,k}(\Omega^{-1}_{k})\) and \(\hat{\psi}_{2,k}(\hat{\Omega}^{-1}_{k})\), the MCav of the bias and standard errors (third column in each panel) of \(\hat{\psi}_{2,k}(\Omega^{-1}_{k})\) and \(\hat{\psi}_{2,k}(\hat{\Omega}^{-1}_{k})\). $b$ and $p$ are estimated by convolutional neural networks. For more details, see Appendix K.3.

Figure 3. Histograms of $UCB^{(1)}(\Omega^{-1}_{k}; \alpha = 0.10, \omega = 0.10)$ and $UCB^{(1)}(\hat{\Omega}^{-1}_{k}; \alpha = 0.10, \omega = 0.10)$ at $k = 2048$ in simulation experiment described in Appendix K.1.
Figure 4. Qqplots of $\hat{IF}_{22,k}(\Omega^{-1}_k)$, $\hat{IF}^{\text{quasi}}_{22,k}(\hat{\Omega}^{\text{est}}_k)^{-1}$, and $\hat{IF}^{\text{shrink}}_{22,k}(\hat{\Omega}^{\text{shrink}}_k)^{-1}$ in simulations in Appendix K.1.

Left panel: Qqplots for $\hat{IF}_{22,k}(\Omega^{-1}_k)$ with $k = 512$ (top), $k = 1024$ (middle), and $k = 2048$ (bottom); Middle panel: Qqnorm plots for $\hat{IF}^{\text{quasi}}_{22,k}(\hat{\Omega}^{\text{est}}_k)^{-1}$ with $k = 512$ (top), $k = 1024$ (middle), and $k = 2048$ (bottom); Right panel: Qqnorm plots for $\hat{IF}^{\text{shrink}}_{22,k}(\hat{\Omega}^{\text{shrink}}_k)^{-1}$ with $k = 1024$ (middle), and $k = 2048$ (bottom).