# ASSUMPTION-LEAN TESTS OF THE VALIDITY OF PUBLISHED DOUBLE-MACHINE-LEARNING WALD CONFIDENCE INTERVALS FOR DOUBLY-ROBUST FUNCTIONALS

LIN LIU[1], RAJARSHI MUKHERJEE[2], AND JAMES M. ROBINS[3]

In this article we study assumption-lean tests that can falsify an analyst's justification for the validity of a reported nominal $(1-\alpha)$ Wald confidence interval centered at a double machine learning (DML) estimator (Chernozhukov et al., 2018a; Farrell et al., 2020) for doubly robust (DR) functionals in the class studied by Rotnitzky et al. (2019). The class of such DR functionals is quite broad. It strictly includes both (i) the class of mean-square continuous functionals that can be written as an expectation of an affine functional of a conditional expectation studied by Chernozhukov et al. (2018c) and the class of functionals studied by Robins et al. (2008). The state-of-the-art estimators for DR functionals $\psi$ are DML estimators $\widehat{\psi}_{cf}$, which combine the benefits of the superior predictive performance of modern machine learning algorithms, the decreased bias of doubly robust estimation, and the analytical tractability and relaxed assumptions on the model complexity of sample splitting with cross fitting. The bias of DR estimators depends on the product of the rates at which two nuisance functions $b$ and $p$ are estimated. Here, for example, $b$ may be the regression of an outcome $Y$ on covariates $X$ and $p$ may be the propensity score (i.e. the regression of a binary treatment $A$ on $X$). However, the associated nominal $(1-\alpha)$ Wald confidence interval $\widehat{\psi}_{cf} \pm z_{\alpha/2}\widehat{\mathsf{s.e.}}(\widehat{\psi}_{cf})$ may still undercover even when the sample size is large, if the bias of the DML estimator is of the same or even greater order than its standard error of order $n^{-1/2}$. We show that it is possible to estimate a part of the bias of $\widehat{\psi}_{cf}$ (which we refer to as the estimable part of the bias) at faster than rate $n^{-1/2}$. The assumption-lean tests developed in this article are consistent tests of the null hypothesis that the estimable part of the bias is $o(n^{-1/2})$, without invoking any complexity-reducing (e.g. sparsity or smoothness) assumptions on $b$ and $p$. In contrast, no consistent estimator of the total bias exists without imposing complexity reducing assumptions that may be incorrect. Thus an analyst's justification for the validity of a their reported nominal $(1-\alpha)$ Wald interval therefore must rely on making complexity-reducing assumptions.

Most commonly analysts justify the validity of their Wald intervals by proving that, under their complexity-reducing assumptions, the Cauchy Schwarz (CS) upper bound $\mathsf{E}[(\widehat{b}(X)-b(X))^2]^{1/2}\mathsf{E}[(\widehat{p}(X)-p(X))^2]^{1/2}$ for the bias of $\widehat{\psi}_{cf}$ is $o(n^{-1/2})$. Thus if we can falsify the null hypothesis that the CS upper bound is $o(n^{-1/2})$, we also falsify the analysts complexity-reducing assumptions and, most importantly, the analyst's justification for the validity of her Wald intervals. We prove that falsification of the hypothesis that the estimable bias is $o(n^{-1/2})$ implies falsification of the hypothesis that the CS bound is $o(n^{-1/2})$. Hence, when our consistent test of the former hypothesis rejects then we can also reject the analyst's justification for the validity of their Wald interval. However, in the absence of complexity reducing assumptions, there exists no consistent test of the hypothesis

---

1. Institute of Natural Sciences, School of Mathematical Sciences, MOE-LSC, SJTU-Yale Joint Center for Biostatistics and Data Sciences, Shanghai Jiao Tong University, linliu@sjtu.edu.cn.

2. Department of Biostatistics, Harvard University, ram521@mail.harvard.edu.

3. Department of Epidemiology and Biostatistics, Harvard University, robins@hsph.harvard.edu.

This work is based on a talk given by the last author in the conference of "Whitney Newey's Contributions to Econometrics" held at MIT in 2019.

that the CS bound is $o(n^{-1/2})$. Hence, our assumption lean test failing to reject cannot be take as evidence for the hypothesis that the CS bound is $o(n^{-1/2})$. We conclude our paper with simulation experiments that demonstrate how the proposed assumption-lean test can be used in practice.

**Key words:** Causal Inference, Machine Learning, Assumption-lean Inference, Hypothesis Testing, Confidence Intervals, Doubly Robust Functionals, Higher-Order Influence Functions, U-Statistics

## 1. Introduction

In this paper we consider the problem of nearly assumption-free/assumption-lean test of the coverage of confidence intervals for a class of low dimensional functionals of high dimensional multivariate distributions that are of great interest in economics and statistics. Specifically, we consider the class of mixed bias functionals (MBFs) studied by Rotnitzky et al. (2019). This class strictly includes both (i) the class of mean-square continuous functionals that can be written as an expectation of an affine functional of a conditional expectation studied by Chernozhukov et al. (2018c) and the class of functionals studied by Robins et al. (2008). The contributions of the paper are three-fold. First we extend the results of Liu et al. (2020a) to the class of MBF. This requires we derive the form of higher order influence function (HOIF) (Robins et al., 2008; van der Vaart, 2014) estimators for functionals in the MBF class. Second we extend the results in Liu et al. (2020a) to the more realistic case where the distribution of the high dimensional covariates $X$ used in the definition of the MBF class is unknown. Liu et al. (2020a) restricted attention to the case where this distribution was known, except for an abbreviated discussion of the unknown case in an online supplement. Third, the results in this paper require much weaker assumptions on the basis functions on which the HOIFs depend than those in Liu et al. (2020a) allowing the proposed methodology to be used more widely in substantive studies.

The formal set up is as follows. We observe $N$ i.i.d. copies of the data vector $O = (W, X)$ drawn from some probability distribution $\mathsf{P}_\theta$ belonging to a model $\mathcal{M} = \{\mathsf{P}_\theta; \theta \in \Theta\}$ where $\Theta$ is a large, infinite dimensional parameter space and $X$ is a $d$-dimensional random vector with compact support. We assume $\mathcal{M}$ is locally nonparametric in the sense that the tangent space for the model at each $\theta \in \Theta$ is equal to $L_2(\mathsf{P}_\theta)$. To avoid technicalities that are orthogonal to the issues studied in this paper we assume that observed data $O$ is bounded with probability 1 (see Remark 3.3 for further discussion). We consider functionals (i.e. parameters) $\psi : \theta \mapsto \psi(\theta)$ that possess a (first order) influence function[1] (Bickel et al., 1998; van der Vaart, 2002, 1998) $\mathsf{IF}_{1,\psi}(\theta) = \mathsf{if}_{1,\psi}(O; \theta)$ (and thus a positive semiparametric variance bound (Newey, 1990)) and are contained in the mixed bias (MB) class of functionals of Rotnitzky et al. (2019) defined as follows.

[1]The term "influence function" when used without further qualification is to be understood to be the first order influence function.

**Definition 1.1** (Definition of a mixed bias (MB) functional (Definition 1 of Rotnitzky et al. (2019))). *$\psi(\theta)$ is an MB functional if, for each $\theta \in \Theta$ there exists $b : x \mapsto b(x) \in \mathcal{B}$ and $p : x \mapsto p(x) \in \mathcal{P}$ such that (i) $\theta = (b, p, \theta_{\setminus(b,p)})$ and $\Theta = \mathcal{B} \times \mathcal{P} \times \Theta_{\setminus(\mathcal{B},\mathcal{P})}$ and (ii) for any $\theta, \theta$‘*

$$\psi(\theta) - \psi(\theta') + \mathsf{E}_{\theta}\left[\mathsf{IF}_{1,\psi}(\theta')\right] = \mathsf{E}_{\theta}\left[S_{bp}(b(X) - b'(X))(p(X) - p'(X))\right] \tag{1.1}$$

*where $S_{bp} \equiv s_{bp}(O)$ with $o \mapsto s_{bp}(o)$ a known function that does not depend on $\theta$ or $\theta'$ satisfying either $\mathsf{P}_{\theta}(S_{bp} \geqslant 0) = 1$ or $\mathsf{P}_{\theta}(S_{bp} \leqslant 0) = 1$.*

To understand the importance of the MB property, let $\mathsf{P}_N$ be the empirical expectation operator and suppose $\theta'$ were an estimate of $\theta$ from a separate sample (which we regard as fixed, i.e. non-random). It then follows that the one step estimator $\psi(\theta') + \mathsf{P}_N\left[\mathsf{IF}_{1,\psi}(\theta')\right]$ is doubly robust (Bang and Robins, 2005; Robins and Rotnitzky, 2001; Scharfstein et al., 1999a,b). That is, by (1.1), it is unbiased for $\psi(\theta)$ under $\mathsf{P}_{\theta}$ if either $b = b'$ or $p = p'$. Because of this fact we will use the term MB functional and doubly robust (DR) functional interchangeably, as is done in much of the current literature. To ease notation, we will restrict consideration to DR functionals for which $\mathsf{P}_{\theta}(S_{bp} \geqslant 0) = 1$. For a DR functional $\psi^{\dagger}(\theta)$ of substantive interest for which $\mathsf{P}_{\theta}(S_{bp} \leqslant 0) = 1$, we will instead analyze the DR functional $\psi(\theta) = -\psi^{\dagger}(\theta)$.

Let $W = (Y, A)$. Below are some examples of DR functionals.

(1) The counterfactual mean $\mathsf{E}_{\theta}\left[Y(a = 1)\right]$ of $Y$ when $\{0, 1\}$-valued $A$ is set to $1$ (under strong ignorability) is identified from the distribution of $W = (A, AY, X)$ by the DR functional $\psi^{\dagger}(\theta) = \mathsf{E}_{\theta}[b(X)]$ with $b(x) = \mathsf{E}_{\theta}\left[Y|X = x, A = 1\right]$, $p(x) = 1/\mathsf{E}_{\theta}[A|X = x]$, and $S_{bp} = -A$. Here $p(x)$ is the inverse of the propensity score $\pi(x) = \mathsf{E}_{\theta}[A|X = x]$. Because $S_{bp} = -A$, we instead analyze the DR functional
$$\psi(\theta) = -\psi^{\dagger}(\theta) = -\mathsf{E}_{\theta}[b(X)] = -\mathsf{E}_{\theta}[Y(a = 1)]$$
for which $S_{bp} = A$. We will use $\psi(\theta) = -\mathsf{E}_{\theta}[b(X)] = -\mathsf{E}_{\theta}[Y(a = 1)]$ as a running example below[2].

(2) The expected conditional covariance
$$\psi(\theta) = \mathsf{E}_{\theta}\left[(Y - b(X))(A - p(X))\right]$$
with $b(x) = \mathsf{E}_{\theta}\left[Y|X = x\right], p(x) = \mathsf{E}_{\theta}\left[A|X = x\right]$ is a DR functional (equivalently, a MB functional) with $S_{bp} = 1$.

The state-of-the-art estimators for DR functionals such as $\psi(\theta) = -\mathsf{E}_{\theta}[b(X)] = -\mathsf{E}_{\theta}[Y(a = 1)]$ are double machine learning (DML) estimators or, equivalently, doubly robust machine learning (DRML) estimators $\widehat{\psi}_{\mathsf{cf},1}$ (Ayyagari, 2010; Chernozhukov et al., 2018a; Zheng and van der Laan, 2011), which combine the benefits of the superior predictive performance of modern machine learning algorithms,

[2] Chernozhukov et al. (2018c) used an alternative decomposition under which they included our $A, X$ in their $X$. See Example 1 of Rotnitzky et al. (2019).

the decreased bias of doubly robust estimation (Bang and Robins, 2005; Robins and Rotnitzky, 2001; Scharfstein et al., 1999a,b), and the analytical tractability and the decreased bias of sample splitting with cross fitting. More specifically, the data is randomly divided into two (or more) samples - the estimation sample of size $n$ and the training or nuisance sample of size $n_{\mathsf{tr}} = N - n$ with $1 - c > n/N > c$ for some $c \in (0, 1)$. To simplify the exposition we will take $c$ to be $1/2$. Machine learning estimators $\widehat{b}(x)$ of the outcome regression $b(x)$ and $\widehat{p}(x)$ of the inverse propensity score $p(x)$ are fit using the training sample data. The one step estimator of $\psi(\theta)$ is $\widehat{\psi}_1 = \psi(\widehat{\theta}) + \mathsf{P}_n \left[\mathsf{IF}_{1,\psi}(\widehat{\theta})\right] = \mathsf{P}_n \left[-\widehat{b}(X) - A\widehat{p}(X)(Y - \widehat{b}(X))\right]$ of $\psi(\theta)$ is computed from the estimation sample treating. Here $\psi(\widehat{\theta})$ is a plug-in estimator of $\psi(\theta)$. 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 Donsker class – a condition often needed for valid inference when sample splitting is not employed. Under conditions given in Theorem 2.3 below, the efficiency lost due to sample splitting can be recovered by cross fitting. The cross-fit estimator $\widehat{\psi}_{\mathsf{cf},1}$ averages $\widehat{\psi}_1$ with its 'twin' obtained by exchanging the roles of the estimation and training sample. However, even using DML estimators, the associated nominal $(1 - \alpha)$ Wald confidence interval $\widehat{\psi}_{\mathsf{cf},1} \pm z_{\alpha/2}\widehat{\mathsf{s.e.}}(\widehat{\psi}_{\mathsf{cf},1})$ may still undercover $\psi(\theta)$ even when the sample size is large, if the bias of the DML estimator $\widehat{\psi}_{\mathsf{cf},1}$ is of the same or even greater order than its standard error of order $n^{-1/2}$. In the following we will say that a nominal $(1 - \alpha)$ confidence interval is valid if its coverage probability under repeated sampling is greater than or equal to $(1 - \alpha)$.

In general a data analyst who reports the Wald confidence interval $\widehat{\psi}_{\mathsf{cf},1} \pm z_{\alpha/2}\widehat{\mathsf{s.e.}}(\widehat{\psi}_{\mathsf{cf},1})$ for a DR functional justifies its validity by (i) imposing complexity-reducing assumptions on the nuisance functions $b$ and $p$ (e.g. in terms of smoothness (Farrell et al., 2020; Robins et al., 2009b) or sparsity (Bradic et al., 2019; Smucler et al., 2019)) and then (ii) appealing to theorems wherein the asymptotic validity of the Wald confidence interval has been proved under these assumptions. However, these assumptions may be incorrect about the true but unknown functions $b$ and $p$ generating the data. In particular, Robins and Ritov (1997), Robins et al. (2009b) and Ritov et al. (2014) proved that, if the parameter space $\Theta$ does not entail any restriction on the complexity of $b$ and $p$ [beyond the assumption that the propensity score $1/p(X)$ is bounded away from zero, which is needed to insure a positive semiparametric variance bound], then (i) the bias of any estimator of $\psi(\theta) = -\mathsf{E}_\theta[b(X)] = -\mathsf{E}_\theta[Y(a = 1)]$ may be order 1 [i.e., there does not exist an estimator of $\psi(\theta)$ such that the bias of estimator converges to zero as $N \to \infty$ for all $\theta \in \Theta$]; and (ii) there does not exist a uniform (in $\theta \in \Theta$) consistent test of the composite null hypothesis that $b$ and $p$ satisfy given smoothness or sparsity assumptions. From (ii) it follows that no matter the sample size $N$, for any $\alpha^\dagger$-level test of this composite null, there exists $\theta = (b, p, \theta_{\setminus(b,p)})$ in the complement to the null such that the power under $\mathsf{P}_\theta$ is less than or equal to $\alpha^\dagger$. From (i) it follows that there exists no valid nominal $(1-\alpha)$ confidence interval whose median length converges to $0$ for all $\theta \in \Theta$ as the sample size $N \to \infty$. In this sense, without imposing possibly incorrect restrictive assumptions on

$b$ and $p$, no meaningful inference concerning $\psi(\theta)$ is possible. Nonetheless, in the sequel, unless stated otherwise we will not impose any complexity reducing assumptions on $b$ or $p$.

In this paper we adopt a *Skeptic's Approach to Statistics* whose goal is to empirically demonstrate, when possible, that the bias of the analyst's estimator $\widehat{\psi}_{\mathsf{cf},1}$ (under the unknown true law $\mathsf{P}_\theta$) is of the same or greater order than its standard error. This would prove to her not only that her assumptions must be incorrect but, more importantly, that her Wald interval $\widehat{\psi}_{\mathsf{cf},1} \pm z_{\alpha/2}\widehat{\mathsf{s.e.}}(\widehat{\psi}_{\mathsf{cf},1})$ cannot be valid. However, for parameters in the DR functional (equivalently MB functional) class this is only possible under a the so-called faithfulness assumption given in Section 3.2.1. Heuristically, faithfulness is the assumption that near perfect cancelling of the bias of two separate components of the total bias of $\widehat{\psi}_{\mathsf{cf},1}$ will essentially never occur.

If we do not assume faithfulness, the Skeptic's approach has the less ambitious goal of empirically demonstrating to the analyst, when possible, that her complexity reducing assumptions on $b$ and $p$ are incorrect [without being able to ever empirically prove the bias of her estimator is of the order of its standard error or greater]. We shall see, this goal is often achievable for parameters in the DR functional class. If successfully achieved, the analyst would then have to admit that she can no longer justify her earlier claim of validity for her state-of-the-art confidence interval. The approach described here is one of being in dialogue with current practices and practitioners. This is not surprising, as it is the statements of the practitioners that the skeptic is critiquing.

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

$$\mathsf{CSBias}_\theta(\widehat{\psi}_1) = \mathsf{E}_\theta\left[A\left(\widehat{b}(X) - b(X)\right)^2\right]^{1/2} \mathsf{E}_\theta\left[A\left(\widehat{p}(X) - p(X)\right)^2\right]^{1/2}$$

is $o(n^{-1/2})$ conditional on the training sample data $\mathbf{O}_{\mathsf{tr}}$[3] (and thus also on the functions $\widehat{b}, \widehat{p}$ computed from the training sample) and then noting the CS bias upper bounds the absolute conditional bias

$$\left|\mathsf{E}_\theta\left[A\left(\widehat{b}(X) - b(X)\right)\left(\widehat{p}(X) - p(X)\right)\right]\right|$$

of $\widehat{\psi}_1$. It then follows if we empirically show that Cauchy Schwarz bias $\mathsf{CSBias}_\theta(\widehat{\psi}_1)$ exceeds some given multiple $\delta > 0$, e.g. $\delta = 3/4$, times $\widehat{\psi}_1$'s conditional standard error of order $n^{-1/2}$, then we have falsified the justification for the analyst's claim that his Wald confidence interval is valid.

More precisely, we shall construct $\alpha^\dagger$-level tests of the null hypothesis $\mathsf{CSBias}_\theta(\widehat{\psi}_1) < \mathsf{s.e.}_\theta(\widehat{\psi}_1)\delta$ via HOIF tests and estimators introduced in Mukherjee et al. (2017); Robins et al. (2008, 2017). In particular,

[3] In this paper, essentially all expectations and probabilities are to be understood as being conditional on the training sample (except when explicitly noted otherwise). Hence we can and do omit this conditioning event in our notation. The randomness in the training sample will also be ignored.

the null hypothesis $\mathsf{CSBias}_\theta(\widehat{\psi}_1) < \mathsf{s.e.}_\theta(\widehat{\psi}_1)\delta$, if true, guarantees that the nominal $(1-\alpha)$ Wald confidence interval covers $\psi(\theta)$ with probability at least

$$\Phi(z_{\alpha/2} - \delta) - \Phi(-z_{\alpha/2} - \delta)$$[4]

in large sample. For example, when $\alpha = 0.10, \delta = 3/4$, $\Phi(z_{\alpha/2} - \delta) - \Phi(-z_{\alpha/2} - \delta) \approx 80\%$. Hence, if we reject this null hypothesis, we conclude that the actual coverage of a nominal 90% interval is less than 80%.

We refer the interested readers to works such as Robins et al. (2008, 2016); van der Vaart (2014). The HOIF tests and estimators we shall use are higher order U-statistics; in particular, an $m$-th order influence function is an $m$-th order U-statistic. To construct these tests we require access to the study data and the functions $\widehat{b}$ and $\widehat{p}$ used by the analyst. However, our tests are constructed without: i) refitting, modifying, or even having knowledge of the ML algorithms that have been employed to compute $\widehat{b}, \widehat{p}$ from the training sample 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 $b$ or propensity score $1/p$.

We will show there is an unavoidable limitation to what can be achieved with our or any other method. No test, including ours, of the null hypothesis $\mathsf{CSBias}_\theta(\widehat{\psi}_1) < \mathsf{s.e.}_\theta(\widehat{\psi}_1)\delta$ can be uniformly (in $\theta$) consistent [without making additional complexity reducing assumptions]. Thus, when our $\alpha^\dagger$-level test rejects this null hypothesis for $\alpha^\dagger$ small, we have strong evidence that the null hypothesis is false; nonetheless when the test does not reject, we cannot conclude that there is good evidence that $\mathsf{CSBias}_\theta(\widehat{\psi}_1) < \mathsf{s.e.}_\theta(\widehat{\psi}_1)\delta$, no matter how large the sample size. This reflects the fact that, because we make (essentially) no assumptions, $\mathsf{CSBias}_\theta(\widehat{\psi}_1)$ may be order 1 and thus $n^{1/2}$ times greater than $\mathsf{s.e.}_\theta(\widehat{\psi}_1)$ under the law $\mathsf{P}_\theta$ generating the data and yet our $\alpha^\dagger$-level test, as we show in Sections 3 and 4, may be powerless to detect this fact!

*Our Approach.* All claims made in this subsection will be proved later in the paper. DML estimators are based on the influence function of the DR functional $\psi(\theta)$ (van der Vaart, 2002, 1998). Our proposed approach begins by computing a second order influence function estimator $\widehat{\mathbb{IF}}_{22,k}$ (defined later) of the estimable part of the conditional bias $\mathsf{E}_\theta[\widehat{\psi}_1 - \psi(\theta)|\mathbf{O}_{\mathsf{tr}}]$ of $\widehat{\psi}_1$ given $\mathbf{O}_{\mathsf{tr}}$. Our bias corrected estimator is $\widehat{\psi}_{2,k} \equiv \widehat{\psi}_1 - \widehat{\mathbb{IF}}_{22,k}$. The statistic $\widehat{\mathbb{IF}}_{22,k}$ is a second-order U-statistic that depends on a choice of $k$ (with $k = o(n^2)$ for reasons explained in Comment 3.4), a vector of basis functions $\bar{Z}_k \equiv \bar{\mathsf{z}}_k(X) \equiv (z_1(X), \ldots, z_k(X))^\top$ of the d-dimensional vector of potential confounders $X$ and an estimator $\widehat{\Sigma}_k^{-1}$ computed from the training sample of the inverse expected weighted outer product $\Sigma_k^{-1} := \{\mathsf{E}_\theta[\lambda(X)\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top]\}^{-1}$, with weight defined as

$$\lambda(x) \equiv \lambda(x;\theta) := \mathsf{E}_\theta[S_{bp}|X = x].$$

[4]We use $\Phi$ to denote standard normal cdf

In this paper we choose $\widehat{\Sigma}_k^{-1} = \mathsf{P}_{n_{\mathrm{tr}}}\left[S_{bp}\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top\right]$ to be the inverse of the average over the training sample of $S_{bp}\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top$ unless noted otherwise. Henceforth we denote $\widehat{\mathbb{IF}}_{22,k}$ and $\widehat{\psi}_{2,k}$ as $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$ and $\widehat{\psi}_{2,k}(\widehat{\Sigma}_k^{-1})$ to make explicit the dependence on the estimator $\widehat{\Sigma}_k^{-1}$. The oracle estimators $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ and $\widehat{\psi}_{2,k}(\Sigma_k^{-1})$ replace $\widehat{\Sigma}_k^{-1}$ by the true $\Sigma_k^{-1}$. The oracle estimators are generally infeasible. The estimators $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$, $\widehat{\psi}_{2,k}(\Sigma_k^{-1})$, $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$ and $\widehat{\psi}_{2,k}(\widehat{\Sigma}_k^{-1})$ will be asymptotically normal conditional on the training sample when, as in our asymptotic set-up, $k = k(n) \to \infty$ and $k = o(n^2)$ as $n \to \infty$. Furthermore, the variance of $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ and $\widehat{\psi}_{2,k}(\Sigma_k^{-1})$ are at most of order $1/n$.

The fraction of the conditional bias of $\widehat{\psi}_1$ corrected by $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$ depends critically on (i) the choice of $k$, (ii) the accuracy of the estimator $\widehat{\Sigma}_k^{-1}$ of $\Sigma_k^{-1}$, and (iii) the particular $k$-vector of (basis) functions $\bar{\mathsf{Z}}_k \equiv \bar{\mathsf{z}}_k(X)$ chosen from a much larger, possibly countably infinite, dictionary of candidate functions. Data adaptive choice of $\bar{\mathsf{z}}_k(x)$ to optimize the power of our test is an important open problem that is beyond the scope of the current paper.

In this paper we shall always choose $k$ to be less than the sample size $n = N/2$ of the estimation sample for the following three reasons: $k < n$ is necessary (i) for $\widehat{\mathbb{IF}}_{22,k}$'s standard error of order $\max\{k^{1/2}/n, 1/\sqrt{n}\}$ to be smaller than or equal to the order $n^{-1/2}$ of the standard error of $\widehat{\psi}_1$, thereby creating the possibility of detecting, for any given $\delta > 0$, that the (absolute value) of the ratio of the estimable part of the bias of $\widehat{\psi}_1$ to its standard error exceeds $\delta$, when the sample size $n$ is sufficiently large , (ii) to be able to use the inverse of the sample covariance matrix $\widehat{\Sigma}_k^{-1}$ computed from the training sample to estimate $\Sigma_k^{-1}$ accurately without imposing, possibly incorrect, additional smoothness or sparsity assumptions required for accurate estimation when $k \geqslant n$, and iii) for bias-corrected confidence intervals centered at $\widehat{\psi}_{2,k}(\widehat{\Sigma}_k^{-1}) \equiv \widehat{\psi}_1 - \widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$ to have length of the same order of $n^{-1/2}$ as $\widehat{\psi}_1$.

The importance of (i) is two-fold. First, if $|\mathsf{E}_\theta[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]| \geqslant \mathsf{s.e.}_\theta(\widehat{\psi}_1)\delta$, then under faithfulness, the analysts nominal $(1-\alpha)$ Wald confidence interval is invalid. Further $|\mathsf{E}_\theta[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]| \geqslant \mathsf{s.e.}_\theta(\widehat{\psi}_1)\delta$ implies $\mathsf{CSBias}_\theta(\widehat{\psi}_1) \geqslant \mathsf{s.e.}_\theta(\widehat{\psi}_1)\delta$. Thus, even without the faithfulness assumption, $|\mathsf{E}_\theta[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]| \geqslant \mathsf{s.e.}_\theta(\widehat{\psi}_1)\delta$ falsifies the analyst's justification for the validity of his nominal $(1-\alpha)$ Wald interval, if his justification needed $\mathsf{CSBias}_\theta(\widehat{\psi}_1)$ be $o(n^{-1/2})$.

*Organization of the paper.* The remainder of the paper is organized as follows. Section 2 reviews needed results on DR functionals obtained by Rotnitzky et al. (2019), the state of the art DML estimators, and the statistical properties of these estimators.

In Section 3, we study and derive theoretical statistical properties of the following: the oracle estimator $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ of the bias of the DML estimator $\widehat{\psi}_1$, the oracle biased-corrected estimator $\widehat{\psi}_{2,k}(\Sigma_k^{-1})$, the $\alpha^\dagger$-level oracle test $\widehat{\chi}_{2,k}(\Sigma_k^{-1}; z_{\alpha^\dagger/2}, \delta)$ of the null hypothesis $|\mathsf{E}_\theta[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]| \geqslant \mathsf{s.e.}_\theta(\widehat{\psi}_1)\delta$.

In Section 4, we no longer assume $\Sigma_k^{-1}$ is known and replace it by $\widehat{\Sigma}_k^{-1}$ computed from the training sample. We show that, in contrast with the oracle test, the test $\widehat{\chi}_{2,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^\dagger/2}, \delta)$ that replaces $\Sigma_k$ by its training sample estimate $\widehat{\Sigma}_k$ fails to protect the $\alpha$-level when testing $\mathsf{CSBias}_\theta(\widehat{\psi}_1) > \mathsf{s.e.}_\theta(\widehat{\psi}_1)\delta$ for a fixed positive $\delta$. However we show that, under additional conditions, $\widehat{\chi}_{3,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^\dagger/2}, \delta)$ based on a third order U-statistic recovers the statistical properties of the oracle test $\widehat{\chi}_{2,k}(\Sigma_k^{-1}; z_{\alpha^\dagger/2}, \delta)$.

In Section 5, we extend the derivation of HOIFs for parameters covered in Robins et al. (2008) to the HOIFs for the functionals in the MBF class (i.e. DR functionals). We study the statistical properties of HOIFs of DR functional, in terms of their biases, variances and asymptotic distributions.

In Section 6, we use the results of Section 5 to construct a new test of the null hypothesis $\mathsf{CSBias}_\theta(\widehat{\psi}_1) < \mathsf{s.e.}_\theta(\widehat{\psi}_1)\delta$ that *almost* has the level and power properties of the oracle test in Section 4, with a much relaxed version of the additional conditions needed for the validity of $\widehat{\chi}_{3,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^\dagger/2}, \delta)$. However, this new test is based on HOIF associated U-statistics whose order increases with sample size and thus may be computationally infeasible. To handle this difficulty, we propose an "early stopping" procedure that can save computational cost while still protecting the level of the test.

In Section 7 we present results of simulation studies to evaluate the finite sample performance of our methods. Section 8 concludes with discussion of some open problems. Many of the technical details are deferred to the on-line appendix.

We next define the observed data structure and collect some frequently used notation.

1.1. **Observation schemes and notations.** As mentioned above, we observe $N$ i.i.d. copies of data $O = (W, X)$ drawn from some unknown probability distribution $\mathsf{P}_\theta$ belonging to a locally nonparametric model

$$\mathcal{M} = \left\{\mathsf{P}_\theta; \theta = (b, p, \theta_{\backslash(b,p)}) \in \Theta = \mathcal{B} \times \mathcal{P} \times \Theta_{\backslash(\mathcal{B},\mathcal{P})}\right\}$$

parameterized by the parameter $\theta = (b, p, \theta_{\backslash(b,p)})$ with $b, p, \theta_{\backslash(b,p)}$ variation independent. Here the maps $b : x \mapsto b(x) \in \mathcal{B}$ and $p : x \mapsto p(x) \in \mathcal{P}$ have range contained in $\mathbb{R}$.

To avoid extraneous technical issues, we assume $W$ is bounded with $\mathsf{P}_\theta$-probability 1 and that $X$ is a $d$-dimensional random vector with compact support whose density is bounded away from 0 and $\infty$ on its support. We use $\mathsf{E}_\theta[\cdot], \mathsf{var}_\theta[\cdot]$ and etc. to denote the expectation, variance, and etc. with respect to $\mathsf{P}_\theta$. The data is randomly divided into an the estimation sample and a training (equivalently, nuisance) sample of size $n = N/2$. To avoid notational clutter, all expectations, variances, and probabilities are conditional on the training sample unless otherwise stated. Hence when we write $\mathsf{E}_\theta[X]$ we mean $\mathsf{E}_\theta[X|\mathbf{O}_{\mathsf{tr}}]$ with $\mathbf{O}_{\mathsf{tr}}$ the training sample data. For a (random) vector $V$, $\|V\|_\theta \equiv \mathsf{E}_\theta[V^{\otimes 2}]^{1/2} = \mathsf{E}_\theta[V^\top V]^{1/2}$ denotes its $L_2(\mathsf{P}_\theta)$ norm conditioning on the training sample, $\|V\| \equiv (V^{\otimes 2})^{1/2} = (V^\top V)^{1/2}$ denotes its $\ell_2$ norm and $\|V\|_\infty$ denotes its $L_\infty(\mathsf{P}_\theta)$ norm. For any matrix $A$, $\|A\|$ will be reserved for its operator norm. Given an integer $k$, and a random vector $\bar{\mathsf{Z}}_k = \bar{\mathsf{z}}_k(X)$, $\Pi_\theta\left[\cdot|\bar{\mathsf{Z}}_k\right]$ denotes the population linear projection operator onto the space spanned by $\bar{\mathsf{Z}}_k$ conditioning on the training sample, and $\Pi_\theta^\perp\left[\cdot|\bar{\mathsf{Z}}_k\right] = (\mathsf{I} - \Pi_\theta)\left[\cdot|\bar{\mathsf{Z}}_k\right]$ is the projection onto the orthogonal complement of $\bar{\mathsf{Z}}_k$ in the Hilbert space $L_2\left(\mathsf{P}_{\theta,X}\right)$ where $\mathsf{P}_{\theta,X}$ denotes the law of the $X$ under $\theta$. That is, for a random variable $V$,

$$\Pi_\theta\left[V|\bar{\mathsf{Z}}_k\right]=\bar{\mathsf{Z}}_k^\top\left(\mathsf{E}_\theta[\bar{\mathsf{Z}}_k\bar{\mathsf{Z}}_k^\top]\right)^{-1}\mathsf{E}_\theta\left[\bar{\mathsf{Z}}_kV\right],\Pi_\theta^\perp\left[V|\bar{\mathsf{Z}}_k\right]=V-\Pi_\theta\left[V|\bar{\mathsf{Z}}_k\right]. \tag{1.2}$$

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

## 2. DR functionals and the statistical properties of DML estimators

We begin this section by reviewing some useful results on DR functionals (see Definition 1.1) established in Rotnitzky et al. (2019). First, the theorem below characterizes the (nonparametric) influence functions of DR functionals.

**Theorem 2.1** (Theorem 1 of Rotnitzky et al. (2019))**.** *Suppose $\psi(\theta)$ for $\theta=\left(b,p,\theta_{\backslash(b,p)}\right)\in\Theta\equiv\mathcal{B}\times\mathcal{P}\times\Theta_{\backslash(\mathcal{B},\mathcal{P})}$ is a DR functional (equivalently MB functional) according to Definition 1.1; further the regularity Condition 1 of Rotnitzky et al. (2019) holds. Then there exist a statistic $S_0$ and maps $h\mapsto m_1(O,h)$ for $h\in\mathcal{B}$ and $h\mapsto m_2(O,h)$ for $h\in\mathcal{P}$ independent of $\theta$ satisfying the following:*

- *the maps $h\in\mathcal{B}\mapsto\mathsf{E}_\theta[m_1(O,h)]$ and $h\in\mathcal{P}\mapsto\mathsf{E}_\theta[m_2(O,h)]$ are linear;*
- $$\psi(\theta)=\mathsf{E}_\theta[m_1(O,b)]+\mathsf{E}_\theta[S_0]=\mathsf{E}_\theta[m_2(O,p)]+\mathsf{E}_\theta[S_0];$$
- *and*

$$\mathsf{IF}_{1,\psi}(\theta)=\mathcal{H}(b,p)-\psi(\theta) \tag{2.1}$$

*where $\mathcal{H}(b,p):=S_{bp}b(X)p(X)+m_1(O,b)+m_2(O,p)+S_0$.*

*Furthermore*

$$\begin{cases}\mathsf{E}_\theta[S_{bp}h(X)p(X)+m_1(O,h)]=0 \text{ for all } h\in\mathcal{B},\\ \mathsf{E}_\theta[S_{bp}b(X)h(X)+m_2(O,h)]=0 \text{ for all } h\in\mathcal{P}.\end{cases} \tag{2.2}$$

*In addition, with $\mathsf{P}_{\theta,X}$ denoting the marginal law of $X$ under $\mathsf{P}_\theta$, suppose for all $h\in L_2(\mathsf{P}_{\theta,X})$, $m_1(O,h)$ and $m_2(O,h)$ are in $L_2(\mathsf{P}_\theta)$. Then the maps $h\in\mathcal{B}\mapsto m_1(O,h)$ and $h\in\mathcal{P}\mapsto m_2(O,h)$ are linear.*

Under very slightly stronger regularity conditions on the maps $m_1$ and $m_2$, the converse of Theorem 2.1 also holds, together with several additional results.

**Theorem 2.2** (Theorem 2 of Rotnitzky et al. (2019))**.** *Suppose the regularity Condition 1 of Rotnitzky et al. (2019) holds. Suppose a functional $\psi(\theta)$ has influence function $\mathsf{IF}_{1,\psi}(\theta)$ of the form (2.1) for $m_1$ and $m_2$ such that all $\theta = \left(b, p, \theta_{\setminus(b,p)}\right) \in \Theta$, the maps $h \mapsto \mathsf{E}_\theta[m_1(O,h)]$ and $h \mapsto \mathsf{E}_\theta[m_2(O,h)]$ for $h \in L_2(\mathsf{P}_{\theta,X})$ are continuous and linear with Riesz representers $\mathcal{R}_1(X) \equiv \mathcal{R}_1(X;\theta)$ and $\mathcal{R}_2(X) \equiv \mathcal{R}_2(X;\theta)$[5] respectively. Moreover, suppose that $b(X)$, $p(X)$, $\lambda(X)b(X)$ and $\lambda(X)p(X)$ belong to $L_2(\mathsf{P}_{\theta,X})$ for all $\theta \in \Theta$, where $\lambda(X) = \mathsf{E}_\theta[S_{bp}|X]$. Then, for all $\theta = \left(b, p, \theta_{\setminus(b,p)}\right) \in \Theta$*

(1) *the identity (1.1) holds for any $\theta'$ and hence $\psi(\theta)$ is an DR functional (equivalently MB functional);*

(2) *for all $h \in L_2(\mathsf{P}_{\theta,X})$*

$$\mathsf{E}_\theta[S_{bp}bh + m_2(O,h)] = 0 \quad \text{and} \quad \mathsf{E}_\theta[S_{bp}hp + m_1(O,h)] = 0.$$

(3) *$b(X) = -\mathcal{R}_2(X)/\lambda(X)$ and $p(X) = -\mathcal{R}_1(X)/\lambda(X)$.*

(4) *$\psi(\theta) = \mathsf{E}_\theta\left[m_2(O,p) + S_0\right] = \mathsf{E}_\theta\left[m_1(O,b) + S_0\right] = \mathsf{E}_\theta\left[-S_{bp}bp + S_0\right]$.*

(5) *$b$ and $p$ are the (global) minimizers of the following minimization problems*

$$b = \arg\min_{h \in L_2(\mathsf{P}_{\theta,X})} \mathsf{E}_\theta\left[S_{bp}\frac{h^2}{2} + m_2(O,h)\right], p = \arg\min_{h \in L_2(\mathsf{P}_{\theta,X})} \mathsf{E}_\theta\left[S_{bp}\frac{h^2}{2} + m_1(O,h)\right].$$

We now give the influence functions $\mathsf{IF}_{1,\psi}(\theta)$ and Riesz representers for several DR functionals.

**Example 1** (Some examples of Riesz representers for DR functionals)**.**

(1) $\psi(\theta) = -\mathsf{E}_\theta[b(X)] = -\mathsf{E}_\theta[Y(a=1)]$. Then $\mathsf{IF}_{1,\psi}(\theta) = \mathcal{H}(b,p) - \psi(\theta)$ with

$$\mathcal{H}(b,p) = -\left(-Ab(X)p(X) + b(X) + AYp(X)\right).$$

Here $S_{bp} = A$ and $\lambda(X) = p(X)^{-1}$, $m_1(O,b) = -b(X)$, $m_2(O,p) = -AYp(X)$ and $S_0 = 0$. Then $\mathcal{R}_1(X) = -1$ and $\mathcal{R}_2(X) = -b(X)/p(X)$ since

$$\begin{cases} \mathsf{E}_\theta\left[m_1(O,h)\right] = \mathsf{E}_\theta[-h(X)], \\ \mathsf{E}_\theta\left[m_2(O,h)\right] = \mathsf{E}_\theta[-AYh(X)] = \mathsf{E}_\theta\left[-\left\{p(X)\right\}^{-1} b(X)h(X)\right]. \end{cases}$$

(2) $\psi(\theta) = \mathsf{E}_\theta\left[(Y - b(X))(A - p(X))\right]$ – expected conditional covariance between $A$ and $Y$ given $X$ with $b(X) = \mathsf{E}_\theta[Y|X], p(X) = \mathsf{E}_\theta[A|X]$: $\mathsf{IF}_{1,\psi}(\theta) = \mathcal{H}(b,p) - \psi(\theta)$ with $\mathcal{H}(b,p) = b(X)p(X) - Ab(X) - Yp(X) + AY$, $S_{bp} = 1$ and $\lambda(X) = 1$, $m_1(O,b) = -Ab(X)$, $m_2(O,p) = -Yp(X)$, and $S_0 = AY$. Then $\mathcal{R}_1(X) = -p(X)$ and $\mathcal{R}_2(X) = -b(X)$ since

$$\begin{cases} \mathsf{E}_\theta\left[m_1(O,h)\right] = \mathsf{E}_\theta[-Ah(X)] = \mathsf{E}_\theta[-p(X)h(X)], \\ \mathsf{E}_\theta\left[m_2(O,h)\right] = \mathsf{E}_\theta[-Yh(X)] = \mathsf{E}_\theta[-b(X)h(X)]. \end{cases}$$

[5] The Riesz representer $\mathcal{R}(X)$ of a continuous linear functional $h \mapsto E[m(O,h)]$ for $h \in L_2(\mathsf{P}_X)$ is, by definition, the function of $X$ satisfying $\mathsf{E}[m(O,h)] = \mathsf{E}\left[\mathcal{R}(X)h(X)\right]$; see Chernozhukov et al. (2018b,c); Rotnitzky et al. (2019).

The following algorithm defines the DML estimators $\widehat{\psi}_1$ and $\widehat{\psi}_{\mathsf{cf},1}$ of a DR functional $\psi(\theta)$ satisfying the suppositions of Theorem 2.2.

(i) The $N$ study subjects are randomly split into two 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) Since $b$ and $p$ are generally unknown, their estimates $\widehat{b}$ and $\widehat{p}$ are constructed from the training sample data using ML methods, such as deep neural networks (Chen et al., 2020; Farrell et al., 2020) and define $\widehat{\theta} \coloneqq \left(\widehat{b}, \widehat{p}, \theta_{\backslash(b,p)}\right)$. [Note that unlike $(\widehat{b}, \widehat{p})$, $\theta_{\backslash(b,p)}$ in the definition of $\widehat{\theta}$ is not estimated from data.]

(iii) Denote $\mathbb{IF}_1 \equiv \mathbb{IF}_1(\theta) \coloneqq \frac{1}{n}\sum_{i=1}^n \mathsf{IF}_{1,\psi}(\theta)$ and $\widehat{\mathbb{IF}}_1 \equiv \mathbb{IF}_1(\widehat{\theta}) \coloneqq \frac{1}{n}\sum_{i=1}^n \mathsf{IF}_{1,\psi}(\widehat{\theta})$. Then

$$\begin{aligned}\widehat{\psi}_1 = \widehat{\mathbb{IF}}_1 + \psi(\widehat{\theta}) &= \frac{1}{n}\sum_{i=1}^n \mathcal{H}_i(\widehat{b}, \widehat{p}) \\ &= \frac{1}{n}\sum_{i=1}^n \left(S_{bp,i}\widehat{b}(X_i)\widehat{p}(X_i) + m_1(O_i, \widehat{b}) + m_2(O_i, \widehat{p}) + S_{0,i}\right)\end{aligned}$$

from $n$ subjects in the estimation sample and

$$\widehat{\psi}_{\mathsf{cf},1} = \frac{1}{2}\left(\widehat{\psi}_1 + \overline{\widehat{\psi}}_1\right)$$

where $\overline{\widehat{\psi}}_1$ is $\widehat{\psi}_1$ but with the training and estimation samples reversed. [Note that $\widehat{\psi}_1 = \frac{1}{n}\sum_{i=1}^n \mathcal{H}_i(\widehat{b}, \widehat{p})$ does not depend on $\theta_{\backslash(b,p)}$ and depends on the training sample data only through the functions $\widehat{b}$ and $\widehat{p}$.]

**Comment 2.1.** *In this paper, we only need $\widehat{\theta}$ to depend on estimated $b$ and $p$ because none of the estimators/tests in this paper will depend on estimates of $\theta_{\backslash\{b,p\}}$, except for the infeasible oracle estimators/tests in Section 3.*

Then the following theorem provides conditional and unconditional asymptotic properties of $\widehat{\psi}_1$.

**Theorem 2.3.** *Under the conditions of Theorem 2.2, conditional on the training sample, $\widehat{\psi}_1$ is asymptotically normal with conditional bias*

$$\mathsf{Bias}_\theta(\widehat{\psi}_1) \coloneqq \mathsf{E}_\theta\left[\widehat{\psi}_1 - \psi(\theta)\right] = \mathsf{E}_\theta\left[S_{bp}\left\{b(X) - \widehat{b}(X)\right\}\left\{p(X) - \widehat{p}(X)\right\}\right]. \tag{2.3}$$

*If further, a) $\mathsf{Bias}_\theta(\widehat{\psi}_1)$ is $o(n^{-1/2})$ and b) $\widehat{b}(x)$ and $\widehat{p}(x)$ converge to $b(x)$ and $p(x)$ in $L_2(\mathsf{P}_\theta)$, then*

*(1)*

$$\begin{aligned}\widehat{\psi}_1 - \psi(\theta) &= n^{-1}\sum_{i=1}^n \mathsf{IF}_{1,\psi,i}(\theta) + o(n^{-1/2}) \\ \widehat{\psi}_{\mathsf{cf},1} - \psi(\theta) &= N^{-1}\sum_{i=1}^N \mathsf{IF}_{1,\psi,i}(\theta) + o(N^{-1/2}).\end{aligned}$$

*Further* $n^{1/2}(\widehat{\psi}_1 - \psi(\theta))$ *converges conditionally and unconditionally to a normal distribution with mean zero;* $\widehat{\psi}_{\mathsf{cf},1}$ *is a regular, asymptotically linear estimator; i.e.,* $N^{1/2}(\widehat{\psi}_{\mathsf{cf},1} - \psi(\theta))$ *converges unconditionally to a normal distribution with mean zero and variance equal to the semiparametric variance bound* $\mathsf{var}_\theta\left[\mathsf{IF}_{1,\psi}(\theta)\right]$ *(Bickel et al., 1998; Newey, 1990).*

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

$$\mathsf{CI}_\alpha(\widehat{\psi}_1) \coloneqq \widehat{\psi}_1 \pm z_{\alpha/2}\widehat{\mathsf{s.e.}}[\widehat{\psi}_1] \tag{2.4}$$

$$\mathsf{CI}_\alpha(\widehat{\psi}_{\mathsf{cf},1}) \coloneqq \widehat{\psi}_{\mathsf{cf},1} \pm z_{\alpha/2}\widehat{\mathsf{s.e.}}[\widehat{\psi}_{\mathsf{cf},1}] \tag{2.5}$$

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

$$\widehat{\mathsf{var}}[\widehat{\psi}_1] = \frac{1}{n^2}\sum_{i=1}^{n}\left(\mathsf{IF}_{1,\psi,i}(\widehat{\theta}) - \frac{1}{n}\sum_{i=1}^{n}\mathsf{IF}_{1,\psi,i}(\widehat{\theta})\right)^2$$
$$\widehat{\mathsf{var}}[\widehat{\psi}_{\mathsf{cf},1}] = \frac{1}{4}\{\widehat{\mathsf{var}}[\widehat{\psi}_1] + \widehat{\mathsf{var}}[\overline{\widehat{\psi}}_1]\}.$$

The proof of Theorem 2.3 is straightforward and can be found in Chernozhukov et al. (2018a), Smucler et al. (2019) or Farrell et al. (2020, Theorem 3). However, in reality, it is often not enough just to invoke Theorem 2.3 and claim the validity of $\mathsf{CI}_\alpha(\widehat{\psi}_1)$. First, owing to the fact that we make no complexity reducing assumptions on $b$ and $p$ it follows that because of their asymptotic nature, there is no finite sample size $n$ at which any test could empirically reject either the hypothesis $\mathsf{Bias}_\theta(\widehat{\psi}_1) = o(n^{-1/2})$ or the hypothesis that $n^{1/2}(\widehat{\psi}_1 - \psi(\theta))$ 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 observed sample size $n$, the actual conditional coverage of the nominal $(1-\alpha)$ two-sided interval $\widehat{\psi}_1 \pm z_{\alpha/2}\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]$ under the distribution $\mathsf{P}_\theta$ that generated the data is less than a fraction $\varrho < 1$ of its *nominal coverage*. Second, since the ML estimators $\widehat{b}(x)$ and $\widehat{p}(x)$ have unknown statistical properties and we are interested in developing assumption-lean statistical procedure relying only on empirical evidence, our inferential statements regard the training sample as fixed rather than random. In particular, the randomness referred to in subsequent sections is that of the estimation sample. In fact, our inferences will rely on being in 'asymptopia', but 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. $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ defined later) are close to the quantiles of a normal. Construction of a non-asymptotic version of our proposed procedure relying on concentration or deviation inequalities is an important open problem.

## 3. The oracle procedure when $\Sigma_k^{-1} \coloneqq \{\mathsf{E}_\theta[\lambda(X)\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top]\}^{-1}$ is known

It is pedagogically useful to first consider an oracle procedure that would only be implementable if $\Sigma_k^{-1}$ were known. We will focus on such situation in this section.

3.1. **Truncated parameters and their unbiased estimators.** Though there does not exist uniformly (in $\theta$) consistent test of the null hypothesis that the bias $\mathsf{Bias}_\theta(\widehat{\psi}_1)$ of $\widehat{\psi}_1$ for estimating $\psi(\theta)$ is smaller than a fraction $\delta$ of its standard error $\mathsf{s.e.}_\theta(\widehat{\psi}_1)$, this is not the case for the bias of $\widehat{\psi}_1$ as an estimator of the truncated parameter $\widetilde{\psi}_k(\theta)$ of $\psi(\theta)$ (Robins et al. (2008, Section 3.2)), defined next. Recall from Section 1 that the bias of $\widehat{\psi}_1$ for $\widetilde{\psi}_k(\theta)$ is the estimable part of the bias of $\widehat{\psi}_1$ for $\psi(\theta)$.

**Definition 3.1** (The truncated parameter $\widetilde{\psi}_k(\theta)$, truncation bias $\mathsf{TB}_k(\theta)$, and the projected bias $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$)**.** *Given estimators $\widehat{b} \in \mathcal{B}$ and $\widehat{p} \in \mathcal{P}$ of $b$ and $p$ computed from the training sample and some $k$ dimensional basis functions $\bar{\mathsf{z}}_k(x)$ in $L_2(\mathsf{P}_{\theta,X})$:*

- *Define the (conditional) truncated parameter of a DR functional $\psi(\theta)$ as*

$$\widetilde{\psi}_k(\theta) \coloneqq \mathsf{E}_\theta\left[\mathcal{H}\left(\widetilde{b}_{k,\theta}, \widetilde{p}_{k,\theta}\right)\right],$$

*where $\widetilde{b}_{k,\theta}$ and $\widetilde{p}_{k,\theta}$ are defined as follows. Consider the working linear models*

$$\begin{cases} b_k^*(X;\bar{\zeta}_{b,k}) = \widehat{b}(X) + \bar{\zeta}_{b,k}^\top \bar{\mathsf{z}}_k(X), \\ p_k^*(X;\bar{\zeta}_{p,k}) = \widehat{p}(X) + \bar{\zeta}_{p,k}^\top \bar{\mathsf{z}}_k(X). \end{cases}$$

*Define $\widetilde{\bar{\zeta}}_b(\theta)$ and $\widetilde{\bar{\zeta}}_p(\theta)$ as the solutions to the following system of equations*

$$\begin{cases} 0 = \mathsf{E}_\theta\left[\dfrac{\partial \mathcal{H}\left(b_k^*(X;\bar{\zeta}_{b,k}), p^*(X;\bar{\zeta}_{p,k})\right)}{\partial \bar{\zeta}_{b,k}}\right] = \mathsf{E}_\theta\left[S_{bp} p_k^*(X;\bar{\zeta}_{p,k})\bar{\mathsf{z}}_k(X) + m_1(O,\bar{\mathsf{z}}_k)\right] \\ 0 = \mathsf{E}_\theta\left[\dfrac{\partial \mathcal{H}\left(b_k^*(X;\bar{\zeta}_{b,k}), p^*(X;\bar{\zeta}_{p,k})\right)}{\partial \bar{\zeta}_{p,k}}\right] = \mathsf{E}_\theta\left[S_{bp} b_k^*(X;\bar{\zeta}_{b,k})\bar{\mathsf{z}}_k(X) + m_2(O,\bar{\mathsf{z}}_k)\right]. \end{cases}$$

*where, for $j = 1, 2$, $m_j(O,\bar{\mathsf{z}}_k) = (m_j(O,z_1), \ldots, m_j(O,z_k))$. Hence*

$$\begin{cases} \widetilde{\bar{\zeta}}_{b,k}(\theta) = -\Sigma_k^{-1}\mathsf{E}_\theta\left[\left(S_{bp}\widehat{b}(X)\bar{\mathsf{z}}_k(X) + m_2(O,\bar{\mathsf{z}}_k)\right)\right], \\ \widetilde{\bar{\zeta}}_{p,k}(\theta) = -\Sigma_k^{-1}\mathsf{E}_\theta\left[(S_{bp}\widehat{p}(X)\bar{\mathsf{z}}_k(X) + m_1(O,\bar{\mathsf{z}}_k))\right] \end{cases} \tag{3.1}$$

*where*

$$\Sigma_k \coloneqq \mathsf{E}_\theta\left[S_{bp}\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top\right] \equiv \mathsf{E}_\theta\left[\lambda(X)\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top\right].$$

*Now define*

$$\begin{cases} \widetilde{b}_{k,\theta}(X) \coloneqq b_k^*(X;\widetilde{\bar{\zeta}}_{b,k}(\theta)) \equiv \widehat{b}(X) + \widetilde{\bar{\zeta}}_{b,k}(\theta)^\top \bar{\mathsf{z}}_k(X), \\ \widetilde{p}_{k,\theta}(X) \coloneqq p_k^*(X;\widetilde{\bar{\zeta}}_{p,k}(\theta)) \equiv \widehat{p}(X) + \widetilde{\bar{\zeta}}_{p,k}(\theta)^\top \bar{\mathsf{z}}_k(X). \end{cases}$$

- *Define the difference between $\widetilde{\psi}_k(\theta)$ and $\psi(\theta)$ as the truncation bias $\mathsf{TB}_k(\theta)$:*

$$\mathsf{TB}_k(\theta) \coloneqq \widetilde{\psi}_k(\theta) - \psi(\theta) = \mathsf{E}_\theta\left[S_{bp}\left(\widetilde{b}_{k,\theta}(X) - b(X)\right)\left(\widetilde{p}_{k,\theta}(X) - p(X)\right)\right]. \tag{3.2}$$

- *Define the bias of $\widehat{\psi}_1$ as an estimator of $\widetilde{\psi}_k(\theta)$ as*

$$\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1) \coloneqq \mathsf{E}_\theta\left[\widehat{\psi}_1 - \widetilde{\psi}_k(\theta)\right]. \tag{3.3}$$

By definition, $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1) = \mathsf{Bias}_\theta(\widehat{\psi}_1) - \mathsf{TB}_k(\theta)$. Combining equation (3.2), we have $\widetilde{\psi}_k(\theta) - \psi(\theta) \equiv \mathsf{Bias}_\theta(\widehat{\psi}_1) - \mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$. As discussed further below, $\mathsf{TB}_k(\theta)$ is not consistently estimable without imposing smoothness/sparsity assumptions on the functions $b$ and $p$. But $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$ can be unbiasedly estimated by the following oracle (i.e. $\Sigma_k^{-1}$ known) second-order U-statistic:

$$\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1}) \coloneqq \frac{1}{n(n-1)} \sum_{1 \leqslant i_1 \neq i_2 \leqslant n} \widehat{\mathsf{IF}}_{22,k,\bar{\imath}_2}(\Sigma_k^{-1})$$

where

$$\widehat{\mathsf{IF}}_{22,k,\bar{\imath}_2}(\Sigma_k^{-1}) \coloneqq \left[\mathcal{E}_{\widehat{b},m_2}(\bar{\mathsf{z}}_k)(O)\right]_{i_1}^\top \Sigma_k^{-1} \left[\mathcal{E}_{\widehat{p},m_1}(\bar{\mathsf{z}}_k)(O)\right]_{i_2}.$$

Here $\mathcal{E}_{1,b'}(h)$ and $\mathcal{E}_{2,p'}(h)$ are linear functionals defined as: given any $h \in L_2(\mathsf{P}_{\theta,X})$, $b' \in \mathcal{B}$ and $p' \in \mathcal{P}$

$$\mathcal{E}_{b',m_2}(h)(O) \coloneqq S_{bp} b'(X) h(X) + m_2(O, h) \tag{3.4}$$

and

$$\mathcal{E}_{p',m_1}(h)(O) \coloneqq S_{bp} p'(X) h(X) + m_1(O, h). \tag{3.5}$$

When the range of $h$ is multidimensional (e.g. $h = \bar{\mathsf{z}}_k$), the linear functionals $\mathcal{E}_{1,b'}$ and $\mathcal{E}_{2,p'}$ act on $h$ component-wise.

For the examples covered in Example 1, $\mathcal{E}_{\widehat{b},m_2}(\bar{\mathsf{z}}_k)(O)$ and $\mathcal{E}_{\widehat{p},m_1}(\bar{\mathsf{z}}_k)(O)$ are

(1) Example 1(1):

$$\begin{cases} \mathcal{E}_{\widehat{b},m_2}(\bar{\mathsf{z}}_k)(O) = A\widehat{b}(X)\bar{\mathsf{z}}_k(X) - AY\bar{\mathsf{z}}_k(X), \\ \mathcal{E}_{\widehat{p},m_1}(\bar{\mathsf{z}}_k)(O) = A\widehat{p}(X)\bar{\mathsf{z}}_k(X) - \bar{\mathsf{z}}_k(X). \end{cases}$$

(2) Example 1(2):

$$\begin{cases} \mathcal{E}_{\widehat{b},m_2}(\bar{\mathsf{z}}_k)(O) = \widehat{b}(X)\bar{\mathsf{z}}_k(X) - Y\bar{\mathsf{z}}_k(X), \\ \mathcal{E}_{\widehat{p},m_1}(\bar{\mathsf{z}}_k)(O) = \widehat{p}(X)\bar{\mathsf{z}}_k(X) - A\bar{\mathsf{z}}_k(X). \end{cases}$$

The following lemma proves that (i) $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ is an unbiased estimator of $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$ and (ii) $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$ is the expected product of the $L_2(\mathsf{P}_\theta)$ projections of the weighted residuals

$$\Delta_{1,\widehat{b}} \coloneqq \lambda(X)^{1/2}\left(\widehat{b}(X) - b(X)\right),$$
$$\Delta_{2,\widehat{p}} \coloneqq \lambda(X)^{1/2}\left(\widehat{p}(X) - p(X)\right)$$

onto the subspace spanned by the $\lambda^{1/2}$-weighted basis functions $\bar{v}_k \coloneqq \lambda^{1/2}\bar{\mathsf{z}}_k$:

**Lemma 3.1.** *Define* $\beta_{\widehat{b},k} \coloneqq \mathsf{E}_\theta\left[\mathcal{E}_{\widehat{b},m_2}(\bar{\mathsf{z}}_k)(O)\right]^\top \Sigma_k^{-1}$ *and* $\beta_{\widehat{p},k} \coloneqq \mathsf{E}_\theta\left[\mathcal{E}_{\widehat{p},m_1}(\bar{\mathsf{z}}_k)(O)\right]^\top \Sigma_k^{-1}$*. Under the conditions of Theorem 2.2, we have*

$$
\begin{aligned}
\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1) &= \mathsf{E}_\theta\left[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})\right] \\
&= \mathsf{E}_\theta\left[\Pi_\theta\left[\Delta_{1,\widehat{b}}|\bar{v}_k\right](X)\Pi_\theta\left[\Delta_{2,\widehat{p}}|\bar{v}_k\right](X)\right] = \beta_{\widehat{b},k}^\top\Sigma_k\beta_{\widehat{p},k}
\end{aligned} \tag{3.6}
$$

*and*

$$
\mathsf{TB}_k(\theta) \coloneqq \widetilde{\psi}_k(\theta) - \psi(\theta) = \mathsf{Bias}_\theta(\widehat{\psi}_1) - \mathsf{Bias}_{\theta,k}(\widehat{\psi}_1) = \mathsf{E}_\theta\left[\Pi_\theta^\perp\left[\Delta_{1,\widehat{b}}|\bar{v}_k\right](X)\Pi_\theta^\perp\left[\Delta_{2,\widehat{p}}|\bar{v}_k\right](X)\right]. \tag{3.7}
$$

*where* $\Pi_\theta^\perp[\cdot|\bar{v}_k]$ *is the projection operator onto the orthocomplement of the linear span of* $\bar{v}_k$*.*

*Furthermore, define the bias corrected estimator* $\widehat{\psi}_{2,k}(\Sigma_k^{-1}) \coloneqq \widehat{\psi}_1 - \widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$*. Then*

$$
\begin{aligned}
\mathsf{Bias}_\theta(\widehat{\psi}_{2,k}(\Sigma_k^{-1})) &\equiv \mathsf{E}_\theta\left[\widehat{\psi}_{2,k}(\Sigma_k^{-1}) - \psi(\theta)\right] = \mathsf{TB}_k(\theta), \\
\text{and thus } \mathsf{E}_\theta\left[\widehat{\psi}_{2,k}(\Sigma_k^{-1})\right] &= \psi(\theta) + \mathsf{TB}_k(\theta) = \widetilde{\psi}_k(\theta).
\end{aligned} \tag{3.8}
$$

*Proof.* First, we show the first part of equation (3.6).

$$
\begin{aligned}
\mathsf{TB}_k(\theta) &= \mathsf{E}_\theta\left[S_{bp}\left(\widetilde{b}_{k,\theta}(X) - b(X)\right)(\widetilde{p}_{k,\theta}(X) - p(X))\right] \\
&= \mathsf{E}_\theta\left[\begin{array}{c} S_{bp}\left\{\widehat{b}(X) - b(X) - \mathsf{E}_\theta\left[(S_{bp}\widehat{b}(X)\bar{\mathsf{z}}_k(X) + m_2(O,\bar{\mathsf{z}}_k))\right]^\top\Sigma_k^{-1}\bar{\mathsf{z}}_k(X)\right\} \\ \times\left\{\widehat{p}(X) - p(X) - \mathsf{E}_\theta\left[(S_{bp}\widehat{p}(X)\bar{\mathsf{z}}_k(X) + m_1(O,\bar{\mathsf{z}}_k))\right]^\top\Sigma_k^{-1}\bar{\mathsf{z}}_k(X)\right\}\end{array}\right] \\
&= \mathsf{E}_\theta\left[S_{bp}(\widehat{b}(X) - b(X))(\widehat{p}(X) - p(X))\right] \\
&\quad - \mathsf{E}_\theta\left[S_{bp}(\widehat{b}(X) - b(X))\bar{\mathsf{z}}_k(X)\right]^\top\Sigma_k^{-1}\mathsf{E}_\theta\left[S_{bp}(\widehat{p}(X) - p(X))\bar{\mathsf{z}}_k(X)\right] \\
&\quad - \mathsf{E}_\theta\left[S_{bp}(\widehat{p}(X) - p(X))\bar{\mathsf{z}}_k(X)\right]^\top\Sigma_k^{-1}\mathsf{E}_\theta\left[S_{bp}(\widehat{b}(X) - b(X))\bar{\mathsf{z}}_k(X)\right] \\
&\quad + \mathsf{E}_\theta\left[S_{bp}(\widehat{p}(X) - p(X))\bar{\mathsf{z}}_k(X)\right]^\top\Sigma_k^{-1}\mathsf{E}_\theta\left[S_{bp}(\widehat{b}(X) - b(X))\bar{\mathsf{z}}_k(X)\right] \\
&= \underbrace{\mathsf{E}_\theta\left[S_{bp}(\widehat{b}(X) - b(X))(\widehat{p}(X) - p(X))\right]}_{\equiv \mathsf{Bias}_\theta(\widehat{\psi}_1)} \\
&\quad - \mathsf{E}_\theta\left[S_{bp}(\widehat{b}(X) - b(X))\bar{\mathsf{z}}_k(X)\right]^\top\Sigma_k^{-1}\mathsf{E}_\theta\left[S_{bp}(\widehat{p}(X) - p(X))\bar{\mathsf{z}}_k(X)\right]
\end{aligned}
$$

where in the third equality we apply Theorem 2.2(2). It follows that

$$
\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1) = \mathsf{E}_\theta\left[S_{bp}(\widehat{b}(X) - b(X))\bar{\mathsf{z}}_k(X)\right]^\top\Sigma_k^{-1}\mathsf{E}_\theta\left[S_{bp}(\widehat{p}(X) - p(X))\bar{\mathsf{z}}_k(X)\right].
$$

Again by Theorem 2.2(2),

$$\mathsf{E}_\theta\left[\mathcal{E}_{\widehat{b},m_2}(\bar{\mathsf{z}}_k)(O)\right] = \mathsf{E}_\theta\left[S_{bp}\widehat{b}(X)\bar{\mathsf{z}}_k(X) + m_2(O,\bar{\mathsf{z}}_k)\right] = \mathsf{E}_\theta\left[S_{bp}\left(\widehat{b}(X) - b(X)\right)\bar{\mathsf{z}}_k(X)\right]$$
$$\mathsf{E}_\theta\left[\mathcal{E}_{\widehat{p},m_1}(\bar{\mathsf{z}}_k)(O)\right] = \mathsf{E}_\theta\left[S_{bp}\widehat{p}(X)\bar{\mathsf{z}}_k(X) + m_1(O,\bar{\mathsf{z}}_k)\right] = \mathsf{E}_\theta\left[S_{bp}\left(\widehat{p}(X) - p(X)\right)\bar{\mathsf{z}}_k(X)\right]$$

and thus

$$\mathsf{E}_\theta\left[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})\right] = \mathsf{E}_\theta\left[\mathcal{E}_{\widehat{b},m_2}(\bar{\mathsf{z}}_k)(O)\right]^\top \Sigma_k^{-1}\mathsf{E}_\theta\left[\mathcal{E}_{\widehat{p},m_1}(\bar{\mathsf{z}}_k)(O)\right] \equiv \mathsf{Bias}_{\theta,k}(\widehat{\psi}_1).$$

In terms of the second part of equation (3.6):

$$\begin{aligned}
\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1) &= \mathsf{E}_\theta\left[S_{bp}(\widehat{b}(X) - b(X))\bar{\mathsf{z}}_k(X)\right]^\top \Sigma_k^{-1}\mathsf{E}_\theta\left[S_{bp}(\widehat{p}(X) - p(X))\bar{\mathsf{z}}_k(X)\right] \\
&= \mathsf{E}_\theta\left[\lambda(X)(\widehat{b}(X) - b(X))\bar{\mathsf{z}}_k(X)\right]^\top \Sigma_k^{-1}\mathsf{E}_\theta\left[\lambda(X)(\widehat{p}(X) - p(X))\bar{\mathsf{z}}_k(X)\right] \\
&= \mathsf{E}_\theta\left[\lambda(X)^{1/2}(\widehat{b}(X) - b(X))\bar{v}_k(X)\right]^\top \Sigma_k^{-1}\mathsf{E}_\theta\left[\lambda(X)^{1/2}(\widehat{p}(X) - p(X))\bar{v}_k(X)\right] \\
&= \mathsf{E}_\theta\left[\Delta_{1,\widehat{b}}\bar{v}_k(X)\right]^\top \Sigma_k^{-1}\mathsf{E}_\theta\left[\Delta_{2,\widehat{p}}\bar{v}_k(X)\right] \\
&= \mathsf{E}_\theta\left[\Pi_\theta\left[\Delta_{1,\widehat{b}}|\bar{v}_k\right](X)\Pi_\theta\left[\Delta_{2,\widehat{p}}|\bar{v}_k\right](X)\right].
\end{aligned}$$

where in the last line we use the definition of the projection operator $\Pi_\theta\left[\cdot|\bar{v}_k\right]$ (see equation (1.2)).

Then equation (3.7) immediately follows:

$$\begin{aligned}
\mathsf{TB}_k(\theta) &= \mathsf{Bias}_\theta(\widehat{\psi}_1) - \mathsf{Bias}_{\theta,k}(\widehat{\psi}_1) \\
&= \mathsf{E}_\theta\left[\Delta_{1,\widehat{b}}\Delta_{2,\widehat{p}}\right] - \mathsf{E}_\theta\left[\Pi_\theta\left[\Delta_{1,\widehat{b}}|\bar{v}_k\right](X)^\top\Pi_\theta\left[\Delta_{2,\widehat{p}}|\bar{v}_k\right](X)\right] \\
&= \mathsf{E}_\theta\left[\Pi_\theta^\perp\left[\Delta_{1,\widehat{b}}|\bar{v}_k\right](X)\right]^\top \mathsf{E}_\theta\left[\Pi_\theta^\perp\left[\Delta_{2,\widehat{p}}|\bar{v}_k\right](X)\right].
\end{aligned}$$

Finally, we prove equation (3.8):

$$\begin{aligned}
\mathsf{Bias}_\theta(\widehat{\psi}_{2,k}(\Sigma_k^{-1})) &= \mathsf{E}_\theta\left[\widehat{\psi}_{2,k}(\Sigma_k^{-1}) - \psi(\theta)\right] \\
&= \mathsf{E}_\theta\left[\widehat{\psi}_1 - \psi(\theta) - \widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})\right] \\
&= \mathsf{Bias}_\theta(\widehat{\psi}_1) - \mathsf{Bias}_{\theta,k}(\widehat{\psi}_1) = \mathsf{TB}_k(\theta)
\end{aligned}$$

where the third line follows from the definition of $\mathsf{Bias}_\theta(\widehat{\psi}_1)$ and equation (3.6). ∎

**Comment 3.1.** *By* $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1) = \mathsf{E}_\theta\left[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})\right] = \beta_{\widehat{b},k}^\top\Sigma_k^{-1}\beta_{\widehat{p},k}$*, we can view* $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ *as an unbiased estimator of the bilinear functional* $\beta_{\widehat{b},k}^\top\Sigma_k^{-1}\beta_{\widehat{p},k}$*, where* $\beta_{\widehat{b},k}$ *and* $\beta_{\widehat{p},k}$ *are the population coefficients of the following* *linear regressions:*

$$\text{regress } \lambda^{1/2}(\widehat{b} - b) \text{ against } \lambda^{1/2}\bar{\mathsf{z}}_k,$$
$$\text{regress } \lambda^{1/2}(\widehat{p} - p) \text{ against } \lambda^{1/2}\bar{\mathsf{z}}_k.$$

**Comment 3.2.** *Under the above setup, we can relate* $\widehat{\psi}_{2,k}(\Sigma_k^{-1}) - \widetilde{\psi}_k(\theta) = \widehat{\psi}_1 - \widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1}) - \widetilde{\psi}_k(\theta)$ *to the second order influence function of the truncated parameter* $\widetilde{\psi}_k(\theta)$ *of a DR functional, following Robins et al. (2008, Definition 2.1). The proof can be found in Appendix A.1.*

- $\widehat{\psi}_{2,k}(\Sigma_k^{-1}) - \widetilde{\psi}_k(\theta)$ *is the second order influence function of* $\widetilde{\psi}_k(\theta)$ *under the law* $\mathsf{P}_{\widehat{\theta}}$*;*
- $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ *is the second order influence function of* $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$ *under the law* $\mathsf{P}_{\widehat{\theta}}$*.*

*The forms of* $m$*-th order influence functions of the DR functionals for general* $m$ *are presented in Theorem 5.2 in Section 6. The above results are simple corollary of Theorem 5.2 and Remark 5.1(2).*

We introduce the following additional notation for various norms which will be useful for stating the statistical properties of $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ and $\widehat{\psi}_{2,k}(\Sigma_k^{-1})$:

$$\begin{aligned}
\mathbb{L}_{\theta,2,\widehat{b}} &\coloneqq \left\{\mathsf{E}_\theta[\Delta_{1,\widehat{b}}^2]\right\}^{1/2}, \mathbb{L}_{\theta,2,\widehat{b},k} \coloneqq \left\{\mathsf{E}_\theta[\Pi_\theta[\Delta_{1,\widehat{b}}|\bar{v}_k](X)^2]\right\}^{1/2}, \\
\mathbb{L}_{\theta,2,\widehat{p}} &\coloneqq \left\{\mathsf{E}_\theta[\Delta_{1,\widehat{p}}^2]\right\}^{1/2}, \mathbb{L}_{\theta,2,\widehat{p},k} \coloneqq \left\{\mathsf{E}_\theta[\Pi_\theta[\Delta_{1,\widehat{p}}|\bar{v}_k](X)^2]\right\}^{1/2}, \\
\mathbb{L}_{\theta,\infty,\widehat{b},k} &\coloneqq \|\Pi_\theta[\Delta_{1,\widehat{b}}|\bar{v}_k]\|_\infty, \mathbb{L}_{\theta,\infty,\widehat{p},k} \coloneqq \|\Pi_\theta[\Delta_{1,\widehat{p}}|\bar{v}_k]\|_\infty, \\
\mathbb{L}_{\theta,2,\widehat{\Sigma},k} &\coloneqq \|\widehat{\Sigma}_k - \Sigma_k\|.
\end{aligned}$$

Henceforth, we further impose the following weak conditions (Condition SW).

**Condition SW.**

(1) *All the eigenvalues of* $\Sigma_k$ *are bounded away from 0 and* $\infty$*.*

(2) *The observed data* $O = (W, X)$*, the true nuisance functions* $b(X)$ *and* $p(X)$*, the estimated nuisance functions* $\widehat{b}(X)$ *and* $\widehat{p}(X)$ *are bounded with* $\mathsf{P}_\theta$*-probability 1, and* $\lambda(X)$ *are bounded away from 0 and* $\infty$ *with* $\mathsf{P}_\theta$*-probability 1.*

(3) $\|\bar{\mathsf{z}}_k^\top \bar{\mathsf{z}}_k\|_\infty \leqslant Bk$ *for some constant* $B > 0$*.*

Note that Condition SW should also be compared to the following slightly stronger Condition W in Liu et al. (2020a), which is the same to Condition SW except (3) shall be replaced by the following (3'):

**Condition W.**

(3') $\|\bar{\mathsf{z}}_k^\top \bar{\mathsf{z}}_k\|_\infty \leqslant Bk$ *for some constant* $B > 0$*,* $\|\Pi_\theta[\Delta_{1,\widehat{b}}|\bar{v}_k]\|_\infty < \infty$ *and* $\|\Pi_\theta[\Delta_{1,\widehat{p}}|\bar{v}_k]\|_\infty < \infty$*.*

**Comment 3.3.** *Condition W(3') is stronger than Condition SW(3) and it holds for Cohen-Vial-Daubechies wavelets, local polynomial partition and B-spline series (Belloni et al., 2015). But it does not hold in general* *for Fourier series or monomial transformation of the covariates $X$ when $X$ is compactly supported (Belloni et al., 2015). We also refer interested readers to Kennedy et al. (2020, Section 6) and Liu et al. (2020b, Section 2.1) for further motivation on why we decide to relax Condition W by Condition SW. However, as we will see, when $\Sigma_k$ is unknown and need to be estimated from the training sample, violation of Condition W(3') but not Condition SW(3), will incur a loss in the power of the test, but the validity of the test (i.e. level) is still guaranteed. Nevertheless, such loss in power happens (or equivalently, Condition SW holds but Condition W fails to hold), only if the infinity norms of the projections $\Pi_\theta[\Delta_{1,\widehat{b}}|\bar{v}_k]$ and $\Pi_\theta[\Delta_{1,\widehat{p}}|\bar{v}_k]$ are not bounded even though those of $\Delta_{1,\widehat{b}}$ and $\Delta_{1,\widehat{p}}$ are bounded.*

The statistical properties of the oracle estimator $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ and the oracle biased-corrected estimator $\widehat{\psi}_{2,k}(\Sigma_k^{-1})$ are given by the following theorem.

**Theorem 3.1.** *Under the conditions of Theorem 2.2 and Condition SW, with $k, n \to \infty$, and $k = o(n^2)$, conditional on the training sample, we have*

(1) $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ *is unbiased for* $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$ *with variance of order*

$$\frac{1}{n}\left(\frac{k}{n} + \mathbb{L}^2_{\theta,2,\widehat{b},k} + \mathbb{L}^2_{\theta,2,\widehat{p},k}\right).$$

(2) $\frac{\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1}) - \mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]} \xrightarrow{d} N(0,1)$, *where* $\xrightarrow{d}$ *stands for convergence in distribution. Further,* $\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})] := \mathsf{var}_\theta^{1/2}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]$ *can be estimated by the bootstrap estimator* $\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})] := \widehat{\mathsf{var}}^{1/2}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]$ *defined in Appendix A.3 .*

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

$$\frac{1}{\sqrt{n}}\left(\sqrt{\frac{k}{n}} + \mathbb{L}_{\theta,2,\widehat{b},k} + \mathbb{L}_{\theta,2,\widehat{p},k}\right).$$

*Proof.* The variance order of $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ is proved in Appendix A.2. When $k = o(n^2)$ and $k \to \infty$ as $n \to \infty$, the conditional asymptotic normality of $\frac{\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1}) - \mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]}$ follows directly from Hoeffding decomposition, with the conditional asymptotic normality of the degenerate second-order U-statistic part implied by Bhattacharya and Ghosh (1992, Corollary 1.2). Appendix A.3 proves that the bootstrap standard error estimators $\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]$ satisfy $\frac{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]} = 1 + o_{\mathsf{P}_\theta}(1)$. ∎

**Comment 3.4.** *When $k \gtrsim n^2$, the Gaussian limit of $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ does not hold. Further if $k \gg n^2$, $\mathsf{var}_\theta[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})] = O(\frac{k}{n^2})$ is of order greater than 1, and therefore $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ cannot consistently estimate $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$ regardless of its magnitude.*

3.2. **Which null hypothesis should be tested?** We have touched upon this problem in a previous work (Liu et al., 2020a). We nonetheless revisit the issue here for the sake of completeness. Recall that in the introduction, we pose the following question:

*Given the training sample used to estimate the nuisance functions $b$ and $p$, can we use empirical information to test the null hypothesis $\mathsf{NH}_0 : \mathsf{Bias}_\theta(\widehat{\psi}_1) = o(n^{-1/2})$, the violation of which implies under coverage of $\mathsf{CI}_\alpha(\widehat{\psi}_1)$?*

In fact, we shall have to settle for statements that are "in dialogue" with current practices and literature. In the current literature, Theorem 2.3 provides the theoretical guarantees on the asymptotic coverage validity of a nominal $(1-\alpha)$ two-sided Wald confidence interval $\mathsf{CI}_\alpha(\widehat{\psi}_1)$ based on a DML estimator $\widehat{\psi}_1$, under the "hypothesis" that $\mathsf{NH}_0 : \mathsf{Bias}_\theta(\widehat{\psi}_1) = o(n^{-1/2})$ i.e. the bias of the DML estimator $\widehat{\psi}_1$ for $\psi(\theta)$ is dominated by the standard error of $\widehat{\psi}_1$ of order $n^{-1/2}$. This "hypothesis" is often proved under complexity-reducing assumptions on the nuisance parameters $b$ and $p$, such as smoothness or sparsity - assumptions which may be incorrect.

However $\mathsf{NH}_0$ involves an asymptotic statement which cannot be quantified or operationalized by a data analyst in finite samples. We propose the following strategy for operationalizing an asymptotic null hypothesis: whenever a null hypothesis (e.g., $\mathsf{NH}_0$) is defined in terms of an asymptotic rate of convergence (e.g. $o(n^{-1/2})$) in the training sample data, we will (1) ask the authors to specify a positive number $\delta = \delta(N)$ possibly depending on the actual sample size $N$ and (2) then operationalize the asymptotic null hypothesis as non-asymptotic null hypothesis (e.g $\mathsf{H}_0(\delta)$ below) quantified by $\delta$. In particular for $\mathsf{NH}_0$, we propose the following operationalized pair

$$\mathsf{NH}_0 \quad : \quad \mathsf{Bias}_\theta(\widehat{\psi}_1) = o(n^{-1/2}); \tag{3.9}$$

$$\mathsf{H}_0(\delta) \quad : \quad \frac{\left|\mathsf{Bias}_\theta(\widehat{\psi}_1)\right|}{\mathsf{s.e.}_\theta(\widehat{\psi}_1)} < \delta. \tag{3.10}$$

operationalized as follows: if $\mathsf{H}_0(\delta)$ is rejected (accepted), we, by convention, will declare $\mathsf{NH}_0$ rejected (accepted). The authors' choice of $\delta$ depends on the degree of under coverage they are willing to tolerate. For example, as we mentioned in Section 1, if 80% is the minimum actual coverage they would tolerate for a 90% two-sided Wald confidence interval $\mathsf{CI}_{\alpha=0.05}(\widehat{\psi}_1)$ (2.4), then under normality they would select $\delta = 0.75$.

Similarly, we have the *surrogate* operationalized pair

$$\mathsf{NH}_{0,k} \quad : \quad \mathsf{Bias}_{\theta,k}(\widehat{\psi}_1) = o(n^{-1/2}); \tag{3.11}$$

$$\mathsf{H}_{0,k}(\delta) \quad : \quad \frac{\left|\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)\right|}{\mathsf{s.e.}_\theta(\widehat{\psi}_1)} < \delta. \tag{3.12}$$

Note that $\mathsf{H}_{0,k}(\delta)$ specifies $|\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)|$, *not* $|\mathsf{Bias}_\theta(\widehat{\psi}_1)|$, to be less than $\delta$ fraction of $\mathsf{s.e.}_\theta(\widehat{\psi}_1)$. Given the statistical properties of $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ summarized in Theorem 3.1, $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$ can be unbiasedly estimated by $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$, whose standard error shrinks to zero at rate at most $1/\sqrt{n}$. This observation motivates the following oracle test of $\mathsf{H}_{0,k}(\delta)$ (3.12):

$$\widehat{\chi}_{2,k}(\Sigma_k^{-1}; \varsigma_k, \delta) \coloneqq \mathbb{1}\left\{ \frac{|\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})|}{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]} - \varsigma_k \frac{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]}{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]} > \delta \right\}, \tag{3.13}$$

for user-specified $\varsigma_k$ and $\delta > 0$. In fact, later we will show in Section 3.3 that $\widehat{\chi}_{2,k}(\Sigma_k^{-1}; \varsigma_k, \delta)$ is a consistent test of $\mathsf{H}_{0,k}(\delta)$ (3.12). For now, we take this conclusion as given and discuss the implication of the test result on the original null hypothesis of interest $\mathsf{NH}_0$ (3.9) in Sections 3.2.1 and 3.2.2 .

3.2.1. *Faithfulness assumption.* Suppose now a data analyst agrees that in reporting $\mathsf{CI}_\alpha(\widehat{\psi}_1)$ (2.4) as a $(1-\alpha)$ two sided confidence interval for $\psi(\theta)$, their implicit or explicit null hypothesis is $\mathsf{NH}_0$ (3.9), that is, the bias $\mathsf{Bias}_\theta(\widehat{\psi}_1)$ of their estimator is $o(n^{-1/2})$. Further suppose the test $\widehat{\chi}_k^{(2)}(\Sigma_k^{-1}; z_{\alpha^\dagger/2}, \delta)$ rejects $\mathsf{H}_{0,k}(\delta)$ (3.12), equivalently $\mathsf{NH}_{0,k}$ (3.11). However, rejecting $\mathsf{NH}_{0,k}$ (3.11) does *not* logically imply rejecting $\mathsf{NH}_0$ (3.9) in general. One "solution" is to adopt an additional "faithfulness" assumption under which rejection of $\mathsf{NH}_{0,k}$ does logically imply rejection of $\mathsf{NH}_0$.

**Condition Faithfulness.** *Given a fixed $k$, $\frac{\mathsf{Bias}_\theta(\widehat{\psi}_1)}{\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)} = 1 + \frac{\mathsf{TB}_k(\theta)}{\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)}$ is not $o(1)$.*

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

**Condition Faithfulness($\delta'$).** *For a given $k$, $\left|\frac{\mathsf{Bias}_\theta(\widehat{\psi}_1)}{\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)}\right| = \left|1 + \frac{\mathsf{TB}_k(\theta)}{\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)}\right| \geqslant \delta'$.*

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

3.2.2. *An alternative to faithfulness?* In this section, we discuss an alternative "solution" that avoids imposing Condition Faithfulness or Condition Faithfulness($\delta'$). The "solution" described in Section 1 is to disprove the complexity reducing assumptions on $b$ and $p$ used by an analyst to justify the validity of her reported Wald interval for $\psi(\theta)$. We restrict consideration to those cases in which an analyst justifies the validity of her Wald $(1-\alpha)$ confidence interval $\mathsf{CI}_\alpha(\widehat{\psi}_1) = \widehat{\psi}_1 \pm z_{\alpha/2}\widehat{\mathsf{s.e.}}(\widehat{\psi}_1)$ for $\psi(\theta)$ by first proving that, under her complexity reducing assumptions, the Cauchy Schwarz (CS) bias functional

$$\mathsf{CSBias}_\theta(\widehat{\psi}_1) \coloneqq \{\mathsf{E}_\theta[\lambda(X)\{b(X) - \widehat{b}(X)\}^2]\mathsf{E}_\theta[\lambda(X)\{p(X) - \widehat{p}(X)\}^2]\}^{1/2} \equiv \mathbb{L}_{\theta,2,\widehat{b}}\mathbb{L}_{\theta,2,\widehat{p}}. \tag{3.14}$$

is $o(n^{-1/2})$ and then noting that this implies $|\mathsf{Bias}_\theta(\widehat{\psi}_1)| = \left| \mathsf{E}_\theta \left[ \lambda(X) \left( \widehat{b}(X) - b(X) \right) (\widehat{p}(X) - p(X)) \right] \right|$ is $o(n^{-1/2})$ by the Cauchy-Schwarz inequality. In this section we provide a test that has power to reject (the operationalized version of) the hypothesis $\mathsf{CSBias}_\theta(\widehat{\psi}_1) = o(n^{-1/2})$. If this test rejects, we have also rejected the analyst's complexity reducing assumptions and therefore her justification for her claim that her Wald interval covers at a rate at least the nominal. We begin by noting that any analysis who follows the squared error loss minimization strategy defined below must indeed be using the criteria $\mathsf{CSBias}_\theta(\widehat{\psi}_1) = o(n^{-1/2})$ to justify the validity of her Wald interval.

> **Squared Error Loss Minimization Strategy**: This strategy is any strategy in which $\widehat{b}(x)$ and $\widehat{p}(x)$ are near minimizers of the conditional expected $\lambda$-weighted ($\mathsf{E}_\theta[\lambda(X)\{\widehat{b}(X) - b(X)\}^2]$ and $\mathsf{E}_\theta[\lambda(X)\{\widehat{p}(X) - p(X)\}^2]$) squared error losses [or unweighted squared error losses $\mathsf{E}_\theta[\{\widehat{b}(X) - b(X)\}^2]$ and $\mathsf{E}_\theta[\{\widehat{p}(X) - p(X)\}^2]$, as discussed later] over the set of functions computable by the algorithm.

Because $b(X)$ and/or $p(X)$ are not necessarily conditional expectations, we will show that it may be more natural to minimize the expected $\lambda$-weighted squared error loss than the expected unweighted squared error loss. To see why, recall that for DR functionals, Theorem 2.2(5) says that the nuisance functions $b$ and $p$ are the solutions to the following minimization problems:

$$\begin{cases} b = \mathsf{arg\ min}_{h \in L_2(\mathsf{P}_{\theta,X})} \mathsf{E}_\theta \left[ S_{bp} \frac{h^2}{2} + m_2(O, h) \right], \\ p = \mathsf{arg\ min}_{h \in L_2(\mathsf{P}_{\theta,X})} \mathsf{E}_\theta \left[ S_{bp} \frac{h^2}{2} + m_1(O, h) \right]. \end{cases}$$

As a result, to construct $\widehat{b}$ and $\widehat{p}$, it is natural to use an ML algorithm to try to obtain the global minimizers:

$$\widehat{b} = \underset{h \in \mathcal{F}}{\mathsf{arg\ min}} \mathsf{P}_{n_{\mathsf{tr}}} \left[ S_{bp} \frac{h^2}{2} + m_2(O, h) \right], \tag{3.15}$$

$$\widehat{p} = \underset{h \in \mathcal{F}}{\mathsf{arg\ min}}\ \mathsf{P}_{n_{\mathsf{tr}}} \left[ S_{bp} \frac{h^2}{2} + m_1(O, h) \right]. \tag{3.16}$$

where $\mathcal{F}$ is the set of functions computable by the ML algorithm. However, we now show that equation (2.2) in Theorem 2.1 implies that the loss minimization criterion in equations (3.15) and (3.16) are equivalent to minimizing expected weighted squared error loss with weight $\lambda(\cdot)$. Thus researchers who use the ML algorithm with the above objective function are explicitly acknowledging that they have adopted the **Squared Error Loss Minimization Strategy**. The following algebra makes this statement explicit:

$$\begin{aligned} \frac{1}{2} \mathsf{E}_\theta[\lambda(X)(b(X) - h(X))^2] &= \frac{1}{2} \mathsf{E}_\theta[S_{bp}(b(X) - h(X))^2] \\ &= \mathsf{E}_\theta[S_{bp} \frac{h(X)^2}{2} - S_{bp} b(X) h(X)] + \frac{1}{2} \mathsf{E}_\theta[S_{bp} b(X)^2] \\ &= \mathsf{E}_\theta[S_{bp} \frac{h(X)^2}{2} + m_2(O, h)] + \frac{1}{2} \mathsf{E}_\theta[S_{bp} b(X)^2] \end{aligned}$$

where in the third line we use equation (2.2) and the extra term $\frac{1}{2}\mathsf{E}_\theta[S_{bp}b(X)^2]$ is unrelated to minimizing $\frac{1}{2}\mathsf{E}_\theta[\lambda(X)(b(X)-h(X))^2]$. By symmetry,

$$\frac{1}{2}\mathsf{E}_\theta[\lambda(X)(p(X)-h(X))^2] = \mathsf{E}_\theta[S_{bp}\frac{h(X)^2}{2} + m_1(O,h)] + \frac{1}{2}\mathsf{E}_\theta[S_{bp}p(X)^2].$$

The following Lemma shows $\mathsf{CSBias}_\theta(\widehat{\psi}_1) = o(n^{-1/2})$ if and only if

$$\mathsf{CSBias}_\theta^{uw}(\widehat{\psi}_1) \coloneqq \left\{\mathsf{E}_\theta[(b(X)-\widehat{b}(X))^2]\mathsf{E}_\theta[(p(X)-\widehat{p}(X))^2]\right\}^{1/2} = o(n^{-1/2})$$

so rejection of either of these two hypotheses implies rejection of the other.

**Lemma 3.2.**

$$\mathsf{CSBias}_\theta(\widehat{\psi}_1) \lesssim \mathsf{CSBias}_\theta^{uw}(\widehat{\psi}_1) \lesssim \mathsf{CSBias}_\theta(\widehat{\psi}_1).$$

*That is, there exists* $1 > c > 0$ *such that* $c > \mathsf{CSBias}_\theta(\widehat{\psi}_1)/\mathsf{CSBias}_\theta^{uw}(\widehat{\psi}_1) < 1/c$.

*Proof.* By Condition SW, we have

$$\begin{aligned}
&\mathsf{E}_\theta[(b(X)-\widehat{b}(X))^2] = \mathsf{E}_\theta\left[\frac{1}{\lambda(X)}\lambda(X)(b(X)-\widehat{b}(X))^2\right] \\
\Rightarrow &\inf_x |\lambda^{-1}(x)|\mathsf{E}_\theta[\lambda(X)(b(X)-\widehat{b}(X))^2] \leqslant \mathsf{E}_\theta[(b(X)-\widehat{b}(X))^2] \leqslant \sup_x |\lambda^{-1}(x)|\mathsf{E}_\theta[\lambda(X)(b(X)-\widehat{b}(X))^2]
\end{aligned}$$

and similarly

$$\inf_x |\lambda^{-1}(x)|\mathsf{E}_\theta[\lambda(X)(p(X)-\widehat{p}(X))^2] \leqslant \mathsf{E}_\theta[(p(X)-\widehat{p}(X))^2] \leqslant \sup_x |\lambda^{-1}(x)|\mathsf{E}_\theta[\lambda(X)(p(X)-\widehat{p}(X))^2].$$

Thus combining the above calculations, we have

$$\mathsf{CSBias}_\theta(\widehat{\psi}_1) \lesssim \mathsf{CSBias}_\theta^{uw}(\widehat{\psi}_1) \lesssim \mathsf{CSBias}_\theta(\widehat{\psi}_1).$$

∎

Thus the $\lambda$-weighted and unweighted squared error loss functions are equivalent up to constants by Condition SW. Henceforth we restrict to consideration of $\mathsf{CSBias}_\theta(\widehat{\psi}_1)$.

Consider the operationalized pair of null hypotheses.

$$\mathsf{NH}_{0,\mathsf{CS}} \quad : \quad \mathsf{CSBias}_\theta(\widehat{\psi}_1) = o(n^{-1/2}) \tag{3.17}$$

$$\mathsf{H}_{0,\mathsf{CS}}(\delta) \quad : \quad \frac{\mathsf{CSBias}_\theta(\widehat{\psi}_1)}{\mathsf{s.e.}_\theta(\widehat{\psi}_1)} < \delta. \tag{3.18}$$

We now show that we will are able to construct a test that has power to reject $\mathsf{H}_{0,\mathsf{CS}}(\delta)$ under certain alternatives based on the following Lemma.

**Lemma 3.3.**

(1) $\mathsf{NH}_{0,\mathsf{CS}} \Rightarrow \mathsf{NH}_0$, *and similarly* $\mathsf{H}_{0,\mathsf{CS}}(\delta) \Rightarrow \mathsf{H}_0(\delta)$*;*

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

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

$$
\begin{aligned}
|\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)| &= \left|\mathsf{E}_\theta\left[\Pi_\theta[\lambda^{1/2}(b-\widehat{b})|\bar{\mathsf{z}}_k](X)\Pi_\theta[\lambda^{1/2}(p-\widehat{p})|\bar{\mathsf{z}}_k](X)\right]\right| \\
&\leqslant \mathbb{L}_{\theta,2,\widehat{b},k}\mathbb{L}_{\theta,2,\widehat{p},k} \leqslant \mathbb{L}_{\theta,2,\widehat{b}}\mathbb{L}_{\theta,2,\widehat{p}}
\end{aligned} \tag{3.19}
$$

where the first inequality follows from CS inequality and the second inequality is a consequence of the fact that a projection contracts $L_2(\mathsf{P}_\theta)$ norms. ∎

Hence falsification of $\mathsf{H}_{0,k}(\delta)$ and $\mathsf{NH}_{0,k}$ implies falsification of $\mathsf{NH}_{0,\mathsf{CS}}$ and $\mathsf{H}_{0,\mathsf{CS}}(\delta)$, respectively. However, note the converse statements of Lemma 3.3 are not always true: for example, $\mathsf{NH}_0$ may be true (and thus, by Theorem 2.3 the above the Wald confidence interval centered at $\widehat{\psi}_1$ is valid) even when the CS null hypothesis is false. Even so, suppose we empirically falsify the *justification* $\mathsf{NH}_{0,\mathsf{CS}}$ ($\mathsf{H}_{0,\mathsf{CS}}(\delta)$) for the null hypothesis of actual interest $\mathsf{NH}_0$ ($\mathsf{H}_0(\delta)$). Then, although logically $\mathsf{NH}_0$ may be true, there seems, to us, neither a substantive nor a philosophical reason to assume $\mathsf{NH}_0$ is true in the absence of $\mathsf{NH}_{0,\mathsf{CS}}$. Hence we take the following philosophical stance:

**Condition CS.** *If the CS null hypothesis* $\mathsf{NH}_{0,\mathsf{CS}}$ *and* $\mathsf{H}_{0,\mathsf{CS}}(\delta)$ *being true is used as the justification for the validity of the Wald* $(1-\alpha)$ *confidence interval* $\mathsf{CI}_\alpha(\widehat{\psi}_1)$*, but in fact are false, one should refuse to support claims whose validity rests on the truth of* $\mathsf{NH}_0$ *or* $\mathsf{H}_0(\delta)$*; in particular, the claims that the Wald confidence intervals centered at* $\widehat{\psi}_1$ *have true coverage greater than or equal to their nominal.*

Since it follows from Lemma 3.3(2) that the rejection of the surrogate $\mathsf{H}_{0,k}(\delta)$ implies rejection of $\mathsf{H}_{0,\mathsf{CS}}(\delta)$, we are only to show that the above test $\widehat{\chi}_{2,k}(\Sigma_k^{-1};\varsigma_k,\delta)$ of the surrogate null hypothesis $\mathsf{H}_{0,k}(\delta)$ is a consistent test of surrogate hypothesis with good power properties.

3.3. **Statistical properties of $\widehat{\chi}_{2,k}(\Sigma_k^{-1};\varsigma_k,\delta)$.** As mentioned above, based on the statistical properties of $\widehat{\psi}_1$ and $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ summarized in Theorems 2.3 and 3.1, for a DR functional $\psi(\theta)$, we study the statistical property of the oracle two-sided test $\widehat{\chi}_{2,k}(\Sigma_k^{-1};\varsigma_k,\delta)$ as a test of the surrogate null hypothesis $\mathsf{H}_{0,k}(\delta)$.

The following theorem characterizes the asymptotic level and power of $\widehat{\chi}_{2,k}(\Sigma_k^{-1};\varsigma_k,\delta)$ for $\mathsf{H}_{0,k}(\delta)$ (3.12).

**Theorem 3.2.** *For a DR functional* $\psi(\theta)$*, under the conditions in Theorem 2.2, Theorem 3.1 and Condition SW, when* $k\to\infty$ *but* $k=o(n)$*, for any given* $\varsigma_k,\delta>0$*, suppose that* $\frac{|\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)|}{\mathsf{s.e.}_\theta[\widehat{\psi}_1]}=\gamma$ *for some (sequence)* $\gamma=\gamma(n)$ *(where* $\gamma(n)$ *can diverge with* $n$*), then the rejection probability of* $\widehat{\chi}_{2,k}(\Sigma_k^{-1};\varsigma_k,\delta)$ *converges to*

$$
2-\Phi\left(\varsigma_k-\lim_{n\to\infty}(\gamma-\delta)\frac{\mathsf{s.e.}_\theta[\widehat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]}\right)-\Phi\left(\varsigma_k+\lim_{n\to\infty}(\gamma+\delta)\frac{\mathsf{s.e.}_\theta[\widehat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]}\right) \tag{3.20}
$$

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

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

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

(1′) *If* $\mathbb{L}_{\theta,2,\widehat{b},k}$ *and* $\mathbb{L}_{\theta,2,\widehat{p},k}$ *converge to 0, under the following alternative to* $\mathsf{H}_{0,k}(\delta)$: $\gamma = \delta + c$, *for any fixed* $c > 0$ *or any diverging sequence* $c = c(n) \to \infty$, $\widehat{\chi}_{2,k}(\Sigma_k^{-1}; \varsigma_k, \delta)$ *has rejection probability converging to 1, as* $n \to \infty$.

**Comment 3.5.** *In Appendix A.4, we prove equation (3.20). We now prove that equation (3.20) implies Theorem 3.2(1)-(2) and (2′).*

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

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

  *which implies that the rejection probability is less than or equal to* $2 - 2\Phi(\varsigma_k)$. *Choose* $\varsigma_k = z_{\alpha^\dagger/2}$, $2(1 - \Phi(\varsigma_k)) = 2\alpha^\dagger/2 = \alpha^\dagger$ *and conclude that the test is a valid level* $\alpha^\dagger$ *test of the null.*
- *Under the alternative to* $\mathsf{H}_{0,k}(\delta)$ *with* $\gamma = \delta + c$ *for some* $c > 0$, *it follows from Theorem 3.1 and equation (3.20) that the rejection probability of* $\widehat{\chi}_{2,k}(\Sigma_k^{-1}; \varsigma_k, \delta)$, *as* $n \to \infty$, *is no smaller than*

$$2 - \Phi\left(\varsigma_k - c\Theta(b, p, \widehat{b}, \widehat{p}, f_X, \bar{\mathsf{z}}_k)\left\{\frac{k}{n} + \mathbb{L}_{\theta,2,\widehat{b},k} + \mathbb{L}_{\theta,2,\widehat{p},k}\right\}^{-1}\right) - \Phi(\infty)$$

  *where* $\Theta(b, p, \widehat{b}, \widehat{p}, f_X, \bar{\mathsf{Z}}_k)$ *is some positive constant depending on the true regression functions* $b$ *and* $p$, *the estimated functions* $\widehat{b}, \widehat{p}$ *from the training sample, the density* $f_X$ *of* $X$ *and the chosen basis functions* $\bar{\mathsf{z}}_k$. *To have power approaching 1 to reject* $\mathsf{H}_{0,k}(\delta)$, *we need one of the following:*
  - *If one of* $\mathbb{L}_{\theta,2,\widehat{b},k}$ *and* $\mathbb{L}_{\theta,2,\widehat{p},k}$ *is* $O(1)$, *we need* $c \to \infty$ *to guarantee that the rejection probability of* $\widehat{\chi}_{2,k}(\Sigma_k^{-1}; \varsigma_k, \delta)$ *converges to* $1 - \Phi(-\infty) = 1$. *Hence we have Theorem 3.2(2).*
  - *If* $c$ *is fixed, we need both* $\mathbb{L}_{\theta,2,\widehat{b},k}$ *and* $\mathbb{L}_{\theta,2,\widehat{p},k}$ *to be* $o(1)$ *to guarantee that the rejection probability of* $\widehat{\chi}_{2,k}(\Sigma_k^{-1}; \varsigma_k, \delta)$ *converges to* $1 - \Phi(-\infty) = 1$. *Hence we have Theorem 3.2(2′).*

Theorem 3.2 implies that $\widehat{\chi}_{2,k}(\Sigma_k^{-1}; z_{\alpha^\dagger/2}, \delta)$ is an asymptotically valid level $\alpha^\dagger$ two-sided test of the surrogate null $\mathsf{H}_{0,k}(\delta)$, and hence by Lemma 3.3(2) it is also an asymptotically $\alpha^\dagger$ level test of $\mathsf{H}_{0,\mathsf{CS}}(\delta)$. Thus when $\widehat{\chi}_{2,k}(\Sigma_k^{-1}; z_{\alpha^\dagger/2}, \delta)$ rejects $\mathsf{H}_{0,k}(\delta)$, it also rejects $\mathsf{H}_{0,\mathsf{CS}}(\delta)$ and by adopting Condition CS, we conclude that we have no justification for the coverage validity of $\mathsf{CI}_\alpha(\widehat{\psi}_1)$. It should be noted that $\widehat{\chi}_{2,k}(\Sigma_k^{-1}; z_{\alpha^\dagger/2}, \delta)$ can be a powerless test for $\mathsf{H}_{0,\mathsf{CS}}(\delta)$ under certain laws $\mathsf{P}_\theta$.

If one is willing to assume the faithfulness Condition Faithfulness($\delta'$), then $\widehat{\chi}_{2,k}(\Sigma_k^{-1}; z_{\alpha^\dagger/2}, \delta)$ is an asymptotically valid level $\alpha^\dagger$ test of $\mathsf{H}_0(\delta\delta')$. As above, it can be a powerless test of $\mathsf{H}_0(\delta\delta')$.

## 4. Assumption-lean test when $\Sigma_k^{-1}$ is unknown

Heretofore we have assumed that $\Sigma_k^{-1}$ is known. In this section, we will focus on the impact of estimating $\Sigma_k^{-1}$ on the test $\widehat{\chi}_{2,k}(\Sigma_k^{-1}; \varsigma_k, \delta)$ for testing the null hypothesis $\mathsf{H}_{0,k}(\delta)$.

One approach to proceed is to construct an estimator of the density of $X$ (Liu et al., 2016; Robins et al., 2008, 2009a, 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 $\Sigma_k^{-1}$ by $\widehat{\Sigma}_k^{-1}$ where $\widehat{\Sigma}_k = n^{-1} \sum_{i \in \mathsf{tr}} S_{bp,i} \bar{\mathsf{z}}_k(X_i) \bar{\mathsf{z}}_k(X_i)^\top$ is the sample Gram matrix estimator from the training sample. They show that, unlike the oracle $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$, $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$ is a biased estimator of $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$ with bias $\mathsf{EB}_{\theta,2,k}(\widehat{\Sigma}_k^{-1}) \equiv \mathsf{E}_\theta[\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1}) - \widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]$[6] of order

$$O\left(\mathbb{L}_{\theta,2,\widehat{b},k} \mathbb{L}_{\theta,2,\widehat{p},k} \mathbb{L}_{\theta,2,\widehat{\Sigma},k}\right) \lesssim O\left(\mathbb{L}_{\theta,2,\widehat{b},k} \mathbb{L}_{\theta,2,\widehat{p},k} \sqrt{\frac{k \mathsf{log}(k)}{n}}\right)$$

under Condition SW or W. [Note that $\mathsf{EB}_{\theta,2,k}(\widehat{\Sigma}_k^{-1})$ is also the bias of $\widehat{\psi}_{2,k}(\widehat{\Sigma}_k^{-1})$ as an estimator of the $\widetilde{\psi}_k(\theta)$.] It follows that $\mathsf{EB}_{\theta,2,k}(\widehat{\Sigma}_k^{-1}) \to 0$ as $n \to 0$ if (i) $k \mathsf{log}(k) = o(n)$ and (ii) $\mathbb{L}_{\theta,2,\widehat{b},k}$ and $\mathbb{L}_{\theta,2,\widehat{p},k}$ are bounded. However, as we will show in this section, we need to add an extra bias correction term to ensure that the test $\widehat{\chi}_{2,k}(\Sigma_k^{-1}; \varsigma_k, \delta)$ is still valid under $\mathsf{H}_{0,k}(\delta)$ (3.12) when $\Sigma_k^{-1}$ is estimated by $\widehat{\Sigma}_k^{-1}$. We define

$$\widehat{\mathbb{IF}}_{33,k}(\widehat{\Sigma}_k^{-1}) \coloneqq \frac{1}{n(n-1)(n-2)} \sum_{1 \leqslant i_1 \neq i_2 \neq i_3 \leqslant n} \widehat{\mathsf{IF}}_{33,k,\bar{i}_3}(\widehat{\Sigma}_k^{-1}) \tag{4.1}$$

where

$$\widehat{\mathbb{IF}}_{33,k,\bar{i}_3}(\widehat{\Sigma}_k^{-1}) \coloneqq \left[\mathcal{E}_{\widehat{b},m_2}(\bar{\mathsf{z}}_k)(O)\right]_{i_1}^\top \widehat{\Sigma}_k^{-1} \left\{\left[S_{bp} \bar{\mathsf{z}}_k(X) \bar{\mathsf{z}}_k(X)^\top\right]_{i_3} - \widehat{\Sigma}_k\right\} \widehat{\Sigma}_k^{-1} \left[\mathcal{E}_{\widehat{p},m_1}(\bar{\mathsf{z}}_k)(O)\right]_{i_2}. \tag{4.2}$$

**Comment 4.1.** *Analogous to Remark 3.2, we relate $\widehat{\psi}_{3,k}(\widehat{\Sigma}_k^{-1}) - \psi(\theta) \equiv \widehat{\psi}_1 - \widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1}) - \psi(\theta)$ to the third order influence function of the truncated parameter $\widetilde{\psi}_k(\theta)$ of a DR functional, following Robins et al. (2008, Definition 2.1).*

- *$\widehat{\psi}_{3,k}(\Sigma_k^{-1}) - \widetilde{\psi}_k(\theta)$ is the third order influence function of $\widetilde{\psi}_k(\theta)$ under the law $\mathsf{P}_{\widehat{\theta}}$;*
- *$\widehat{\mathbb{IF}}_{22\to 33,k}(\Sigma_k^{-1})$ is the third order influence function of $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$ under the law $\mathsf{P}_{\widehat{\theta}}$.*

*We again emphasize that the forms of the HOIFs of the DR functionals are presented in Theorem 5.2 of Section 6.*

Note that neither $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ nor $\widehat{\mathbb{IF}}_{22\to 33,k}(\Sigma_k^{-1})$ are feasible because of the dependence on $\Sigma_k^{-1}$. Therefore to construct estimators/tests based on HOIFs, we need their estimated version and in this paper we simply replace the only unknown quantity $\Sigma_k^{-1}$ by $\widehat{\Sigma}_k^{-1}$.

[6]We denote the bias due to estimating $\Sigma_k^{-1}$ by $\mathsf{EB}_{\theta,2,k}(\widehat{\Sigma}_k^{-1})$ because this bias is termed as Estimation Bias in Robins et al. (2008) and Mukherjee et al. (2017)

Then we have the following results on the variances and the estimation biases of $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$ and $\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})$ as an estimator of $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1) \equiv \mathsf{E}_\theta[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]$: the estimation bias bounds are proved in Mukherjee et al. (2017) and the variance bounds are proved in Appendix A.2:

**Proposition 4.1.** *Under the conditions of Theorem 3.2, when $k = o(n)$, the followings hold on the event that $\widehat{\Sigma}_k$ is invertible:*

(1) *The estimation bias and variance of $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$ are of the following order:*

$$\left|\mathsf{EB}_{\theta,2,k}(\widehat{\Sigma}_k^{-1})\right| \lesssim \mathbb{L}_{\theta,2,\widehat{b},k}\mathbb{L}_{\theta,2,\widehat{p},k}\mathbb{L}_{\theta,2,\widehat{\Sigma},k} \lesssim \mathbb{L}_{\theta,2,\widehat{b},k}\mathbb{L}_{\theta,2,\widehat{p},k}\sqrt{\frac{k\mathsf{log}(k)}{n}},$$

$$\mathsf{var}_\theta\left[\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})\right] \lesssim \frac{1}{n}\left\{\frac{k}{n} + \mathbb{L}^2_{\theta,2,\widehat{b},k} + \mathbb{L}^2_{\theta,2,\widehat{p},k}\right\}.$$

(2) *The estimation bias and variance of $\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1}) \coloneqq \widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1}) + \widehat{\mathbb{IF}}_{33,k}(\widehat{\Sigma}_k^{-1})$ are of the following order:*

$$\left|\mathsf{EB}_{\theta,3,k}(\widehat{\Sigma}_k^{-1})\right| \lesssim \mathbb{L}_{\theta,2,\widehat{b},k}\mathbb{L}_{\theta,2,\widehat{p},k}\mathbb{L}^2_{\theta,2,\widehat{\Sigma},k} \lesssim \mathbb{L}_{\theta,2,\widehat{b},k}\mathbb{L}_{\theta,2,\widehat{p},k}\frac{k\mathsf{log}(k)}{n},$$

$$\mathsf{var}_\theta\left[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})\right] \lesssim \frac{1}{n}\left\{\frac{k}{n} + \mathbb{L}^2_{\theta,2,\widehat{b},k} + \mathbb{L}^2_{\theta,2,\widehat{p},k} + k\mathbb{L}^2_{\theta,2,\widehat{b},k}\mathbb{L}^2_{\theta,2,\widehat{p},k}\right\}$$

*where*

$$\mathsf{EB}_{\theta,3,k}(\widehat{\Sigma}_k^{-1}) \coloneqq \mathsf{E}_\theta[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})] - \mathsf{Bias}_{\theta,k}(\widehat{\psi}_1).$$

*If, however, Condition SW is replaced by Condition W*

$$\mathsf{var}_\theta\left[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})\right] \lesssim \frac{1}{n}\left\{\frac{k}{n} + \mathbb{L}^2_{\theta,2,\widehat{b},k} + \mathbb{L}^2_{\theta,2,\widehat{p},k}\right\}.$$

(3) *As $k \to \infty$, $n \to \infty$, but $k = o(n^2)$,*

$$\frac{\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1}) - \mathsf{E}_\theta\left[\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})\right]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})]} \text{ and } \frac{\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1}) - \mathsf{E}_\theta\left[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})\right]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})]} \xrightarrow{d} N(0,1).$$

*When we estimate $\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$ and $\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})$ respectively by their bootstrap estimators $\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$ and $\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})$ described in detail in Appendix A.3, we also have*

$$\frac{\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1}) - \mathsf{E}_\theta\left[\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})\right]}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})]} \text{ and } \frac{\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1}) - \mathsf{E}_\theta\left[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})\right]}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})]} \xrightarrow{d} N(0,1).$$

We now define the following candidate test statistics when $\Sigma_k^{-1}$ is replaced by $\widehat{\Sigma}_k^{-1}$:

$$\widehat{\chi}_{2,k}(\widehat{\Sigma}_k^{-1}; \varsigma_k, \delta) \coloneqq \mathbb{1}\left\{\frac{|\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})|}{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]} - \varsigma_k \frac{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})]}{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]} > \delta\right\}, \tag{4.3}$$

$$\widehat{\chi}_{3,k}(\widehat{\Sigma}_k^{-1}; \varsigma_k, \delta) \coloneqq \mathbb{1}\left\{\frac{|\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})|}{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]} - \varsigma_k \frac{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})]}{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]} > \delta\right\}, \tag{4.4}$$

As a consequence of the asymptotic statistical properties of $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$ and $\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})$ in Proposition 4.1, theoretically one can guarantee that $\widehat{\chi}_{3,k}(\widehat{\Sigma}_k^{-1}; \varsigma_k, \delta)$, but not $\widehat{\chi}_{2,k}(\widehat{\Sigma}_k^{-1}; \varsigma_k, \delta)$, is asymptotically equivalent to the oracle test $\widehat{\chi}_{2,k}(\Sigma_k^{-1}; \varsigma_k, \delta)$. The proof can be found in Appendix A.5.

**Proposition 4.2.** *Under the conditions in Proposition 4.1 with Condition W, $k\log(k) = o(n)$ together with the additional restriction*

$$\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1) \neq o\left(\mathbb{L}_{\theta,2,\widehat{b},k}\mathbb{L}_{\theta,2,\widehat{p},k}\right), \tag{4.5}$$

*for any given $\varsigma_k, \delta > 0$, suppose that $\frac{|\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)|}{\mathsf{s.e.}_\theta[\widehat{\psi}_1]} = \gamma$ for some (sequence) $\gamma = \gamma(n)$ (where $\gamma(n)$ can diverge with $n$), then the rejection probability of $\widehat{\chi}_{3,k}(\widehat{\Sigma}_k^{-1}; \varsigma_k, \delta)$ converges to*

$$2 - \Phi\left(\varsigma_k - \lim_{n\to\infty}(\gamma - \delta)\frac{\mathsf{s.e.}_\theta[\widehat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})]}\right) - \Phi\left(\varsigma_k + \lim_{n\to\infty}(\gamma + \delta)\frac{\mathsf{s.e.}_\theta[\widehat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})]}\right) \tag{4.6}$$

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

(1) *under $\mathsf{H}_{0,k}(\delta) : \gamma \leqslant \delta$, $\widehat{\chi}_{3,k}(\widehat{\Sigma}_k^{-1}; \varsigma_k, \delta)$ rejects the null with probability less than or equal to $2(1 - \Phi(\varsigma_k))$, as $n \to \infty$;*

(2) *under the following alternative to $\mathsf{H}_{0,k}(\delta)$: $\gamma = \delta + c$, for any diverging sequence $c = c(n) \to \infty$, $\widehat{\chi}_{3,k}(\widehat{\Sigma}_k^{-1}; \varsigma_k, \delta)$ rejects the null with probability converging to 1, as $n \to \infty$.*

   (1′) *If $\mathbb{L}_{\theta,2,\widehat{b},k}$ and $\mathbb{L}_{\theta,2,\widehat{p},k}$ converge to 0, under the following alternative to $\mathsf{H}_{0,k}(\delta)$: $\gamma = \delta + c$, for any fixed $c > 0$ or any diverging sequence $c = c(n) \to \infty$, $\widehat{\chi}_{3,k}(\widehat{\Sigma}_k^{-1}; \varsigma_k, \delta)$ has rejection probability converging to 1, as $n \to \infty$.*

*If Condition W is replaced by Condition SW, we have the following instead of (1), (2) and (2′):*

(i) *under $\mathsf{H}_{0,k}(\delta) : \gamma \leqslant \delta$, $\widehat{\chi}_{3,k}(\widehat{\Sigma}_k^{-1}; \varsigma_k, \delta)$ rejects the null with probability equal to $2(1 - \Phi(\varsigma_k))$, as $n \to \infty$;*

(ii) *under the following alternative to $\mathsf{H}_{0,k}(\delta)$: $\gamma = \delta + c$, for any diverging sequence $c = c(n) \gg k\mathbb{L}^2_{\theta,2,\widehat{b},k}\mathbb{L}^2_{\theta,2,\widehat{p},k} = k\mathsf{CSBias}_{\theta,k}(\widehat{\psi}_1)$, $\widehat{\chi}_{3,k}(\widehat{\Sigma}_k^{-1}; \varsigma_k, \delta)$ rejects the null with probability converging to 1, as $n \to \infty$.*

**Comment 4.2** (Comments on Proposition 4.2)**.**

(1) *As a consequence, under the conditions of Proposition 4.2, $\widehat{\chi}_{3,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^\dagger/2}, \delta)$ is an asymptotically valid level $\alpha^\dagger$ two-sided test of the surrogate null $\mathsf{H}_{0,k}(\delta)$, and hence by Lemma 3.3(2) it is also an asymptotically $\alpha^\dagger$ level test of $\mathsf{H}_{0,\mathsf{CS}}(\delta)$. Thus when $\widehat{\chi}_{3,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^\dagger/2}, \delta)$ rejects $\mathsf{H}_{0,k}(\delta)$, it also rejects $\mathsf{H}_{0,\mathsf{CS}}(\delta)$ and by*

*adopting Condition CS, we conclude that we have no justification for the coverage validity of* $\mathsf{CI}_\alpha(\widehat{\psi}_1)$. *It should be noted that* $\widehat{\chi}_{3,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^\dagger/2}, \delta)$ *can be a powerless test for* $\mathsf{H}_{0,\mathsf{CS}}(\delta)$ *under certain laws* $\mathsf{P}_\theta$.

*If one is willing to assume the faithfulness Condition Faithfulness*($\delta'$), *then* $\widehat{\chi}_{3,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^\dagger/2}, \delta)$ *is an asymptotically valid level* $\alpha^\dagger$ *test of* $\mathsf{H}_0(\delta\delta')$. *As above, it can be a powerless test of* $\mathsf{H}_0(\delta\delta')$.

(2) *We cannot show that* $\widehat{\chi}_{2,k}(\widehat{\Sigma}_k^{-1}; \varsigma_k, \delta)$ *is asymptotically equivalent to the oracle test* $\widehat{\chi}_{2,k}(\Sigma_k^{-1}; \varsigma_k, \delta)$. *It would be true if the bias* $\mathsf{EB}_{\theta,2,k}(\widehat{\Sigma}_k^{-1})$ *is dominated by* $\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})] \asymp \frac{1}{\sqrt{n}}\left(\sqrt{\frac{k}{n}} + \mathbb{L}_{\theta,2,\widehat{b},k} + \mathbb{L}_{\theta,2,\widehat{p},k}\right)$. *Consider the following scenario: Under* $\mathsf{H}_{0,k}(\delta)$ (3.12), *the further restriction* $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1) \neq o\left(\mathbb{L}_{\theta,2,\widehat{b},k}\mathbb{L}_{\theta,2,\widehat{p},k}\right)$ *and* $\mathbb{L}_{\theta,2,\widehat{b},k} \asymp \mathbb{L}_{\theta,2,\widehat{p},k} \leqslant n^{-1/4}$, *however, we are only able to obtain the following upper bound:*

$$\begin{aligned}\mathsf{EB}_{\theta,2,k}(\widehat{\Sigma}_k^{-1}) &\lesssim \mathbb{L}_{\theta,2,\widehat{b},k}\mathbb{L}_{\theta,2,\widehat{p},k}\sqrt{\frac{k\log(k)}{n}} \\ &\ll \sqrt{\frac{1}{n}}\sqrt{\frac{k\log(k)}{n}} = \frac{\sqrt{k\log(k)}}{n}\end{aligned}$$

*with an extra* $\sqrt{\log(k)}$ *factor compared to the desired rate* $\sqrt{k}/n$. *It is possible that our upper bound on* $\mathsf{EB}_{\theta,2,k}(\widehat{\Sigma}_k^{-1})$ *is not tight, but this remains an open theoretical problem.*

*The loss in power under Condition SW but not Condition W, compared to the oracle test* $\widehat{\chi}_{2,k}(\Sigma_k^{-1}; \zeta_k, \delta)$, *has been brought up in Section 2.1 of Liu et al. (2020b). When SW(3) holds but Condition W(3′) does not hold, the standard error of an* $m$*-th order influence function estimator has a term of order* $\sqrt{\frac{k}{n}}\mathsf{CSBias}_{\theta,k}(\widehat{\psi}_1)$. *This term comes from the linear term in the Hoeffding decomposition of* $m$*-th order U-statistics and dominates when* $k < n$. *To have power against* $\mathsf{H}_{0,k}(\delta)$ *converging to 1, we need the "signal"* $\gamma$ *to exceed the "noise"* $\frac{\mathsf{s.e.}_\theta[\widehat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})]}$, *that is,*

$$\begin{aligned}&\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)^2 \equiv \left\{\mathbb{E}_\theta\left[\Pi_\theta[\lambda^{1/2}(\widehat{b}-b)|\bar{v}_k](X)\Pi_\theta[\lambda^{1/2}(\widehat{p}-p)|\bar{v}_k](X)\right]\right\}^2 \\ \gg &\left\{\sqrt{\frac{k}{n}}\mathsf{CSBias}_{\theta,k}(\widehat{\psi}_1)\right\}^2 \equiv \frac{k}{n}\mathbb{E}_\theta\left[\Pi_\theta[\lambda^{1/2}(\widehat{b}-b)|\bar{v}_k](X)^2\right]\mathbb{E}_\theta\left[\Pi_\theta[\lambda^{1/2}(\widehat{p}-p)|\bar{v}_k](X)^2\right].\end{aligned}$$

*However, we remark that such loss in power happens only if the infinity norms of the projections* $\Pi_\theta[\Delta_{1,\widehat{b}}|\bar{v}_k]$ *and* $\Pi_\theta[\Delta_{1,\widehat{p}}|\bar{v}_k]$ *are not bounded even though those of* $\Delta_{1,\widehat{b}}$ *and* $\Delta_{1,\widehat{p}}$ *are bounded under Condition SW.*

(3) *The additional restriction* $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1) \neq o(\mathbb{L}_{\theta,2,\widehat{b},k}\mathbb{L}_{\theta,2,\widehat{p},k})$ *will often hold when we estimate both nuisance functions* $b$ *and* $p$ *from a single training sample. However when we further split the training sample to estimate the nuisance functions from separate independent subsamples, it is possible to construct estimators that violate this restriction. For example, see Kennedy (2020); Newey and Robins (2018). In Section 6, we will propose a procedure that relies on even higher-order influence functions and successively relaxes this additional restriction but at the expense of higher computational cost. Therefore we need to first present the general theory of higher order influence functions for DR functionals in next section.*

# 5. Higher order influence functions for the DR functionals

5.1. **A review of the theory of higher order influence functions.** As mentioned above, in this section, we present the higher order influence functions (HOIFs) for the DR functionals (Rotnitzky et al., 2019) under a (locally) nonparametric model of every order with the derivation deferred to Appendix A.1. First, we need to formally define HOIFs of a generic functional $\psi(\theta)$, not necessarily in the MBF/DR class. In particular, we use the characterization of HOIFs under the (locally) nonparametric models given in Theorem 2.3 of Robins et al. (2008). We note that in the following $\mathbb{IF}_m(\theta)$ is a $m$-th order U-statistic and $\mathbb{IF}_{mm}(\theta)$ is a degenerate $m$-th order U-statistic under $\mathsf{P}_\theta$.

**Definition 5.1.** *For a generic functional $\psi(\theta)$, under a (locally) nonparametric model for which the first order influence function $\mathsf{IF}_1(\theta)$ exists. For $m \geqslant 2$, if the $(m-1)$-th order influence function $\mathbb{IF}_{m-1}(\theta)$ exists, then there exists a $m$-th order influence function $\mathbb{IF}_m(\theta)$ of $\psi(\theta)$, defined recursively via the following rule:*

$$\mathbb{IF}_1(\theta) \equiv \mathbb{IF}_{11}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \mathsf{IF}_{1,i}(\theta),$$

$$\mathbb{IF}_m(\theta) \equiv \mathbb{IF}_{m-1}(\theta) + \mathbb{IF}_{mm}(\theta)$$

*where $m\mathbb{IF}_{mm}(\theta)$ is the $m$-th order degenerate U-statistic component of the Hoeffding decomposition of*

$$\frac{(n-m)!}{n!} \sum_{1 \leqslant i_1 \neq \cdots \neq i_m \leqslant n} \mathsf{if}_{1,\mathsf{if}_{m-1,m-1}(O_{i_1},\cdots,O_{i_{m-1}};\cdot)}(O_{i_m};\theta),$$

*if, for each $(o_{i_1}, \cdots, o_{i_{m-1}})$ in the support of $(O_{i_1}, \cdots, O_{i_{m-1}})$, the first order influence function*

$$\mathsf{IF}_{1,\mathsf{if}_{m-1,m-1}(o_{i_1},\cdots,o_{i_{m-1}};\cdot)}(O_{i_m};\theta)$$

*exists for the functional*

$$\mathsf{if}_{m-1,m-1}(o_{i_1}, \cdots, o_{i_{m-1}};\theta))$$

*given by the kernel of the $(m-1)$-th order U-statistic $\mathbb{IF}_{m-1,m-1}(\theta)$ evaluated at fixed data $o_{i_1}, \cdots, o_{i_{m-1}}$. Otherwise $\mathbb{IF}_m(\theta)$ does not exist.*

The general use of HOIF estimators is to decrease bias as given in the following Theorem 2.2 of Robins et al. (2008) .

**Theorem 5.1** (Theorem 2.2 of Robins et al. (2008))**.** *Given an $m$-th order influence function $\mathbb{IF}_m(\theta)$ for $\psi(\theta)$, if both $\mathsf{E}_\theta[\mathbb{IF}_m(\widehat{\theta})]$ and $\psi(\widehat{\theta}) - \psi(\theta) \equiv \widehat{\psi} - \psi(\theta)$ have bounded Fréchet derivatives with respect to $\widehat{\theta}$ to order $m$ for a norm $\|\cdot\|$, then the estimator $\psi(\widehat{\theta}) + \mathbb{IF}_m(\widehat{\theta})$ has bias of order $\|\widehat{\theta} - \theta\|^{m+1}$ as an estimator of $\psi(\theta)$ where $\widehat{\theta}$ and $\mathbb{IF}_m(\cdot)$ are computed from independent samples.*

5.2. **Higher order influence functions of the DR functionals.** With the above preparation, we are ready to state the following theorem characterizing the HOIFs of the DR functionals:

**Theorem 5.2.** *Under the conditions of Theorem 2.2, in a locally nonparametric model, under $\mathsf{P}_{\widehat{\theta}}$, the $m$-th order influence function of $\widetilde{\psi}_k(\theta)$ is $\widehat{\mathbb{IF}}_{m,k}(\Sigma_k^{-1}) = \widehat{\mathbb{IF}}_1 - \widehat{\mathbb{IF}}_{22\to mm,k}(\Sigma_k^{-1})$, where $\widehat{\mathbb{IF}}_{\ell\ell\to mm,k}(\Sigma_k^{-1}) \coloneqq \sum_{j=\ell}^{m} \widehat{\mathbb{IF}}_{jj,k}(\Sigma_k^{-1})$,*

$$\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1}) \coloneqq \frac{1}{n(n-1)} \sum_{1\leqslant i_1 \neq i_2 \leqslant n} \widehat{\mathsf{IF}}_{22,k,\bar{i}_2}(\Sigma_k^{-1})$$

*and for $j > 2$,*

$$\widehat{\mathbb{IF}}_{jj,k}(\Sigma_k^{-1}) \coloneqq \frac{(n-j)!}{n!} \sum_{1\leqslant i_1 \neq \ldots \neq i_j \leqslant n} \widehat{\mathsf{IF}}_{jj,k,\bar{i}_j}(\Sigma_k^{-1}).$$

*Here*

$$\widehat{\mathsf{IF}}_{22,k,\bar{i}_2}(\Sigma_k^{-1}) \equiv \widehat{\mathsf{IF}}_{22,\widetilde{\psi}_k,\bar{i}_2}(\theta) = \left[\mathcal{E}_{1,\widehat{b}}(\bar{\mathsf{z}}_k)(O)\right]_{i_1}^{\top} \Sigma_k^{-1} \left[\mathcal{E}_{2,\widehat{p}}(\bar{\mathsf{z}}_k)(O)\right]_{i_2},$$

*and*

$$\widehat{\mathsf{IF}}_{jj,k,\bar{i}_j}(\Sigma_k^{-1}) \equiv \widehat{\mathsf{IF}}_{jj,\widetilde{\psi}_k,\bar{i}_j}(\theta)$$

$$= (-1)^j \left[\mathcal{E}_{1,\widehat{b}}(\bar{\mathsf{z}}_k)(O)\right]_{i_1}^{\top} \left\{\prod_{s=3}^{j} \Sigma_k^{-1} \left(\left[S_{bp}\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^{\top}\right]_{i_s} - \Sigma_k\right)\right\} \Sigma_k^{-1} \left[\mathcal{E}_{2,\widehat{p}}(\bar{\mathsf{z}}_k)(O)\right]_{i_2}. \tag{5.1}$$

*are the kernels of second order influence function $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ and $j$-th order influence function $\widehat{\mathbb{IF}}_{jj,k}(\Sigma_k^{-1})$ respectively, where $\bar{i}_j \coloneqq \{i_1, \ldots, i_j\}$.*

*Furthermore, define*

$$\widehat{\mathbb{IF}}_{22\to mm,k}(\Sigma_k^{-1}) \coloneqq \sum_{j=2}^{m} \widehat{\mathbb{IF}}_{jj,k}(\Sigma_k^{-1}).$$

*As a consequence, $\widehat{\mathbb{IF}}_{22\to mm,k}(\Sigma_k^{-1})$ is the $m$-th order influence function of $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$ under $\mathsf{P}_{\widehat{\theta}}$.*

The procedure that will be proposed in the next section in order to relax the additional restriction $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1) \neq o(\mathbb{L}_{\theta,2,\widehat{b},k}\mathbb{L}_{\theta,2,\widehat{p},k})$ imposed in Proposition 4.2 is based on the estimates of the above HOIFs of $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$ under the law $\mathsf{P}_{\widehat{\theta}}$. These estimates simply replace $\Sigma_k^{-1}$ by $\widehat{\Sigma}_k^{-1}$.

**Comment 5.1.** *Note that:*

(1) *$\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})$ is an $m$-th order U-statistic. The greater $m$, the higher the computational cost incurred in computing $\widehat{\mathbb{IF}}_{\ell\ell\to mm,k}(\widehat{\Sigma}_k^{-1})$;*

(2) *In the oracle setting (see Section 3), i.e. when $\Sigma_k^{-1}$ is known, the $m$-th order influence function of $\widetilde{\psi}_{k,\theta}$ is 0 when $m > 2$.*

Next, we obtain the following bounds on the estimation bias and variance of $\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})$ as an estimator of $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1) \equiv \mathsf{E}_{\theta}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]$; similar to Proposition 4.1, the estimation bias bounds are proved in Mukherjee et al. (2017) and the variance bounds are an immediate consequence of Lemma A.2 proved in Appendix A.2.

**Proposition 5.1.** *Under the conditions of Theorem 3.2 with Condition W, when $k\log(k) = o(n)$ and $km^2 = o(n)$, the followings hold on the event that $\widehat{\Sigma}_k$ is invertible,*

$$\left|\mathsf{EB}_{\theta,mm,k}(\widehat{\Sigma}_k^{-1})\right| \lesssim \mathbb{L}_{\theta,2,\widehat{b},k}\mathbb{L}_{\theta,2,\widehat{p},k}\mathbb{L}_{\theta,2,\widehat{\Sigma},k}^{m-1} \lesssim \mathbb{L}_{\theta,2,\widehat{b},k}\mathbb{L}_{\theta,2,\widehat{p},k}\left(\frac{k\log(k)}{n}\right)^{\frac{m-1}{2}},$$
$$\mathsf{var}_\theta\left[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})\right] \lesssim \frac{1}{n}\left\{\frac{k}{n} + \mathbb{L}_{\theta,2,\widehat{b},k}^2 + \mathbb{L}_{\theta,2,\widehat{p},k}^2\right\}$$

*where*

$$\mathsf{EB}_{\theta,m,k}(\widehat{\Sigma}_k^{-1}) \coloneqq \mathsf{E}_\theta[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})] - \mathsf{Bias}_{\theta,k}(\widehat{\psi}_1).$$

*If Condition SW holds instead of Condition W:*

$$\mathsf{var}_\theta\left[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})\right] \lesssim \frac{1}{n}\left\{\frac{k}{n} + \mathbb{L}_{\theta,2,\widehat{b},k}^2 + \mathbb{L}_{\theta,2,\widehat{p},k}^2 + k\mathbb{L}_{\theta,2,\widehat{b},k}^2\mathbb{L}_{\theta,2,\widehat{p},k}^2\right\}.$$

*In addition, for any $m \geqslant 3$,*

$$\frac{\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1}) - \mathsf{E}_\theta[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]} \xrightarrow{d} N(0,1).$$

**Comment 5.2.**

- *One can estimate $\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]$ by a bootstrap estimator $\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]$, which can be derived by generalizing the argument given in Appendix A.3.*
- *As $m$ increases, $\mathsf{EB}_{\theta,m,k}(\widehat{\Sigma}_k^{-1})$ decreases. And this is a manifestation of Theorem 5.1.*

## 6. Higher order tests to relax the additional restriction in equation (4.5)

With the forms of HOIFs for the DR functionals presented in Section 5, we are now ready to present the higher order testing procedure that helps relax the additional restriction in equation (4.5). To this end, we first define the following test that has the same property as $\widehat{\chi}_{3,k}(\widehat{\Sigma}_k^{-1};\varsigma_k,\delta)$ under a weakened version of the additional restriction $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1) \neq o(\mathbb{L}_{\theta,2,\widehat{b},k}\mathbb{L}_{\theta,2,\widehat{p},k})$ imposed in Proposition 4.2: for any fixed $m \geqslant 3$, define

$$\widehat{\chi}_{m,k}(\widehat{\Sigma}_k^{-1};\varsigma_k,\delta) \coloneqq \mathbb{1}\left\{\frac{|\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})|}{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]} - \varsigma_k\frac{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]}{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]} > \delta\right\}. \tag{6.1}$$

Proposition 5.1 immediately implies the following:

**Proposition 6.1.** *Every statements in Proposition 4.2 about the test $\widehat{\chi}_{3,k}(\widehat{\Sigma}_k^{-1};\varsigma_k,\delta)$ hold for $\widehat{\chi}_{m,k}(\widehat{\Sigma}_k^{-1};\varsigma_k,\delta)$, under the same conditions in Proposition 4.2, except that the additional restriction $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1) \neq o(\mathbb{L}_{\theta,2,\widehat{b},k}\mathbb{L}_{\theta,2,\widehat{p},k})$ is relaxed to $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1) \neq o\left(\mathbb{L}_{\theta,2,\widehat{b},k}\mathbb{L}_{\theta,2,\widehat{p},k}\left(\frac{k\log(k)}{n}\right)^{\frac{m-3}{2}}\right)$.*

**Comment 6.1.** *Note that the greater $m$ is, the less restrictive is $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1) \neq o\left(\mathbb{L}_{\theta,2,\widehat{b},k}\mathbb{L}_{\theta,2,\widehat{p},k}\left(\frac{k\log(k)}{n}\right)^{\frac{m-3}{2}}\right)$. In fact we can let $m \equiv m(n) \to \infty$ as $n \to \infty$.*

As remarked in Comment 5.1, although increasing $m$ relaxes the restriction (4.5), it incurs higher computational cost. In practice, one might want $m$ to be as small as possible. In view of such computational concern in practice, we propose the following "early-stopping" test that stops increasing $m$ if $\widehat{\chi}_{m,k}(\widehat{\Sigma}_k^{-1}; \varsigma_k, \delta)$ fails to reject $\mathsf{H}_{0,k}(\delta)$ at a smaller $m$: for a fixed integer $M > 0$, define

$$\widehat{\chi}_{M,k}^{es}(\widehat{\Sigma}_k^{-1}; \varsigma_k, \delta) \coloneqq \mathbb{1}\left\{ \frac{|\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})|}{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]} - \varsigma_k \frac{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]}{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]} > \delta : \forall\, m = 2, \ldots, M \right\}. \tag{6.2}$$

If $\widehat{\chi}_{M,k}^{es}(\widehat{\Sigma}_k^{-1}; \varsigma_k, \delta)$ fails to reject $\mathsf{H}_{0,k}(\delta)$ at some $m \leqslant M$, we stop the test at $m$ and claim that we fail to reject $\mathsf{H}_{0,k}(\delta)$. This "early stopping" procedure, with $\varsigma_k$ set to $z_{\alpha^\dagger/2}$ is a conservative $\alpha^\dagger$-level test under much weaker conditions:

**Proposition 6.2.** *Under the conditions of Proposition 4.2, but with (4.5) replaced by*

$$\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1) \neq o\left( \mathbb{L}_{\theta,2,\widehat{b},k} \mathbb{L}_{\theta,2,\widehat{p},k} \left( \frac{k \log(k)}{n} \right)^{\frac{M-3}{2}} \right),$$

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

The power of the above "early-stopping" procedure is more challenging to characterize and it may require more assumptions on the estimation biases $\mathsf{EB}_{\theta,m,k}(\widehat{\Sigma}_k^{-1})$ for each $m$. We leave this problem to the future.

## 7. Monte Carlo experiments

In the Monte Carlo (MC) experiments, we focus on parameter the counterfactual mean of $Y$ when $\{0,1\}$-valued $A$ is set to 1: $\psi^\dagger(\theta) \equiv \mathsf{E}_\theta[Y(a=1)] \equiv \mathsf{E}_\theta[b(X)]$ under ignorability. Here $p(X) = 1/\mathsf{E}_\theta[A|W]$ is the inverse propensity score and $b(X) = \mathsf{E}_\theta[Y|A=1, X]$ is the conditional mean of the outcome in the treatment group. For this parameter, $\Sigma_k = \mathsf{E}_\theta[A \bar{\mathsf{z}}_k(X) \bar{\mathsf{z}}_k(X)^\top]$ and $\widehat{\Sigma}_k = n^{-1} \sum_{i \in \mathsf{tr}} A_i \bar{\mathsf{z}}_k(X_i) \bar{\mathsf{z}}_k(X_i)^\top$. We choose $\psi^\dagger(\theta) \equiv 0$.

We consider two simulation setups. In simulation setup I, we draw $N = 100{,}000$ i.i.d. $X_j$ for $j = 1, \ldots, 4$ (so $d = 4$). The marginal density $f$ of each $X_j$ is supported on $[0,1]$ with $f \in$ Hölder$(0.1 + c)$ for some small $c > 0$, as defined in Appendix B. The correlation between each pair of $X_j$ and $X_k$, with $j \neq k$, is introduced based on the algorithm described in Appendix B.1. We then simulate $Y$ and $A$ according to the following data generating mechanism:

$$Y \sim b(X) + N(0,1) = \sum_{j=1}^{4} \tau_{b,j} h_b(X_j; 0.25) + N(0,1)$$

and

$$A \sim \mathsf{Bernoulli}\left( 1/p(X) \equiv \mathsf{expit}\left\{ \sum_{j=1}^{4} \tau_{p,j} h_p(X_j; 0.25) \right\} \right)$$

where $h_b(\cdot; 0.25)$ and $h_p(\cdot; 0.25)$ have the forms as defined in Appendix B and hence both belong to Hölder$(0.25 + c)$ for some very small $c > 0$. The numerical values for $(\tau_{b,j}, \tau_{p,j})_{j=1}^4$ are provided in Table 5. *We fix half of the* $N = 100,000$ *samples as the training sample so* $n_{\mathsf{tr}} = n = 50,000$ *and only consider the randomness from the estimation sample in the simulation.* In simulation setup II, we consider the same data generating mechanism as in setup I except that we choose $b(X) = \sum_{j=1}^4 \tau_{b,j} h_b(X_j; 0.6)$ and $1/p(X) \equiv \mathsf{expit}\left\{\sum_{j=1}^4 \tau_{p,j} h_p(X_j; 0.6)\right\}$ where $h_b(\cdot; 0.6)$ and $h_p(\cdot; 0.6)$ have the forms as defined in Appendix B and hence both belong to Hölder$(0.6 + c)$ for some very small $c > 0$.

We choose D12 (or equivalently db6) Daubechies wavelets (Daubechies, 1992) at resolutions $\ell \in (6, 7, 8)$ to form the basis functions

$$\bar{\mathsf{z}}_k(X) = (\bar{\mathsf{z}}_{k'}(X_1)^\top, \bar{\mathsf{z}}_{k'}(X_2)^\top, \bar{\mathsf{z}}_{k'}(X_3)^\top, \bar{\mathsf{z}}_{k'}(X_4)^\top)^\top,$$

with the corresponding $k'^6 = 64, 2^7 = 128, 2^8 = 256\}$ and $k \in \{64 \cdot 4 = 256, 128 \cdot 4 = 512, 256 \cdot 4 = 1024\}$. To compute the oracle statistics and tests, we evaluate $\Sigma_k$ through Monte Carlo integration by simulating $L = 10^7$ independent $(A, X)$ from the true data generating law. To investigate the finite sample performance of the statistical procedures developed in this article, all the summary statistics of the Monte Carlo experiments are calculated based on 100 replicates. We estimate the nuisance functions $1/p(x)$ and $b(x)$ using generalized additive models (GAMs) (Hastie and Tibshirani, 1986, 1987). In particular, the smoothing parameters were selected by generalized cross validation, the default setup in gam function from R package mgcv (Wood et al., 2016). We choose db6 father wavelets to construct HOIF estimators when analyzing the simulated data.

7.1. **Finite sample performance of tests for** $\mathsf{H}_{0,k}(\delta)$ **(3.12).** In this section, we consider testing the null hypothesis $\mathsf{H}_{0,k}(\delta)$ (3.12). Henceforth we investigate the finite sample performance of the oracle statistics and tests $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$, $\widehat{\psi}_{2,k}(\Sigma_k^{-1})$, and $\widehat{\chi}_{2,k}(\Sigma_k^{-1})$, together with the statistics and tests relying on $\widehat{\Sigma}_k^{-1}$: $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$, $\widehat{\mathbb{IF}}_{33,k}(\widehat{\Sigma}_k^{-1})$, $\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})$, $\widehat{\psi}_{2,k}(\widehat{\Sigma}_k^{-1})$, $\widehat{\psi}_{3,k}(\widehat{\Sigma}_k^{-1})$, $\widehat{\chi}_{2,k}(\widehat{\Sigma}_k^{-1})$, and $\widehat{\chi}_{3,k}(\widehat{\Sigma}_k^{-1})$.

First, we check the asymptotic normalities of $\frac{\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]}$, $\frac{\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})]}$, $\frac{\widehat{\mathbb{IF}}_{33,k}(\widehat{\Sigma}_k^{-1})}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{33,k}(\widehat{\Sigma}_k^{-1})]}$, and $\frac{\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})]}$ through normal qq-plots displayed in Figures 1 (simulation setup I) and 2 (simulation setup II). We observe that the distributions of most of these statistics are close to normal at different $k$'s ($k = 256$: left panels; $k = 512$: middle panels; $k = 1024$: right panels). In the third row of Figures 1 and 2), we observe some deviation from normality in the tails of $\frac{\widehat{\mathbb{IF}}_{33,k}(\widehat{\Sigma}_k^{-1})}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{33,k}(\widehat{\Sigma}_k^{-1})]}$, but the asymptotic distribution of $\frac{\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})]}$ is not much affected and quite close to normal based on the qqplot.

In the simulation, we use nonparametric bootstrap to estimate the standard errors of $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$, $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$, $\widehat{\mathbb{IF}}_{33,k}(\widehat{\Sigma}_k^{-1})$ and $\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})$, as described in Appendix A.3. Hence we study if the estimated standard errors of $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$, $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$, $\widehat{\mathbb{IF}}_{33,k}(\widehat{\Sigma}_k^{-1})$, and $\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})$ by nonparametric bootstrap are close to their true standard errors (calibrated by the MC variance from 100 replicates in the simulation). We use $B = 100$ bootstrap samples to compute the bootstrapped standard errors as the estimated standard errors for all four statistics. In Tables 1 (simulation setup I) and 3 (simulation setup II), we display the MC standard deviations (the upper numerical values in each cell), accompanied with the MC averages of the estimated standard errors (the lower numerical values outside the parenthesis in each cell) and *MC standard deviations* of the estimated standard errors (the lower numerical values inside the parenthesis in each cell) of all three statistics at $k = 256$ (left panel), $k = 512$ (middle panel) and $k = 1024$ (right panel). From Table 1 and Table 3, we observe that the estimated standard errors only slightly differ from the MC standard deviations.

Then we investigate (1) the finite sample performance of $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$, $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$, and $\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})$ and evaluate how close they are to $\mathsf{Bias}_\theta(\widehat{\psi}_1)$, which is evaluated based on the MC bias of 100 replicates, and (2) the rejection rate of the tests $\widehat{\chi}_{2,k}(\Sigma_k^{-1}; z_{0.10/2}, \delta)$, $\widehat{\chi}_{2,k}(\widehat{\Sigma}_k^{-1}; z_{0.10/2}, \delta)$ and $\widehat{\chi}_{3,k}(\Sigma_k^{-1}; z_{0.10/2}, \delta)$ for the null hypothesis $\mathsf{H}_{0,k}(\delta)$ (3.12). The numerical results are shown in Tables 2 (simulation setup I) and 4 (simulation setup II).

- In the upper panel of Table 2 (simulation setup I), the first row and the third column, we display the MC bias of the DML estimator $\widehat{\psi}_1$ and the MC average of $\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]$, which are $-34.26 \times 10^{-3}$ and $8.77 \times 10^{-3}$. Since the ratio between the bias and standard error is around 4, we expect the associated 90% Wald confidence interval $\widehat{\psi}_1 \pm z_{0.05}\widehat{\mathsf{s.e.}}(\widehat{\psi}_1)$ does not have the nominal coverage. This is indeed the case by reading from the first row, the second column of the upper panel of Table 2, showing the MC coverage probability for the 90% Wald confidence interval of $\widehat{\psi}_1$ is 0%.

  Similarly, in the upper panel of Table 4 (simulation setup II), the first row and the third column, we display the MC bias of the DML estimator $\widehat{\psi}_1$ and the MC average of $\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]$, which are $-8.88 \times 10^{-3}$ and $8.10 \times 10^{-3}$. This is as expected because the true nuisance functions $b$ and $p$ belong to Hölder$(0.6 + c)$ for some $c > 0$ and DML estimator (with $b$ and $p$ estimated in optimal rate in $L_2$ norm) is expected to have bias of $o(n^{-1/2})$ when the average smoothness between $b$ and $p$ is above 0.5 (Robins et al., 2009b). Here the ratio between the bias and standard error is around 1 and we expect the associated 90% Wald confidence interval $\widehat{\psi}_1 \pm z_{0.05}\widehat{\mathsf{s.e.}}(\widehat{\psi}_1)$ has near nominal coverage. This is indeed the case by reading from the first row, the second column of the upper panel of Table 4, showing the MC coverage probability for the 90% Wald confidence interval of $\widehat{\psi}_1$ is 83%.
- In the second column of the upper panel of Table 2, we display the MC averages of $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ and the MC averages of its estimated standard errors; in the middle column, we display those of $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$; and in the lower panel, we display those of $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1}) + \widehat{\mathbb{IF}}_{33,k}(\widehat{\Sigma}_k^{-1})$ after third-order bias correction. In the upper panel, we observe that increasing $k$ from 256 to 512 does improve the amount of bias recovered by $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ (from $-18.76 \times 10^{-3}$ to $-25.34 \times 10^{-3}$), but there is no obvious difference in the MC averages of $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ between $k = 512$ and $k = 1024$

($-25.34 \times 10^{-3}$ and $-25.43 \times 10^{-3}$, about 75% of the total bias). The MC average of the estimated standard error increases with $k$, from $2.45 \times 10^{-3}$ at $k = 256$ to $3.43 \times 10^{-3}$ at $k = 1024$. This is consistent with the theoretical prediction that the variability of $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ should grow with $k$. In the middle, we observe very similar numerical results between $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ and $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$ for all the statistics that we are interested in. Interestingly, we did not see much improvement of using $\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})$ instead of $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$, when compared to $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$, at least in this example.

In the second column of the upper panel of Table 4, in the upper panel, we also observe that increasing $k$ from 256 to 512 also slightly improve the amount of bias recovered by $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ (from $-4.54 \times 10^{-3}$ to $-4.94 \times 10^{-3}$). The MC average of the estimated standard error increases with $k$, from $1.51 \times 10^{-3}$ at $k = 256$ to $2.43 \times 10^{-3}$ at $k = 1024$. In the middle panel, we observe very similar numerical results between $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ and $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$ for all the statistics that we are interested in. Interestingly, we again did not see much improvement of using $\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})$ instead of $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$, when compared to $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$, at least in this example.

- In the third columns of Table 2, we display the MC coverage probabilities of the 90% two-sided confidence intervals of $\widehat{\psi}_{2,k}(\Sigma_k^{-1})$ (upper panel) and $\widehat{\psi}_{3,k}(\widehat{\Sigma}_k^{-1})$ (lower panel). Comparing the oracle bias-corrected estimator $\widehat{\psi}_{2,k}(\Sigma_k^{-1})$ vs. $\widehat{\psi}_1$, the coverage probability improves from 0% to 33% (82%) at $k = 256$ (at $k = 1024$). Similar observations can be made for $\widehat{\psi}_{2,k}(\widehat{\Sigma}_k^{-1})$ and $\widehat{\psi}_{3,k}(\widehat{\Sigma}_k^{-1})$ as well when $\Sigma_k$ is unknown.

  In the third columns of Table 4, we display the MC coverage probabilities of the 90% two-sided confidence intervals of $\widehat{\psi}_{2,k}(\Sigma_k^{-1})$ (upper panel) and $\widehat{\psi}_{3,k}(\widehat{\Sigma}_k^{-1})$ (lower panel). Comparing the oracle bias-corrected estimator $\widehat{\psi}_{2,k}(\Sigma_k^{-1})$ vs. $\widehat{\psi}_1$, the coverage probability improves from 83% to 100%) at $k = 256$ and $k = 1024$. Similar observations can be made for $\widehat{\psi}_{2,k}(\widehat{\Sigma}_k^{-1})$ and $\widehat{\psi}_{3,k}(\widehat{\Sigma}_k^{-1})$ as well when $\Sigma_k$ is unknown.
- In the fourth columns of Tables 2 and 4, we display the MC biases of $\widehat{\psi}_{2,k}(\Sigma_k^{-1})$ (upper panel) and $\widehat{\psi}_{3,k}(\widehat{\Sigma}_k^{-1})$ (lower panel), together with their MC standard deviations (in the parentheses). As expected, the standard deviations of the estimators after bias correction are very similar to those of $\widehat{\psi}_1$. This indeed confirms that we are able to correct bias without drastically inflating the variance.
- In the fifth column of Table 2, we display the MC rejection rates of the test statistic: the upper panel shows the MC rejection rates of the oracle test $\widehat{\chi}_{2,k}(\Sigma_k^{-1}; z_{0.10/2}, \delta = 3/4)$ and the lower panel shows the MC rejection rates of the test $\widehat{\chi}_{3,k}(\widehat{\Sigma}_k^{-1}; z_{0.10/2}, \delta = 3/4)$. These simulation results demonstrate that when a "good" basis (db6 father wavelets) is used, our proposed test does have power to reject the null hypothesis of actual interest $\mathsf{H}_0(\delta)$ (3.10) that the ratio between the bias of $\widehat{\psi}_1$ and its standard error is lower than $\delta$. All the rejection rates in Table 2 are 100% when $\delta = 3/4$,

| $k$ | 256 | 512 | 1024 |
|---|---|---|---|
| $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ | 2.437<br>2.268 (0.218) | 3.011<br>2.716 (0.268) | 3.247<br>2.841 (0.322) |
| $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$ | 2.456<br>2.271 (0.218) | 3.075<br>2.778 (0.278) | 3.546<br>3.032 (0.378) |
| $\widehat{\mathbb{IF}}_{33,k}(\widehat{\Sigma}_k^{-1})$ | 0.495<br>0.603 (0.0598) | 0.814<br>1.126 (0.126) | 1.518<br>1.998 (0.285) |
| $\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})$ | 2.282<br>2.251 (0.227) | 2.745<br>2.738 (0.290) | 2.743<br>3.009 (0.427) |

TABLE 1. For data generating mechanism in simulation setup I: We reported the MC standard deviations $\times 10^{-3}$ (upper values in each cell), the MC averages of the estimated standard errors $\times 10^{-3}$ (lower values in each cell outside the parenthesis) and the MC standard deviations of the estimated standard errors $\times 10^{-3}$ (lower values in each cell inside the parenthesis) through nonparametric bootstrap resampling for $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$, $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$, $\widehat{\mathbb{IF}}_{33,k}(\widehat{\Sigma}_k^{-1})$ and $\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})$.

| k | $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1}) \times 10^{-3}$ | MC Coverage ($\widehat{\psi}_{2,k}(\Sigma_k^{-1})$ 90% Wald CI) | Bias($\widehat{\psi}_{2,k}(\Sigma_k^{-1})) \times 10^{-3}$ | $\widehat{\chi}_{2,k}(\Sigma_k^{-1}; z_{0.10/2}, \delta = 3/4)$ |
|---|---|---|---|---|
| 0 | NA (NA) | 0% | -34.26 (8.77) | NA |
| 256 | -18.76 (2.27) | 31% | -15.50 (8.60) | 100% |
| 512 | -25.34 (2.72) | 83% | -8.92 (8.61) | 100% |
| 1024 | -25.43 (2.84) | 82% | -8.83 (8.79) | 100% |
| k | $\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1}) \times 10^{-3}$ | MC Coverage ($\widehat{\psi}_{3,k}(\widehat{\Sigma}_k^{-1})$ 90% Wald CI) | Bias($\widehat{\psi}_{3,k}(\widehat{\Sigma}_k^{-1})) \times 10^{-3}$ | $\widehat{\chi}_{3,k}(\widehat{\Sigma}_k^{-1}; z_{0.10/2}, \delta = 3/4)$ |
| 256 | -18.54 (2.25) | 30% | -15.72 (8.60) | 100% |
| 512 | -24.65 (2.74) | 81% | -9.61 (8.60) | 100% |
| 1024 | -23.82 (3.01) | 76% | -10.44 (8.76) | 100% |

TABLE 2. For data generating mechanism in simulation setup I: We reported the MC averages of point estimates and standard errors (the first column in each panel) of $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ (upper panel) and $\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})$ (lower panel), together with the coverage probabilities of two-sided 90% Wald confidence intervals (the second column in each panel) of $\widehat{\psi}_{2,k}(\Sigma_k^{-1})$ (upper panel) and $\widehat{\psi}_{3,k}(\widehat{\Sigma}_k^{-1})$ (lower panel), the MC biases and MC standard deviations (the third column in each panel) of $\widehat{\psi}_{2,k}(\Sigma_k^{-1})$ (upper panel) and $\widehat{\psi}_{3,k}(\widehat{\Sigma}_k^{-1})$ (lower panel) and the empirical rejection rates (the fourth column in each panel) of $\widehat{\chi}^{(1)}_{2,k}(\Sigma_k^{-1}; z_{0.10/2}, \delta = 3/4)$ (upper panel) and $\widehat{\chi}^{(1)}_{3,k}(\widehat{\Sigma}_k^{-1}; z_{0.10/2}, \delta = 3/4)$ (lower panel). In the upper panel, for $k = 0$, $\widehat{\psi}_{2,k=0} \equiv \widehat{\psi}_1$.

regardless of whether we are using the oracle test or not. When $\delta = 2$, we obtain nontrivial rejection rates. This is as expected because the ratio between $\widehat{\mathbb{IF}}_{22,k}$ and $\widehat{\text{s.e.}}(\widehat{\psi}_1)$ is around 2 to 3. In the fifth column of Table 4, all the rejection rates are 0% when $\delta = 3/4$. This is also as expected because the ratio between $\widehat{\mathbb{IF}}_{22,k}$ and $\widehat{\text{s.e.}}(\widehat{\psi}_1)$ is close to $1/2$.

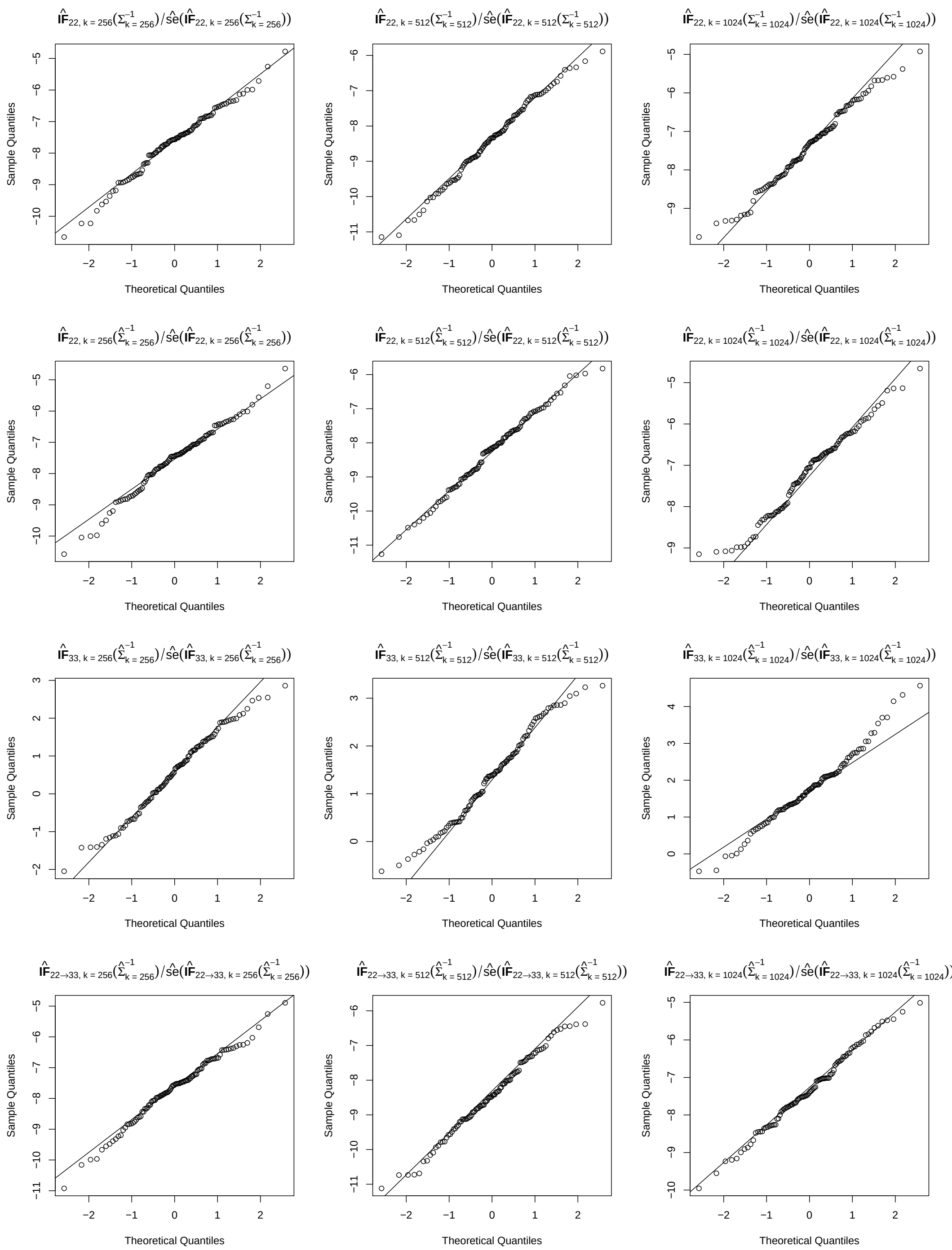


FIGURE 1. For data generating mechanism in simulation setup I: normal qq-plots for $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})/\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]$ (the first row), $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})/\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})]$ (the second row), $\widehat{\mathbb{IF}}_{33,k}(\widehat{\Sigma}_k^{-1})/\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{33,k}(\widehat{\Sigma}_k^{-1})]$ (the third row) and $\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})/\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})]$ (the fourth row) for $k = 256$ (left panels), $k = 512$ (middle panels) and $k = 1024$ (right panels).

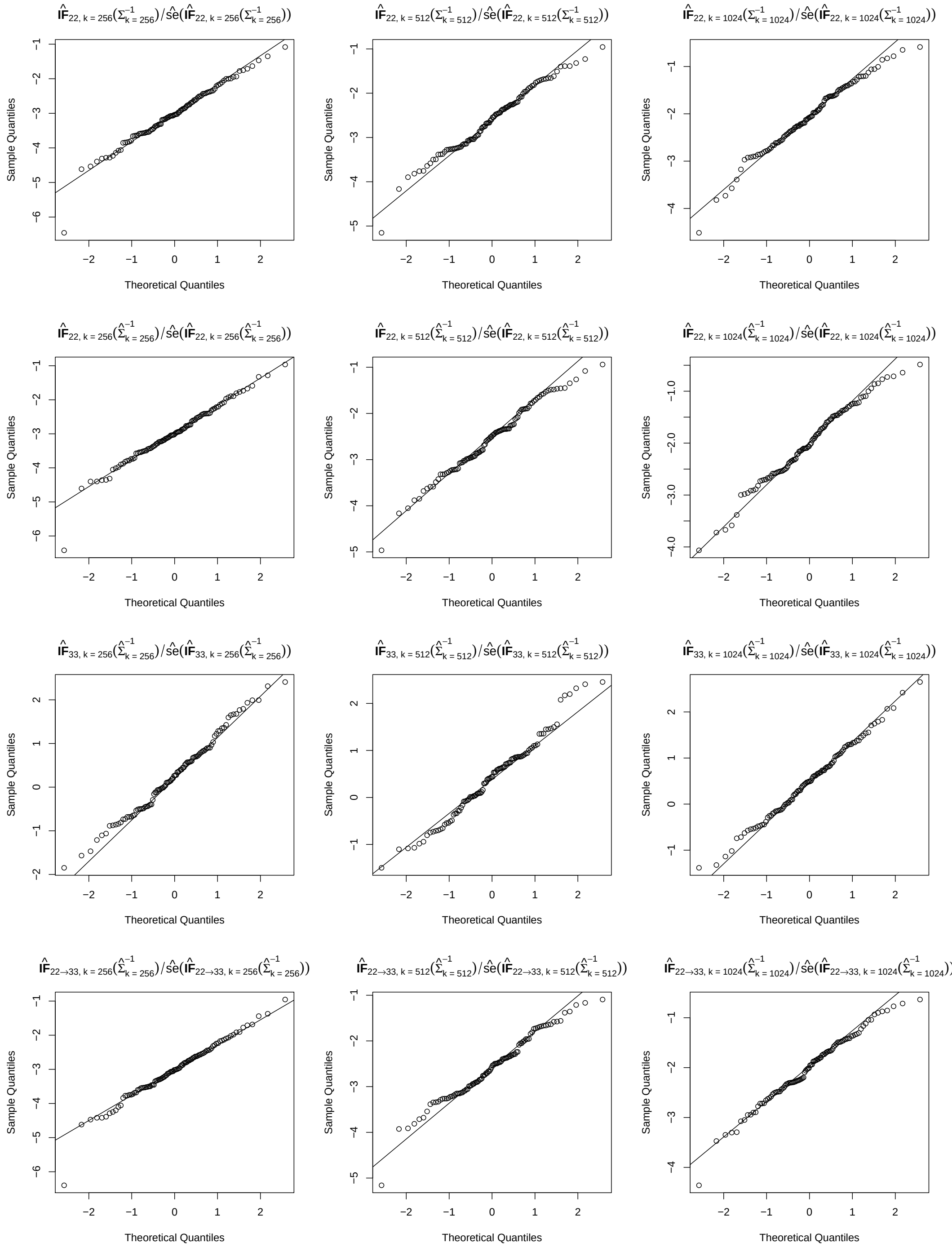


FIGURE 2. For data generating mechanism in simulation setup II: normal qq-plots for $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})/\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]$ (the first row), $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})/\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})]$ (the second row), $\widehat{\mathbb{IF}}_{33,k}(\widehat{\Sigma}_k^{-1})/\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{33,k}(\widehat{\Sigma}_k^{-1})]$ (the third row) and $\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})/\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})]$ (the fourth row) for $k = 256$ (left panels), $k = 512$ (middle panels) and $k = 1024$ (right panels).

| $k$ | 256 | 512 | 1024 |
|---|---|---|---|
| $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ | 1.166<br>1.209 (0.156) | 1.367<br>1.408 (0.191) | 1.673<br>1.584 (0.267) |
| $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$ | 1.187<br>1.219 (0.159) | 1.433<br>1.439 (0.198) | 1.828<br>1.697 (0.285) |
| $\widehat{\mathbb{IF}}_{33,k}(\widehat{\Sigma}_k^{-1})$ | 0.251<br>0.331 (0.0451) | 0.417<br>0.629 (0.0997) | 0.854<br>1.251 (0.247) |
| $\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})$ | 1.109<br>1.186 (0.156) | 1.256<br>1.346 (0.191) | 1.488<br>1.560 (0.383) |

Table 3. For data generating mechanism in simulation setup II: We reported the MC standard deviations $\times 10^{-3}$ (upper values in each cell), the MC averages of the estimated standard errors $\times 10^{-3}$ (lower values in each cell outside the parenthesis) and the MC standard deviations of the estimated standard errors $\times 10^{-3}$ (lower values in each cell inside the parenthesis) through nonparametric bootstrap resampling for $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$, $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$, $\widehat{\mathbb{IF}}_{33,k}(\widehat{\Sigma}_k^{-1})$ and $\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})$.

| k | $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1}) \times 10^{-3}$ | MC Coverage ($\widehat{\psi}_{2,k}(\Sigma_k^{-1})$ 90% Wald CI) | Bias($\widehat{\psi}_{2,k}(\Sigma_k^{-1})$) $\times 10^{-3}$ | $\widehat{\chi}_{2,k}(\Sigma_k^{-1}; z_{0.10/2}, \delta = 3/4)$ |
|---|---|---|---|---|
| 0 | NA (NA) | 83% | -8.88 (8.10) | NA |
| 256 | -4.54 (1.21) | 99% | -4.34 (8.13) | 0% |
| 512 | -4.89 (1.41) | 99% | -3.99 (8.21) | 0% |
| 1024 | -4.94 (1.58) | 100% | -3.94 (8.35) | 0% |
| k | $\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1}) \times 10^{-3}$ | MC Coverage ($\widehat{\psi}_{3,k}(\widehat{\Sigma}_k^{-1})$ 90% Wald CI) | Bias($\widehat{\psi}_{3,k}(\widehat{\Sigma}_k^{-1})$) $\times 10^{-3}$ | $\widehat{\chi}_{3,k}(\widehat{\Sigma}_k^{-1}; z_{0.10/2}, \delta = 3/4)$ |
| 256 | -4.49 (1.19) | 99% | -4.39 (8.13) | 0% |
| 512 | -4.76 (1.35) | 99% | -4.12 (8.20) | 0% |
| 1024 | -4.62 (1.56) | 100% | -4.26 (8.32) | 0% |

Table 4. For data generating mechanism in simulation setup II: We reported the MC averages of point estimates and standard errors (the first column in each panel) of $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ (upper panel) and $\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})$ (lower panel), together with the coverage probabilities of two-sided 90% Wald confidence intervals (the second column in each panel) of $\widehat{\psi}_{2,k}(\Sigma_k^{-1})$ (upper panel) and $\widehat{\psi}_{3,k}(\widehat{\Sigma}_k^{-1})$ (lower panel), the MC biases and MC standard deviations (the third column in each panel) of $\widehat{\psi}_{2,k}(\Sigma_k^{-1})$ (upper panel) and $\widehat{\psi}_{3,k}(\widehat{\Sigma}_k^{-1})$ (lower panel) and the empirical rejection rates (the fourth column in each panel) of $\widehat{\chi}^{(1)}_{2,k}(\Sigma_k^{-1}; z_{0.10/2}, \delta = 3/4)$ (upper panel) and $\widehat{\chi}^{(1)}_{3,k}(\widehat{\Sigma}_k^{-1}; z_{0.10/2}, \delta = 3/4)$ (lower panel). In the upper panel, for $k = 0$, $\widehat{\psi}_{2,k=0} \equiv \widehat{\psi}_1$.

## 8. Concluding remarks

In this paper, we developed nearly assumption-free/assumption-lean tests that can test if the Wald confidence interval $\mathsf{CI}_\alpha(\widehat{\psi}_1)$ associated with an DML estimator $\widehat{\psi}_1$ of a DR functional has nominal coverage. We also develop a test hierarchy that can perform similar tasks for higher-order estimators $\widehat{\psi}_{m,k}$ based on higher-order influence function estimators (Robins et al., 2008). In this paper, we fix the basis $\bar{\mathsf{z}}$ but one might perform multiple testing if a family of candidate basis functions is available to the analysts. It is also interesting to investigate the possibility and consequence of selecting basis functions $\bar{z}(\cdot)$ from a huge dictionary $\mathcal{V}$ of functions using data-driven methods, the goal of which is to maximize the power of the assumption-lean test.

## References


Miguel A Arcones and Evarist Gine. On the bootstrap of $U$ and $V$ statistics. *The Annals of Statistics*, 20 (2):655–674, 1992.

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

Rose Baker. An order-statistics-based method for constructing multivariate distributions with fixed marginals. *Journal of Multivariate Analysis*, 99(10):2312–2327, 2008.

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

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, Chris A J Klaassen, Ya'acov Ritov, and Jon A Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins series in the mathematical sciences. Springer New York, 1998. ISBN 9780387984735.

Jelena Bradic, Victor Chernozhukov, Whitney K Newey, and Yinchu Zhu. Minimax semiparametric learning with approximate sparsity. *arXiv preprint arXiv:1912.12213*, 2019.

Xiaohong Chen, Ying Liu, Shujie Ma, and Zheng Zhang. Efficient estimation of general treatment effects using neural networks with a diverging number of confounders. *arXiv preprint arXiv:2009.07055*, 2020.

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 Newey, and James Robins. Double/de-biased machine learning using regularized Riesz representers. *arXiv preprint arXiv:1802.08667*, 2018b.

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

Ingrid Daubechies. *Ten lectures on wavelets*, volume 61. SIAM, 1992.

Max H Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference. *Econometrica*, 2020. doi: 10.3982/ECTA16901.

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

Trevor Hastie and Robert Tibshirani. Generalized additive models. *Statistical Science*, 1(3):297–310, 1986.

Trevor Hastie and Robert Tibshirani. Generalized additive models: some applications. *Journal of the American Statistical Association*, 82(398):371–386, 1987.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948.

Marie Huskova and Paul Janssen. Consistency of the generalized bootstrap for degenerate $U$-statistics. *The Annals of Statistics*, 21(4):1811–1823, 1993.

Edward H Kennedy. Optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497*, 2020.

Edward H Kennedy, Sivaraman Balakrishnan, and Larry Wasserman. Discussion of "On nearly assumption-free tests of nominal confidence interval coverage for causal parameters estimated by machine learning". *Statistical Science*, 35(3):540–544, 2020.

Lin Liu, Rajarsh Mukherjee, James M Robins, and Eric Tchetgen Tchetgen. Adaptive estimation of nonparametric functionals. *arXiv preprint arXiv:1608.01364*, 2016.

Lin Liu, Rajarshi Mukherjee, and James M Robins. On nearly assumption-free tests of nominal confidence interval coverage for causal parameters estimated by machine learning. *Statistical Science*, 35(3):518–539, 2020a.

Lin Liu, Rajarshi Mukherjee, and James M Robins. Rejoinder: On nearly assumption-free tests of nominal confidence interval coverage for causal parameters estimated by machine learning. *Statistical Science*, 35(3):545–554, 2020b.

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

James E Mosimann. On the compound multinomial distribution, the multivariate $\beta$-distribution, and correlations among proportions. *Biometrika*, 49(1-2):65–82, 1962.

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

Justin T Newcomer, Nagaraj K Neerchal, and Jorge G Morel. Computation of higher order moments from two multinomial overdispersion likelihood models. Technical report, Department of Mathematics and Statistics, University of Maryland, Baltimore, USA, 2008.

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.

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 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, Lingling Li, Eric Tchetgen Tchetgen, and Aad W van der Vaart. Quadratic semiparametric von Mises calculus. *Metrika*, 69(2-3):227–247, 2009a.

James Robins, Eric Tchetgen Tchetgen, Lingling Li, and Aad van der Vaart. Semiparametric minimax rates. *Electronic Journal of Statistics*, 3:1305–1321, 2009b.

James Robins, Lingling Li, Eric Tchetgen Tchetgen, and Aad van der Vaart. Technical report: Higher order influence functions and minimax estimation of nonlinear functionals. *arXiv preprint arXiv:1601.05820*, 2016.

James M Robins and Ya'acov Ritov. Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in Medicine*, 16(3):285–319, 1997.

James M Robins and Andrea Rotnitzky. Comments on "Inference for semiparametric models: some questions and an answer". *Statistica Sinica*, 11(4):920–936, 2001.

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.

Andrea Rotnitzky, Ezequiel Smucler, and James M Robins. Characterization of parameters with a mixed bias property. *arXiv preprint arXiv:1904.03725*, 2019.

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.

Ezequiel Smucler, Andrea Rotnitzky, and James M Robins. A unifying approach for doubly-robust $\ell_1$ regularized estimation of causal contrasts. *arXiv preprint arXiv:1904.03737*, 2019.

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

Aad van der Vaart. Higher order tangent spaces and influence functions. *Statistical Science*, 29(4):679–686, 2014.

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

Simon N Wood, Natalya Pya, and Benjamin Säfken. Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, 111(516):1548–1563, 2016.

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. Derivations and proofs

A.1. **Derivation of the higher order influence functions of $\widetilde{\psi}_k(\theta)$ in Theorem 5.2.** In this section, we give an almost self-contained derivation of higher order influence functions (HOIFs) of $\widetilde{\psi}_k(\theta)$, except for Robins et al. (2008, Theorem 2.3), which provides the "calculus" of deriving $m$-th order influence functions from $(m-1)$-th order influence functions.

**Theorem A.1.** *Under the conditions of Theorem 2.2, the HOIFs of order $m$ of $\widetilde{\psi}_k(\theta)$ are $\mathbb{IF}_{m,k}(\Sigma_k^{-1}) = \mathbb{IF}_1 - \mathbb{IF}_{22\to mm,k}(\Sigma_k^{-1})$, where $\mathbb{IF}_{\ell\ell\to mm,k}(\Sigma_k^{-1}) \coloneqq \sum_{j=\ell}^{m} \mathbb{IF}_{jj,k}(\Sigma_k^{-1})$,*

$$\mathbb{IF}_{22,k}(\Sigma_k^{-1}) \coloneqq \frac{1}{n(n-1)} \sum_{1 \leqslant i_1 \neq i_2 \leqslant n} \mathsf{IF}_{22,k,\bar{i}_2}(\Sigma_k^{-1}) \tag{A.1}$$

*and for $j > 2$,*

$$\mathbb{IF}_{jj,k}(\Sigma_k^{-1}) \coloneqq \frac{(n-j)!}{n!} \sum_{1 \leqslant i_1 \neq \ldots \neq i_j \leqslant n} \mathsf{IF}_{jj,k,\bar{i}_j}(\Sigma_k^{-1}). \tag{A.2}$$

*Here*

$$\mathsf{IF}_{22,k,\bar{i}_2}(\Sigma_k^{-1}) \equiv \mathsf{IF}_{22,\widetilde{\psi}_k,\bar{i}_2}(\theta) = \left[\mathcal{E}_{1,\widetilde{b}_{k,\theta}}(\bar{\mathsf{z}}_k)(O)\right]_{i_1}^{\top} \Sigma_k^{-1} \left[\mathcal{E}_{2,\widetilde{p}_{k,\theta}}(\bar{\mathsf{z}}_k)(O)\right]_{i_2}, \tag{A.3}$$

*and*

$$\begin{aligned} &\mathsf{IF}_{jj,k,\bar{i}_j}(\Sigma_k^{-1}) \equiv \mathsf{IF}_{jj,\widetilde{\psi}_k,\bar{i}_j}(\theta) \\ &= (-1)^j \left[\mathcal{E}_{1,\widetilde{b}_{k,\theta}}(\bar{\mathsf{z}}_k)(O)\right]_{i_1}^{\top} \left\{\prod_{s=3}^{j} \Sigma_k^{-1} \left(\left[S_{bp}\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^{\top}\right]_{i_s} - \Sigma_k\right)\right\} \Sigma_k^{-1} \left[\mathcal{E}_{2,\widetilde{p}_{k,\theta}}(\bar{\mathsf{z}}_k)(O)\right]_{i_2}. \end{aligned} \tag{A.4}$$

*are the kernels of second order influence function $\mathbb{IF}_{22,k}(\Sigma_k^{-1})$ and $j$-th order influence function $\mathbb{IF}_{jj,k}(\Sigma_k^{-1})$ respectively, where $\bar{i}_j \coloneqq \{i_1, \ldots, i_j\}$.*

*Furthermore, define*

$$\mathbb{IF}_{22\to mm,k}(\Sigma_k^{-1}) \coloneqq \sum_{j=2}^{m} \mathbb{IF}_{jj,k}(\Sigma_k^{-1}).$$

*As a consequence, $\mathbb{IF}_{22\to mm,k}(\Sigma_k^{-1})$ is the $m$-th order influence function of $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$ under $\mathsf{P}_{\widehat{\theta}}$.*

*Proof.* Without loss of generality, we assume $\mathsf{P}_\theta(S_{bp} \geqslant 0) = 1$. First, let's recall the results from Theorem 2.1 (or Rotnitzky et al. (2019, Theorems 2 (ii))), the influence function of $\psi(\theta)$ has the following form

$$\mathsf{IF}_{1,\psi(\theta)} = S_{bp} b(X) p(X) + m_1(O, b) + m_2(O, p) + S_0 - \psi(\theta),$$

where the linear functionals $m_1(O, h)$ and $m_2(O, h)$ have the Riesz representers $\mathcal{R}_1(X)$ and $\mathcal{R}_2(X)$ respectively. By definition of Riesz representer, they must obey the following identities:

$$\begin{cases} \mathsf{E}_\theta\left[m_1(O, h)\right] = \mathsf{E}_\theta\left[h(X)\mathcal{R}_1(X)\right] \\ \mathsf{E}_\theta\left[m_2(O, h)\right] = \mathsf{E}_\theta\left[h(X)\mathcal{R}_2(X)\right] \end{cases}$$

for any $h \in L_2(g)$.

To derive the HOIFs for $\widetilde{\psi}_k(\theta)$, we need its influence function. Recall from Definition 3.1 that $\widetilde{\psi}_k(\theta)$ is of the following form:

$$
\begin{aligned}
&\widetilde{\psi}_k(\theta) \\
&\equiv \mathsf{E}_\theta\left[\mathcal{H}\left(\widetilde{b}_{k,\theta}, \widetilde{p}_{k,\theta}\right)\right] \\
&= \mathsf{E}_\theta\left[S_{bp}\widetilde{b}_{k,\theta}(X)\widetilde{p}_{k,\theta}(X) + m_1(O, \widetilde{b}_{k,\theta}) + m_2(O, \widetilde{p}_{k,\theta}) + S_0\right] \\
&= \mathsf{E}_\theta\left[\begin{array}{c} S_{bp}\left(\widehat{b}(X) + \widetilde{\widetilde{\zeta}}_{b,k}^\top(\theta)\bar{\mathsf{z}}_k(X)\right)\left(\widehat{p}(X) + \widetilde{\widetilde{\zeta}}_{p,k}^\top(\theta)\bar{\mathsf{z}}_k(X)^\top\right) \\ + m_1(O, \widehat{b} + \widetilde{\widetilde{\zeta}}_{b,k}^\top(\theta)\bar{\mathsf{z}}_k) + m_2(O, \widehat{p} + \widetilde{\widetilde{\zeta}}_{p,k}^\top(\theta)\bar{\mathsf{z}}_k) + S_0 \end{array}\right] \\
&= \mathsf{E}_\theta\left[\begin{array}{c} S_{bp}\left(\widehat{b}(X) - \Sigma_k^{-1}\mathsf{E}_\theta\left[S_{bp}\widehat{b}(X)\bar{\mathsf{z}}_k(X)^\top + m_2(O, \bar{\mathsf{z}}_k(X))^\top\right]\bar{\mathsf{z}}_k(X)\right) \\ \times\left(\widehat{p}(X) - \Sigma_k^{-1}\mathsf{E}_\theta\left[S_{bp}\widehat{p}(X)\bar{\mathsf{z}}_k(X)^\top + m_1(O, \bar{\mathsf{z}}_k(X))^\top\right]\bar{\mathsf{z}}_k(X)\right) \\ + m_1(O, \widehat{b} - \Sigma_k^{-1}\mathsf{E}_\theta\left[S_{bp}\widehat{b}(X)\bar{\mathsf{z}}_k(X)^\top + m_2(O, \bar{\mathsf{z}}_k(X))^\top\right]\bar{\mathsf{z}}_k) \\ + m_2(O, \widehat{p} - \Sigma_k^{-1}\mathsf{E}_\theta\left[S_{bp}\widehat{p}(X)\bar{\mathsf{z}}_k(X)^\top + m_1(O, \bar{\mathsf{z}}_k(X))^\top\right]\bar{\mathsf{z}}_k) + S_0 \end{array}\right].
\end{aligned}
$$

Then

$$
\begin{aligned}
\mathsf{IF}_{1,\widetilde{\psi}_k(\theta),i}(\theta) &= \mathcal{H}(\widetilde{b}_k(X,\theta), \widetilde{p}_k(X,\theta)) - \widetilde{\psi}_k(\theta) \\
&\quad + \mathsf{E}_\theta\left[\partial\mathcal{H}\left(b_k^*(X, \widetilde{\widetilde{\zeta}}_{b,k}(\theta)), p_k^*(X, \widetilde{\widetilde{\zeta}}_{b,k}(\theta))\right)/\partial\bar{\zeta}_{b,k}^\top\right]\mathsf{IF}_{1,\widetilde{\widetilde{\zeta}}_{b,k}}(\theta) \\
&\quad + \mathsf{E}_\theta\left[\partial\mathcal{H}\left(b_k^*(X, \widetilde{\widetilde{\zeta}}_{b,k}(\theta)), p_k^*(X, \widetilde{\widetilde{\zeta}}_{b,k}(\theta))\right)/\partial\bar{\zeta}_{p,k}^\top\right]\mathsf{IF}_{1,\widetilde{\widetilde{\zeta}}_{p,k}}(\theta) \\
&= \mathcal{H}(\widetilde{b}_{k,\theta}, \widetilde{p}_{k,\theta}) - \widetilde{\psi}_k(\theta) \\
&= S_{bp,i}\widetilde{b}_{k,\theta}(X_i)\widetilde{p}_{k,\theta}(X_i) + m_1(O_i, \widetilde{b}_{k,\theta}) + m_2(O_i, \widetilde{p}_{k,\theta}) + S_{0,i} - \widetilde{\psi}_k(\theta).
\end{aligned}
$$

where the second equality follows from the definitions of $\widetilde{\widetilde{\zeta}}_{b,k}(\theta)$ and $\widetilde{\widetilde{\zeta}}_{p,k}(\theta)$ in equation (3.1). Then the influence function of $-\mathsf{IF}_{1,\widetilde{\psi}_k(\theta),i}(\theta)$, denoted as $\mathsf{IF}_{1,-\mathsf{if}_{1,\widetilde{\psi}_k(\theta),i_1},i_2}(\theta)$, can be calculated as follows:

$$
\mathsf{IF}_{1,-\mathsf{if}_{1,\widetilde{\psi}_k(\theta),i_1},i_2}(\theta) = -\frac{\partial\mathcal{H}_{i_1}\left(\widetilde{b}_{k,\theta}(X_{i_1}), \widetilde{p}_{k,\theta}(X_{i_1})\right)}{\partial\bar{\zeta}_{b,k}^\top}\mathsf{IF}_{1,\widetilde{\widetilde{\zeta}}_{b,k},i_2}(\theta) - \frac{\partial\mathcal{H}_{i_1}\left(\widetilde{b}_{k,\theta}(X_{i_1}), \widetilde{p}_{k,\theta}(X_{i_1})\right)}{\partial\bar{\zeta}_{p,k}^\top}\mathsf{IF}_{1,\widetilde{\widetilde{\zeta}}_{p,k},i_2}(\theta)
$$

where

$$
\left\{\begin{array}{l} \frac{\partial\mathcal{H}_{i_1}\left(\widetilde{b}_{k,\theta}(X_{i_1}), \widetilde{p}_{k,\theta}(X_{i_1})\right)}{\partial\bar{\zeta}_{b,k}} = S_{bp,i_1}\widetilde{p}_{k,\theta}(X_{i_1})\bar{\mathsf{z}}_k(X_{i_1}) + m_1(O_{i_1}, \bar{\mathsf{z}}_k), \\ \frac{\partial\mathcal{H}_{i_1}\left(\widetilde{b}_{k,\theta}(X_{i_1}), \widetilde{p}_{k,\theta}(X_{i_1})\right)}{\partial\bar{\zeta}_{p,k}} = S_{bp,i_1}\widetilde{b}_{k,\theta}(X_{i_1})\bar{\mathsf{z}}_k(X_{i_1}) + m_2(O_{i_1}, \bar{\mathsf{z}}_k) \end{array}\right.
$$

and by Lemma A.1

$$
\left\{\begin{array}{l} \mathsf{IF}_{1,\widetilde{\widetilde{\zeta}}_{b,k},i_2}(\theta) = -\Sigma_k^{-1}\left(S_{bp,i_2}\widetilde{b}_{k,\theta}(X_{i_2})\bar{\mathsf{z}}_k(X_{i_2}) + m_2(O_{i_2}, \bar{\mathsf{z}}_k)\right), \\ \mathsf{IF}_{1,\widetilde{\widetilde{\zeta}}_{p,k},i_2}(\theta) = -\Sigma_k^{-1}\left(S_{bp,i_2}\widetilde{p}_{k,\theta}(X_{i_2})\bar{\mathsf{z}}_k(X_{i_2}) + m_1(O_{i_2}, \bar{\mathsf{z}}_k)\right). \end{array}\right.
$$

Then

$$
\begin{aligned}
&\mathsf{IF}_{1,-\mathsf{if}_{1,\widetilde{\psi}_k(\theta),i_1},i_2}(\theta) \\
=& \left[S_{bp}\widetilde{p}_{k,\theta}(X)\bar{\mathsf{z}}_k(X) + m_1(O,\bar{\mathsf{z}}_k)\right]_{i_1}^{\top} \Sigma_k^{-1} \left[S_{bp}\widetilde{b}_{k,\theta}(X)\bar{\mathsf{z}}_k(X) + m_2(O,\bar{\mathsf{z}}_k)\right]_{i_2} \\
&+ \left[S_{bp}\widetilde{b}_{k,\theta}(X)\bar{\mathsf{z}}_k(X) + m_2(O,\bar{\mathsf{z}}_k)\right]_{i_1}^{\top} \Sigma_k^{-1} \left[S_{bp}\widetilde{p}_{k,\theta}(X)\bar{\mathsf{z}}_k(X) + m_1(O,\bar{\mathsf{z}}_k)\right]_{i_2}.
\end{aligned}
$$

Then by Robins et al. (2008, part 5.c of Theorem 2.3),

$$
\mathbb{IF}_{22,k}(\Sigma_k^{-1}) = \frac{1}{2}\Pi_\theta\left[\mathbb{V}_2\left[\mathsf{IF}_{1,-\mathsf{if}_{1,\widetilde{\psi}_k(\theta),i_1},i_2}(\theta)\right] | \mathcal{U}_1^{\perp,2}(\theta)\right].
$$

Here $\mathcal{U}_m(\theta)$ denotes the Hilbert space of all $m$-th order U-statistics with mean zero and finite variance with inner product defined by covariances w.r.t. the $n$-fold product measure $\mathsf{P}_\theta(\cdot)^n$ and $\mathcal{U}_m^{\perp,m+1}(\theta)$ denotes the orthocomplement of $\mathcal{U}_m(\theta)$ in $\mathcal{U}_{m+1}(\theta)$. The notation $\mathbb{V}_m[U_{i_1,\ldots,i_m}]$ maps an $m$-th order U-statistic kernel $U_{i_1,\ldots,i_m}$ to an $m$-th order U-statistic:

$$
\mathbb{V}_m[U_{i_1,\ldots,i_m}] = \frac{(n-m)!}{n!}\sum_{1\leqslant i_1\neq\cdots\neq i_m\leqslant n} U_{i_1,\ldots,i_m}.
$$

For the 2nd order U-statistic kernel $\mathsf{IF}_{1,\mathsf{if}_{1,\widetilde{\psi}_k(\theta),i_1},i_2}(\theta)$, we thus have

$$
\begin{aligned}
&\mathbb{V}_2\left[\mathsf{IF}_{1,-\mathsf{if}_{1,\widetilde{\psi}_k(\theta),i_1},i_2}(\theta)\right] \\
=& \mathbb{V}_2\left[\begin{array}{c}\left[S_{bp}\widetilde{p}_{k,\theta}(X)\bar{\mathsf{z}}_k(X) + m_1(O,\bar{\mathsf{z}}_k)\right]_{i_1}^{\top} \Sigma_k^{-1} \left[S_{bp}\widetilde{b}_{k,\theta}(X)\bar{\mathsf{z}}_k(X) + m_2(O,\bar{\mathsf{z}}_k)\right]_{i_2} \\ + \left[S_{bp}\widetilde{b}_{k,\theta}(X)\bar{\mathsf{z}}_k(X) + m_2(O,\bar{\mathsf{z}}_k)\right]_{i_1}^{\top} \Sigma_k^{-1} \left[S_{bp}\widetilde{p}_{k,\theta}(X)\bar{\mathsf{z}}_k(X) + m_1(O,\bar{\mathsf{z}}_k)\right]_{i_2}\end{array}\right] \\
=& \mathbb{V}_2\left[2\left[S_{bp}\widetilde{p}_{k,\theta}(X)\bar{\mathsf{z}}_k(X) + m_1(O,\bar{\mathsf{z}}_k)\right]_{i_1}^{\top} \Sigma_k^{-1} \left[S_{bp}\widetilde{b}_{k,\theta}(X)\bar{\mathsf{z}}_k(X) + m_2(O,\bar{\mathsf{z}}_k)\right]_{i_2}\right]
\end{aligned}
$$

where the last equality is due to the symmetrization of $\mathbb{V}_2[\cdot]$. In addition, we have

$$
\Pi_\theta\left[\mathbb{V}_2\left[\mathsf{IF}_{1,-\mathsf{if}_{1,\widetilde{\psi}_k(\theta),i_1},i_2}(\theta)\right] | \mathcal{U}_1(\theta)\right] = 0
$$

since

$$
\begin{cases}
\mathsf{E}_\theta\left[S_{bp}\widetilde{p}_{k,\theta}(X)\bar{\mathsf{z}}_k(X) + m_1(O,\bar{\mathsf{z}}_k)\right] = 0, \\
\mathsf{E}_\theta\left[S_{bp}\widetilde{b}_{k,\theta}(X)\bar{\mathsf{z}}_k(X) + m_2(O,\bar{\mathsf{z}}_k)\right] = 0.
\end{cases}
$$

Thus

$$
\begin{aligned}
&\mathbb{IF}_{22,k}(\Sigma_k^{-1}) = \frac{1}{2}\mathbb{V}_2\left[\mathsf{IF}_{1,-\mathsf{if}_{1,\widetilde{\psi}_k(\theta),i_1},i_2}(\theta)\right] \\
&= \frac{1}{n(n-1)}\sum_{1\leqslant i_1\neq i_2\leqslant n}\left[S_{bp}\widetilde{p}_{k,\theta}(X)\bar{\mathsf{z}}_k(X) + m_1(O,\bar{\mathsf{z}}_k)\right]_{i_1}^{\top} \Sigma_k^{-1} \left[S_{bp}\widetilde{b}_{k,\theta}(X)\bar{\mathsf{z}}_k(X) + m_2(O,\bar{\mathsf{z}}_k)\right]_{i_2}.
\end{aligned}
$$

Hence formally we have

$$
\mathsf{IF}_{22,k,\bar{i}_2}(\Sigma_k^{-1}) = \left[S_{bp}\widetilde{p}_{k,\theta}(X)\bar{\mathsf{z}}_k(X) + m_1(O,\bar{\mathsf{z}}_k)\right]_{i_1}^{\top} \Sigma_k^{-1} \left[S_{bp}\widetilde{b}_{k,\theta}(X)\bar{\mathsf{z}}_k(X) + m_2(O,\bar{\mathsf{z}}_k)\right]_{i_2}
$$

and

$$\widehat{\mathsf{IF}}_{22,k,\bar{i}_2}(\Sigma_k^{-1}) = [S_{bp}\widehat{p}(X)\bar{\mathsf{z}}_k(X) + m_1(O,\bar{\mathsf{z}}_k)]_{i_1}^\top \Sigma_k^{-1} \left[S_{bp}\widehat{b}(X)\bar{\mathsf{z}}_k(X) + m_2(O,\bar{\mathsf{z}}_k)\right]_{i_2}.$$

We now complete the proof by induction. We assume that the form for $\widehat{\mathsf{IF}}_{jj,k,\bar{i}_j}(\Sigma_k^{-1})$ is true and thus by the induction hypothesis

$$\begin{aligned}&\mathsf{IF}_{jj,k,\bar{i}_j}(\Sigma_k^{-1})\\ &= (-1)^j \left[S_{bp}\widetilde{p}_{k,\theta}(X)\bar{\mathsf{z}}_k(X) + m_1(O,\bar{\mathsf{z}}_k)\right]_{i_1}^\top \prod_{s=3}^{j} \Sigma_k^{-1}\left(\left[S_{bp}\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top\right]_{i_s} - \Sigma_k\right)\\ &\quad\times\ \Sigma_k^{-1}\left[S_{bp}\widetilde{b}_{k,\theta}(X)\bar{\mathsf{z}}_k(X) + m_2(O,\bar{\mathsf{z}}_k)\right]_{i_2}\end{aligned}$$

and

$$\mathbb{IF}_{jj,k}(\Sigma_k^{-1}) = \mathbb{V}_j\left[\mathsf{IF}_{jj,k,\bar{i}_j}(\Sigma_k^{-1})\right].$$

Then

$$\begin{aligned}&\mathbb{V}_{j+1}\left[\mathsf{IF}_{(j+1),(j+1),k,\bar{i}_{j+1}}(\Sigma_k^{-1})\right]\\ &= \frac{1}{j+1}\mathbb{V}_{j+1}\left[\Pi_\theta\left[\mathsf{IF}_{1,\mathsf{if}_{jj,k,\bar{i}_j},i_{j+1}}(\theta)|\mathcal{U}_j^{\perp,j+1}(\theta)\right]\right].\end{aligned}$$

Then the derivatives w.r.t. $\theta$'s in $\widetilde{\widetilde{\zeta}}_{b,k}(\theta)$, $\widetilde{\widetilde{\zeta}}_{p,k}(\theta)$ and $(j-1)$ terms of $\Sigma_k^{-1}$ will be contributing terms to $\mathbb{V}_{j+1}\left[\mathsf{IF}_{(j+1),(j+1),k,\bar{i}_{j+1}}(\Sigma_k^{-1})\right]$. But differentiating w.r.t. $(j-2)$ terms of $\Sigma_k$ will not contribute terms to $\mathbb{V}_{j+1}\left[\mathsf{IF}_{(j+1),(j+1),k,\bar{i}_{j+1}}(\Sigma_k^{-1})\right]$ as the contribution of these $(j-2)$ terms to $\mathsf{IF}_{1,\mathsf{if}_{jj,k,\bar{i}_j},i_{j+1}}(\theta)$ is a function of $j$ units' data and is thus an element of $\mathcal{U}_j(\theta)$. Then as in the proof of Lemma A.1,

$$\mathsf{IF}_{1,\Sigma_k^{-1}}(\theta) = -\Sigma_k^{-1}\left(S_{bp}\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top - \Sigma_k\right)\Sigma_k^{-1}$$

and upon permuting the indices, the contribution of each of these $(j-1)$ terms to $\mathsf{IF}_{1,\mathsf{if}_{jj,k,\bar{i}_j},i_{j+1}}(\theta)$ is

$$\begin{aligned}&-(-1)^j\left[S_{bp}\widetilde{p}_{k,\theta}(X)\bar{\mathsf{z}}_k(X) + m_1(O,\bar{\mathsf{z}}_k)\right]_{i_1}^\top\\ &\times\sum_{s=3}^{j+1}\Sigma_k^{-1}\left\{\left(S_{bp}\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top\right)_{i_s} - \Sigma_k\right\}\\ &\times\Sigma_k^{-1}\left[S_{bp}\widetilde{b}_{k,\theta}(X)\bar{\mathsf{z}}_k(X) + m_2(O,\bar{\mathsf{z}}_k)\right]_{i_2}\end{aligned}\tag{A.5}$$

which is degenerate and thus orthogonal to $\mathcal{U}_j(\theta)$. Notice that equation (A.5) is exactly the kernel of $\mathbb{IF}_{(j+1),(j+1),k}(\Sigma_k^{-1})$. Then differentiating w.r.t. $\theta$'s in $\widetilde{p}_{k,\theta}$ and $\widetilde{b}_{k,\theta}$, we further obtain terms of the following form contributing to

$\mathsf{IF}_{1,\mathsf{if}_{jj,k,\bar{i}_j},i_{j+1}}(\theta)$:

$$
\begin{aligned}
&(-1)^j \mathsf{IF}_{1,\widetilde{\widetilde{\zeta}}_{p,k},i_{j+1}}(\theta)^\top \left[S_{bp}\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top\right]_{i_1} \left[\sum_{s=3}^{j} \Sigma_k^{-1} \left\{\left(S_{bp}\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top\right)_{i_s} - \Sigma_k\right\}\right] \\
&\times \Sigma_k^{-1} \left[S_{bp}\widetilde{b}_{k,\theta}(X)\bar{\mathsf{z}}_k(X) + m_2(O,\bar{\mathsf{z}}_k)\right]_{i_2} \\
&+ (-1)^j \left[S_{bp}\widetilde{p}_{k,\theta}(X)\bar{\mathsf{z}}_k(X) + m_1(O,\bar{\mathsf{z}}_k)\right]_{i_1}^\top \left[\sum_{s=3}^{j} \Sigma_k^{-1} \left\{\left(S_{bp}\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top\right)_{i_s} - \Sigma_k\right\}\right] \\
&\times \Sigma_k^{-1} \left[S_{bp}\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top\right]_{i_2} \mathsf{IF}_{1,\widetilde{\widetilde{\zeta}}_{b,k},i_{j+1}}(\theta).
\end{aligned} \tag{A.6}
$$

Again applying Lemma A.1 and projecting onto $\mathcal{U}_{j,\theta}^{\perp,j+1}$, we obtain that

$$
\Pi_\theta\left[\text{equation (A.6)}|\mathcal{U}_{j,\theta}^{\perp,j+1}\right] \equiv 2 \times \text{equation (A.5)}.
$$

Recall that there are $(j-1)$ terms of equation (A.5) in total contributing to $\mathsf{IF}_{1,\mathsf{if}_{jj,k,\bar{i}_j},i_{j+1}}(\theta)$, thus eventually there are $(j+1)$ terms of equation (A.5) in total contributing to $\mathbb{V}_{j+1}\left[\Pi_\theta\left[\mathsf{IF}_{1,\mathsf{if}_{jj,k,\bar{i}_j},i_{j+1}}(\theta)|\mathcal{U}_j^{\perp,j+1}(\theta)\right]\right]$. Further recall that

$$
\begin{aligned}
&\mathbb{V}_{j+1}\left[\mathsf{IF}_{(j+1),(j+1),k,\bar{i}_{j+1}}(\Sigma_k^{-1})\right] \\
&= \frac{1}{j+1}\mathbb{V}_{j+1}\left[\Pi_\theta\left[\mathsf{IF}_{1,\mathsf{if}_{jj,k,\bar{i}_j},i_{j+1}}(\theta)|\mathcal{U}_j^{\perp,j+1}(\theta)\right]\right].
\end{aligned}
$$

Therefore we have

$$
\begin{aligned}
\mathbb{IF}_{(j+1),(j+1),k}(\Sigma_k^{-1}) &\equiv \mathbb{V}_{j+1}\left[\mathsf{IF}_{(j+1),(j+1),k,\bar{i}_{j+1}}(\Sigma_k^{-1})\right] \\
&= \frac{1}{j+1}\mathbb{V}_{j+1}\left[(j+1)\times \text{equation (A.5)}\right] \\
&= \mathbb{V}_{j+1}\left[\text{equation (A.5)}\right].
\end{aligned}
$$

We thus prove the results for general $j$ by induction. ∎

**Lemma A.1.** *As defined in equation (3.1), the influence functions of* $\widetilde{\widetilde{\zeta}}_{b,k}(\theta)$ *and* $\widetilde{\widetilde{\zeta}}_{p,k}(\theta)$ *are*

$$
\begin{cases}
\mathsf{IF}_{1,\widetilde{\widetilde{\zeta}}_{b,k}}(\theta) = -\Sigma_k^{-1}\left(S_{bp}\widetilde{b}_{k,\theta}(X)\bar{\mathsf{z}}_k(X) + m_2(O,\bar{\mathsf{z}}_k)\right), \\
\mathsf{IF}_{1,\widetilde{\widetilde{\zeta}}_{p,k}}(\theta) = -\Sigma_k^{-1}\left(S_{bp}\widetilde{p}_{k,\theta}(X)\bar{\mathsf{z}}_k(X) + m_1(O,\bar{\mathsf{z}}_k)\right).
\end{cases} \tag{A.7}
$$

*Proof.* We only need to show $\widehat{\mathsf{IF}}_{1,\widetilde{\widetilde{\zeta}}_{b,k}}(\theta)$ and $\widehat{\mathsf{IF}}_{1,\widetilde{\widetilde{\zeta}}_{p,k}}(\theta)$ will follow by symmetry. Recall that

$$
\widetilde{\widetilde{\zeta}}_{b,k}(\theta) = -\left\{\mathsf{E}_\theta\left[S_{bp}\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top\right]\right\}^{-1}\mathsf{E}_\theta\left[S_{bp}\widehat{b}(X)\bar{\mathsf{z}}_k(X) + m_2(O,\bar{\mathsf{z}}_k)\right].
$$

Then

$$
\begin{aligned}
&\mathsf{IF}_{1,\widetilde{\widetilde{\zeta}}_{b,k}}(\theta) \\
&= -\left\{\mathsf{E}_\theta\left[S_{bp}\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top\right]\right\}^{-1}\mathsf{IF}_{1,\mathsf{E}.\left[S_{bp}\widehat{b}(X)\bar{\mathsf{z}}_k(X)+m_2(O,\bar{\mathsf{z}}_k)\right]}(\theta)
\end{aligned}
$$

$$
\begin{aligned}
&- \mathsf{IF}_{1,\left\{\mathsf{E}_{\cdot}\left[S_{bp}\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top\right]\right\}^{-1}}(\theta)\mathsf{E}_\theta\left[S_{bp}\widehat{b}(X)\bar{\mathsf{z}}_k(X)+m_2(O,\bar{\mathsf{z}}_k)\right]\\
=&-\Sigma_k^{-1}\left(S_{bp}\widehat{b}(X)\bar{\mathsf{z}}_k(X)+m_2(O,\bar{\mathsf{z}}_k)-\mathsf{E}_\theta\left[S_{bp}\widehat{b}(X)\bar{\mathsf{z}}_k(X)+m_2(O,\bar{\mathsf{z}}_k)\right]\right)\\
&+\Sigma_k^{-1}\left(S_{bp}\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top-\Sigma_k\right)\Sigma_k^{-1}\mathsf{E}_\theta\left[S_{bp}\widehat{b}(X)\bar{\mathsf{z}}_k(X)+m_2(O,\bar{\mathsf{z}}_k)\right]\\
=&-\Sigma_k^{-1}\left(S_{bp}\widehat{b}(X)\bar{\mathsf{z}}_k(X)+m_2(O,\bar{\mathsf{z}}_k)-S_{bp}\bar{\mathsf{z}}_k(X)^\top\Sigma_k^{-1}\mathsf{E}_\theta\left[S_{bp}\widehat{b}(X)\bar{\mathsf{z}}_k(X)+m_2(O,\bar{\mathsf{z}}_k)\right]\bar{\mathsf{z}}_k(X)\right)\\
=&-\Sigma_k^{-1}\left(S_{bp}\widehat{b}(X)\bar{\mathsf{z}}_k(X)+m_2(O,\bar{\mathsf{z}}_k)-S_{bp}\bar{\mathsf{z}}_k(X)^\top\widetilde{\widetilde{\zeta}}_{b,k}(\theta)\bar{\mathsf{z}}_k(X)\right)\\
=&-\Sigma_k^{-1}\left\{S_{bp}\left(\widehat{b}(X)-\bar{\mathsf{z}}_k(X)^\top\widetilde{\widetilde{\zeta}}_{b,k}(\theta)\right)\bar{\mathsf{z}}_k(X)+m_2(O,\bar{\mathsf{z}}_k)\right\}\\
=&-\Sigma_k^{-1}\left(S_{bp}\widetilde{b}_{k,\theta}(X)\bar{\mathsf{z}}_k(X)+m_2(O,\bar{\mathsf{z}}_k)\right)
\end{aligned}
$$

where the fourth equality follows from the definition of $\widetilde{\widetilde{\zeta}}_{b,k}(\theta)$ and the sixth equality follows from the definition of $\widetilde{b}_{k,\theta}(X)$. ∎

Under conditions in Theorem 2.2, $\widetilde{b}_{k,\widehat{\theta}}=\widehat{b}$ and $\widetilde{p}_{k,\widehat{\theta}}=\widehat{p}$ or equivalently $\widetilde{\widetilde{\zeta}}_{b,k}(\widehat{\theta})=\widetilde{\widetilde{\zeta}}_{p,k}(\widehat{\theta})=0$. This follows from results in Theorem 2.1 with $\theta$ evaluated at $\widehat{\theta}$:

$$\mathsf{E}_{\widehat{\theta}}\left[S_{bp}\widehat{b}(X)h(X)+m_2(O,h)\right]=0 \text{ for all } h\in\mathcal{B}, \mathsf{E}_{\widehat{\theta}}\left[S_{bp}\widehat{p}(X)h(X)+m_1(O,h)\right]=0 \text{ for all } h\in\mathcal{P}.$$

Thus under $\widehat{\theta}$, the HOIFs of $\widetilde{\psi}_k(\widehat{\theta})$ are: $\widehat{\mathbb{IF}}_{m,k}(\Sigma_k^{-1})=\widehat{\mathbb{IF}}_1-\widehat{\mathbb{IF}}_{22\to mm,k}(\Sigma_k^{-1})$, where $\widehat{\mathbb{IF}}_{\ell\ell\to mm,k}(\Sigma_k^{-1}):=\sum_{j=\ell}^m\widehat{\mathbb{IF}}_{jj,k}(\Sigma_k^{-1})$,

$$\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1}):=\frac{1}{n(n-1)}\sum_{1\leqslant i_1\neq i_2\leqslant n}\widehat{\mathsf{IF}}_{22,k,\bar{i}_2}(\Sigma_k^{-1}) \tag{A.8}$$

and for $j>2$,

$$\widehat{\mathbb{IF}}_{jj,k}(\Sigma_k^{-1}):=\frac{(n-j)!}{n!}\sum_{1\leqslant i_1\neq\ldots\neq i_j\leqslant n}\widehat{\mathsf{IF}}_{jj,k,\bar{i}_j}(\Sigma_k^{-1}). \tag{A.9}$$

Here

$$\widehat{\mathsf{IF}}_{22,k,\bar{i}_2}(\Sigma_k^{-1})\equiv\mathsf{IF}_{22,\widetilde{\psi}_k,\bar{i}_2}(\widehat{\theta})=\left[\mathcal{E}_{\widehat{b},m_2}(\bar{\mathsf{z}}_k)(O)\right]_{i_1}^\top\Sigma_k^{-1}\left[\mathcal{E}_{\widehat{p},m_1}(\bar{\mathsf{z}}_k)(O)\right]_{i_2}, \tag{A.10}$$

and

$$
\begin{aligned}
&\widehat{\mathsf{IF}}_{jj,k,\bar{i}_j}(\Sigma_k^{-1})\equiv\mathsf{IF}_{jj,\widetilde{\psi}_k,\bar{i}_j}(\widehat{\theta})\\
&=(-1)^j\left[\mathcal{E}_{\widehat{b},m_2}(\bar{\mathsf{z}}_k)(O)\right]_{i_1}^\top\left\{\prod_{s=3}^j\Sigma_k^{-1}\left(\left[S_{bp}\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top\right]_{i_s}-\Sigma_k\right)\right\}\Sigma_k^{-1}\left[\mathcal{E}_{\widehat{p},m_1}(\bar{\mathsf{z}}_k)(O)\right]_{i_2}.
\end{aligned} \tag{A.11}
$$

are the kernels of second order influence function $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ and $j$-th order influence function $\widehat{\mathbb{IF}}_{jj,k}(\Sigma_k^{-1})$ respectively, where $\mathcal{E}_{\widehat{b},m_2}$ and $\mathcal{E}_{\widehat{p},m_1}$ are defined in equations (3.4) and (3.5).

As a consequence of Theorem A.1, the $m$-th order influence function for the truncated parameter $\widetilde{\psi}_k(\theta)$ under $\mathsf{P}_{\widehat{\theta}}$ is simply $\widehat{\psi}_{m,k}(\Sigma_k^{-1})-\widetilde{\psi}_k(\theta)\equiv\widehat{\mathbb{IF}}_{m,k}(\Sigma_k^{-1})$ and the $m$-th order influence function for the bias $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$ is $\widehat{\mathbb{IF}}_{22\to mm,k}(\Sigma_k^{-1})$, with the $b$ and $p$ in $\mathbb{IF}_{m,k}(\Sigma_k^{-1})$ and $\mathbb{IF}_{22\to mm,k}(\Sigma_k^{-1})$ replaced by $\widehat{b}$ and $\widehat{p}$.

A.2. **General formula for the variance order of $\widehat{\mathbb{IF}}_{mm,k}(\widehat{\Sigma}_k^{-1})$.** In this section we derive the following formula for the order of $\mathsf{var}_\theta\left[\widehat{\mathbb{IF}}_{mm,k}(\widehat{\Sigma}_k^{-1})\right]$ for general $m$. Then we have

**Lemma A.2.** *When $m \geqslant 3$,*

(A.12)

$$
\begin{aligned}
\mathsf{var}_\theta\left[\widehat{\mathbb{IF}}_{mm,k}(\widehat{\Sigma}_k^{-1})\right] \overset{\textit{Condition } W}{\lesssim} \frac{1}{n} & \left\{ \begin{array}{c} \left(\frac{k}{n}\right)^{m-1} + \sum_{j=2}^{m-1} \binom{m}{j}^2 \left(\frac{k}{n}\right)^{j-1} \mathbb{L}_{\theta,2,\widehat{\Sigma},k}^{2(m-j-2)} \left\{ \mathbb{L}_{\theta,2,\widehat{\Sigma},k}^4 + \mathbb{L}_{\theta,2,\widehat{b},k}^2 \mathbb{L}_{\theta,2,\widehat{\Sigma},k}^2 + \mathbb{L}_{\theta,2,\widehat{p},k}^2 \mathbb{L}_{\theta,2,\widehat{\Sigma},k}^2 + k \mathbb{L}_{\theta,2,\widehat{b},k}^2 \mathbb{L}_{\theta,2,\widehat{p},k}^2 \right\} \\ + m^2 \mathbb{L}_{\theta,2,\widehat{\Sigma},k}^{2(m-3)} \left\{ \mathbb{L}_{\theta,2,\widehat{b},k}^2 \mathbb{L}_{\theta,2,\widehat{\Sigma},k}^2 + \mathbb{L}_{\theta,2,\widehat{p},k}^2 \mathbb{L}_{\theta,2,\widehat{\Sigma},k}^2 + k \mathbb{L}_{\theta,2,\widehat{b},k}^2 \mathbb{L}_{\theta,2,\widehat{p},k}^2 \right\} \end{array} \right\} \\
\overset{\textit{Condition } SW}{\lesssim} \frac{1}{n} & \left\{ \begin{array}{c} \left(\frac{k}{n}\right)^{m-1} + \sum_{j=2}^{m-1} \binom{m}{j}^2 \left(\frac{k}{n}\right)^{j-1} \mathbb{L}_{\theta,2,\widehat{\Sigma},k}^{2(m-j-2)} \left\{ \begin{array}{c} \mathbb{L}_{\theta,2,\widehat{\Sigma},k}^4 + \mathbb{L}_{\theta,2,\widehat{b},k}^2 \mathbb{L}_{\theta,2,\widehat{\Sigma},k}^2 + \mathbb{L}_{\theta,2,\widehat{p},k}^2 \mathbb{L}_{\theta,2,\widehat{\Sigma},k}^2 \\ + \left\{ \mathbb{L}_{\theta,2,\widehat{b},k}^2 \mathbb{L}_{\theta,\infty,\widehat{p},k}^2 \wedge \mathbb{L}_{\theta,2,\widehat{p},k}^2 \mathbb{L}_{\theta,\infty,\widehat{b},k}^2 \right\} \end{array} \right\} \\ + m^2 \mathbb{L}_{\theta,2,\widehat{\Sigma},k}^{2(m-3)} \left\{ \mathbb{L}_{\theta,2,\widehat{b},k}^2 \mathbb{L}_{\theta,2,\widehat{\Sigma},k}^2 + \mathbb{L}_{\theta,2,\widehat{p},k}^2 \mathbb{L}_{\theta,2,\widehat{\Sigma},k}^2 + \left\{ \mathbb{L}_{\theta,2,\widehat{b},k}^2 \mathbb{L}_{\theta,\infty,\widehat{p},k}^2 \wedge \mathbb{L}_{\theta,2,\widehat{p},k}^2 \mathbb{L}_{\theta,\infty,\widehat{b},k}^2 \right\} \right\} \end{array} \right\}.
\end{aligned}
$$

*Proof.* Equation A.12 follows from Hoeffding decomposition (Hoeffding, 1948) of U-statistics $\widehat{\mathbb{IF}}_{mm,k}(\widehat{\Sigma}_k^{-1}) - \mathsf{E}_\theta\left[\widehat{\mathbb{IF}}_{mm,k}(\widehat{\Sigma}_k^{-1})\right]$. We simply bound each term in the decomposition. The detailed calculations can be found in the forthcoming updated version of Mukherjee et al. (2017). ∎

A.3. **Bootstrapping higher order influence functions.** In this section, we will use higher order moments of multinomial distributions frequently. Two good references are Mosimann (1962); Newcomer et al. (2008). For notational convenience, we suppress the dependence on $\Sigma_k^{-1}$ or $\widehat{\Sigma}_k^{-1}$ in $\widehat{\mathbb{IF}}_{mm,k}(\Sigma_k^{-1})$ and $\widehat{\mathbb{IF}}_{mm,k}(\widehat{\Sigma}_k^{-1})$ as the proposed bootstrap resampling procedure will not depend on $\Sigma_k^{-1}$ or the training sample. We propose the following bootstrap estimator of $\mathsf{var}_\theta[\widehat{\mathbb{IF}}_{22,k}]$: choose some large integer $M$ as the number of bootstrap resamples, and define

$$
\widehat{\mathsf{var}}[\widehat{\mathbb{IF}}_{22,k}] := \underbrace{\frac{1}{M-1} \sum_{m=1}^{M} \left( \widehat{\mathbb{IF}}_{22,k}^{(m)} - \frac{1}{M} \sum_{m=1}^{M} \widehat{\mathbb{IF}}_{22,k}^{(m)} \right)^2}_{T_1} - \underbrace{\frac{2}{M-1} \sum_{m=1}^{M} \left( \widehat{\mathbb{IF}}_{22,k}^{(m),c} - \frac{1}{M} \sum_{m=1}^{M} \widehat{\mathbb{IF}}_{22,k}^{(m),c} \right)^2}_{T_2} \tag{A.13}
$$

where

$$
\begin{aligned}
\widehat{\mathbb{IF}}_{22,k}^{(m)} &= \frac{1}{n(n-1)} \sum_{1 \leqslant i_1 \neq i_2 \leqslant n} W_{i_1}^{(m)} W_{i_2}^{(m)} \widehat{\mathsf{IF}}_{22,k,\bar{i}_2} \\
\widehat{\mathbb{IF}}_{22,k}^{(m),c} &= \frac{1}{n(n-1)} \sum_{1 \leqslant i_1 \neq i_2 \leqslant n} (W_{i_1}^{(m)} - 1)(W_{i_2}^{(m)} - 1) \widehat{\mathsf{IF}}_{22,k,\bar{i}_2}
\end{aligned}
$$

and $\boldsymbol{W}^{(m)} = (W_1^{(m)}, \ldots, W_n^{(m)}) \overset{i.i.d.}{\sim} \mathsf{Multinom}(n, \underbrace{n^{-1}, \ldots, n^{-1}}_{(n-1)'s})$ for $m = 1, \ldots, M$ are $M$ multinomial weights independent of the observed data $\{O_i, i = 1, \ldots, N\}$.

**Comment A.1.** *We now explain why $\widehat{\mathsf{var}}[\widehat{\mathbb{IF}}_{22,k}]$ is proposed as an estimator of $\mathsf{var}_\theta[\widehat{\mathbb{IF}}_{22,k}]$. Conditional on the observed data (including both the training and estimation samples), the expectation over the random multinomial weights $\boldsymbol{W}$ of the term $T_1$ is:*

$$
\mathsf{E}_{\boldsymbol{W}}[T_1 | \{O_i\}_{i=1}^N]
$$

$$
\begin{aligned}
&= \mathsf{E}_{\boldsymbol{W}}\left[\left(\frac{1}{n(n-1)}\sum_{1\leqslant i_1\neq i_2\leqslant n} W_{i_1}^{(m)}W_{i_2}^{(m)}\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\right)^2 |\{O_i\}_{i=1}^N\right] \\
&\quad - \left(\mathsf{E}_{\boldsymbol{W}}\left[\frac{1}{n(n-1)}\sum_{1\leqslant i_1\neq i_2\leqslant n} W_{i_1}^{(m)}W_{i_2}^{(m)}\widehat{\mathsf{IF}}_{22,k,i_1,i_2}|\{O_i\}_{i=1}^N\right]\right)^2 \\
&= \frac{1}{n^2(n-1)^2}\sum_{1\leqslant i_1\neq i_2\leqslant n}\{\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_1,i_2}+\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_2,i_1}\}\{\mathsf{E}_{\boldsymbol{W}}[W_{i_1}^2W_{i_2}^2]-(\mathsf{E}_{\boldsymbol{W}}[W_{i_1}W_{i_2}])^2\} \\
&\quad + \frac{1}{n^2(n-1)^2}\sum_{1\leqslant i_1\neq i_2\neq i_3\leqslant n}\left\{\begin{array}{c}
\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_1,i_3}\{\mathsf{E}_{\boldsymbol{W}}[W_{i_1}^2W_{i_2}W_{i_3}]-\mathsf{E}_{\boldsymbol{W}}[W_{i_1}W_{i_2}]\mathsf{E}_{\boldsymbol{W}}[W_{i_1}W_{i_3}]\} \\
+\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_3,i_1}\{\mathsf{E}_{\boldsymbol{W}}[W_{i_1}^2W_{i_2}W_{i_3}]-\mathsf{E}_{\boldsymbol{W}}[W_{i_1}W_{i_2}]\mathsf{E}_{\boldsymbol{W}}[W_{i_3}W_{i_1}]\} \\
+\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_2,i_3}\{\mathsf{E}_{\boldsymbol{W}}[W_{i_1}W_{i_2}^2W_{i_3}]-\mathsf{E}_{\boldsymbol{W}}[W_{i_1}W_{i_2}]\mathsf{E}_{\boldsymbol{W}}[W_{i_2}W_{i_3}]\} \\
+\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_3,i_2}\{\mathsf{E}_{\boldsymbol{W}}[W_{i_1}W_{i_2}^2W_{i_3}]-\mathsf{E}_{\boldsymbol{W}}[W_{i_1}W_{i_2}]\mathsf{E}_{\boldsymbol{W}}[W_{i_3}W_{i_2}]\}
\end{array}\right\} \\
&\quad + \frac{1}{n^2(n-1)^2}\sum_{1\leqslant i_1\neq i_2\neq i_3\neq i_4\leqslant n}\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_3,i_4}\{\mathsf{E}_{\boldsymbol{W}}[W_{i_1}W_{i_2}W_{i_3}W_{i_4}]-\mathsf{E}_{\boldsymbol{W}}[W_{i_1}W_{i_2}]\mathsf{E}_{\boldsymbol{W}}[W_{i_3}W_{i_4}]\}.
\end{aligned}
$$

*Plugging in the following higher order moments of* $\mathsf{Multinom}(n, \underbrace{n^{-1},\ldots,n^{-1}}_{(n-1)'s})$ (Newcomer et al., 2008)*:*

$$
\begin{aligned}
\mathsf{E}_{\boldsymbol{W}}[W_1W_2] &= \frac{n-1}{n}, \mathsf{E}_{\boldsymbol{W}}[W_1W_2W_3] = \frac{(n-1)(n-2)}{n^2}, \mathsf{E}_{\boldsymbol{W}}[W_1W_2W_3W_4] = \frac{(n-1)(n-2)(n-3)}{n^3}, \\
\mathsf{E}_{\boldsymbol{W}}[W_1^2W_2] &= \frac{2(n-1)^2}{n^2}, \mathsf{E}_{\boldsymbol{W}}[W_1^2] = \frac{2n-1}{n}, \\
\mathsf{E}_{\boldsymbol{W}}[W_1^2W_2W_3] &= \frac{(n-1)(n-2)(n-3)+n(n-1)(n-2)}{n^3} \\
&= \frac{(n-1)(2n^2-7n+6)}{n^3}, \\
\mathsf{E}_{\boldsymbol{W}}[W_1^2W_2^2] &= \frac{(n-1)(n-2)(n-3)+2n(n-1)(n-2)+n^2(n-1)}{n^3} \\
&= \frac{(n-1)(4n^2-9n+6)}{n^3},
\end{aligned}
$$

*we have*

$$
\begin{aligned}
&\mathsf{E}_{\boldsymbol{W}}[T_1|\{O_i\}_{i=1}^N] \\
&= \frac{1}{n^2(n-1)^2}\sum_{1\leqslant i_1\neq i_2\leqslant n}\{\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_1,i_2}+\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_2,i_1}\}\left(3-\frac{11}{n}+\frac{14}{n^2}-\frac{6}{n^3}\right) \\
&\quad + \frac{1}{n^2(n-1)^2}\sum_{1\leqslant i_1\neq i_2\neq i_3\leqslant n}\left\{\begin{array}{c}
\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_1,i_3}+\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_3,i_1} \\
+\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_2,i_3}+\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_3,i_2}
\end{array}\right\}\left(1-\frac{7}{n}+\frac{12}{n^2}-\frac{6}{n^3}\right) \\
&\quad + \frac{1}{n^2(n-1)^2}\sum_{1\leqslant i_1\neq i_2\neq i_3\neq i_4\leqslant n}\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_3,i_4}\left(-\frac{4}{n}+\frac{10}{n^2}-\frac{6}{n^3}\right)
\end{aligned} \tag{A.14}
$$

*The term $T_2$ is the bootstrap variance estimator for second order degenerate U-statistics and it has been proposed previously in Arcones and Gine (1992); Huskova and Janssen (1993). Similarly, conditional on the observed data (including both the* *training and estimation samples), the expectation over the random multinomial weights $\boldsymbol{W}$ of the term $T_2$ is:*

$$
\begin{aligned}
&\mathsf{E}_{\boldsymbol{W}}[T_2|\{O_i\}_{i=1}^N] \\
&= \frac{2}{n^2(n-1)^2}\sum_{1\leqslant i_1\neq i_2\leqslant n}\{\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_1,i_2}+\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_2,i_1}\}\{\mathsf{E}_{\boldsymbol{W}}[(W_{i_1}-1)^2(W_{i_2}-1)^2]-(\mathsf{E}_{\boldsymbol{W}}[(W_{i_1}-1)(W_{i_2}-1)])^2\} \\
&+\frac{2}{n^2(n-1)^2}\sum_{1\leqslant i_1\neq i_2\neq i_3\leqslant n}\left\{\begin{array}{l}\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_1,i_3}\{\mathsf{E}_{\boldsymbol{W}}[(W_{i_1}-1)^2(W_{i_2}-1)(W_{i_3}-1)]-\mathsf{E}_{\boldsymbol{W}}[(W_{i_1}-1)(W_{i_2}-1)]\mathsf{E}_{\boldsymbol{W}}[(W_{i_1}-1)(W_{i_3}-1)]\} \\ +\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_3,i_1}\{\mathsf{E}_{\boldsymbol{W}}[(W_{i_1}-1)^2(W_{i_2}-1)(W_{i_3}-1)]-\mathsf{E}_{\boldsymbol{W}}[(W_{i_1}-1)(W_{i_2}-1)]\mathsf{E}_{\boldsymbol{W}}[(W_{i_3}-1)(W_{i_1}-1)]\} \\ +\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_2,i_3}\{\mathsf{E}_{\boldsymbol{W}}[(W_{i_1}-1)(W_{i_2}-1)^2(W_{i_3}-1)]-\mathsf{E}_{\boldsymbol{W}}[(W_{i_1}-1)(W_{i_2}-1)]\mathsf{E}_{\boldsymbol{W}}[(W_{i_2}-1)(W_{i_3}-1)]\} \\ +\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_3,i_2}\{\mathsf{E}_{\boldsymbol{W}}[(W_{i_1}-1)(W_{i_2}-1)^2(W_{i_3}-1)]-\mathsf{E}_{\boldsymbol{W}}[(W_{i_1}-1)(W_{i_2}-1)]\mathsf{E}_{\boldsymbol{W}}[(W_{i_3}-1)(W_{i_2}-1)]\}\end{array}\right\} \\
&+\frac{2}{n^2(n-1)^2}\sum_{1\leqslant i_1\neq i_2\neq i_3\neq i_4\leqslant n}\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_3,i_4}\{\mathsf{E}_{\boldsymbol{W}}[(W_{i_1}-1)(W_{i_2}-1)(W_{i_3}-1)(W_{i_4}-1)]-\mathsf{E}_{\boldsymbol{W}}[(W_{i_1}-1)(W_{i_2}-1)]\mathsf{E}_{\boldsymbol{W}}[(W_{i_3}-1)(W_{i_4}-1)]\} \\
&= \frac{2}{n^2(n-1)^2}\sum_{1\leqslant i_1\neq i_2\leqslant n}\{\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_1,i_2}+\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_2,i_1}\}\{\mathsf{E}_{\boldsymbol{W}}[(W_1^2-2W_1+1)(W_2^2-2W_2+1)]-(\mathsf{E}_{\boldsymbol{W}}[(W_1-1)(W_2-1)])^2\} \\
&+\frac{2}{n^2(n-1)^2}\sum_{1\leqslant i_1\neq i_2\neq i_3\leqslant n}\left\{\begin{array}{l}\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_1,i_3}+\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_3,i_1} \\ +\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_2,i_3}+\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_3,i_2}\end{array}\right\}\{\mathsf{E}_{\boldsymbol{W}}[(W_1-1)^2(W_2-1)(W_3-1)]-(\mathsf{E}_{\boldsymbol{W}}[(W_1-1)(W_2-1)])^2\} \\
&+\frac{2}{n^2(n-1)^2}\sum_{1\leqslant i_1\neq i_2\neq i_3\neq i_4\leqslant n}\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_3,i_4}\{\mathsf{E}_{\boldsymbol{W}}[(W_1-1)(W_2-1)(W_3-1)(W_4-1)]-(\mathsf{E}_{\boldsymbol{W}}[(W_1-1)(W_2-1)])^2\} \\
&= \frac{2}{n^2(n-1)^2}\sum_{1\leqslant i_1\neq i_2\leqslant n}\{\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_1,i_2}+\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_2,i_1}\}\left(1-\frac{3}{n}+\frac{7}{n^2}-\frac{6}{n^3}-\frac{1}{n^2}\right) \\
&+\frac{2}{n^2(n-1)^2}\sum_{1\leqslant i_1\neq i_2\neq i_3\leqslant n}\left\{\begin{array}{l}\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_1,i_3}+\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_3,i_1} \\ +\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_2,i_3}+\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_3,i_2}\end{array}\right\}\left(-\frac{1}{n}+\frac{5}{n^2}-\frac{6}{n^3}-\frac{1}{n^2}\right) \\
&+\frac{2}{n^2(n-1)^2}\sum_{1\leqslant i_1\neq i_2\neq i_3\neq i_4\leqslant n}\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_3,i_4}\left(\frac{3}{n^2}-\frac{6}{n^3}-\frac{1}{n^2}\right).
\end{aligned}
$$

*Hence $T_2$ serves as a correction term of the inflated factor 3 appeared in equation (A.14).*

*Combining the above computations, we have*

$$
\begin{aligned}
&\mathsf{E}_{\boldsymbol{W}}[\widehat{\mathsf{var}}[\widehat{\mathbb{IF}}_{22,k}]|\{O_i\}_{i=1}^N]=\mathsf{E}_{\boldsymbol{W}}[T_1|\{O_i\}_{i=1}^N]-\mathsf{E}_{\boldsymbol{W}}[T_2|\{O_i\}_{i=1}^N] \\
&=\frac{1}{n^2(n-1)^2}\sum_{1\leqslant i_1\neq i_2\leqslant n}\{\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_1,i_2}+\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_2,i_1}\}\left(1-\frac{5}{n}+\frac{2}{n^2}+\frac{6}{n^3}\right) \\
&\quad+\frac{1}{n^2(n-1)^2}\sum_{1\leqslant i_1\neq i_2\neq i_3\leqslant n}\left\{\begin{array}{l}\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_1,i_3}+\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_3,i_1} \\ +\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_2,i_3}+\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_3,i_2}\end{array}\right\}\left(1-\frac{5}{n}+\frac{6}{n^3}\right) \\
&\quad-\frac{4}{n}\frac{1}{n^2(n-1)^2}\sum_{1\leqslant i_1\neq i_2\neq i_3\neq i_4\leqslant n}\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_3,i_4}\left(1-\frac{1.5}{n}-\frac{1.5}{n^2}\right).
\end{aligned}
$$

*Finally, it is not difficult to see that*

$$
\mathsf{E}_\theta[\mathsf{E}_{\boldsymbol{W}}[\widehat{\mathsf{var}}[\widehat{\mathbb{IF}}_{22,k}]|\{O_i\}_{i=1}^N]]=\mathsf{var}_\theta[\widehat{\mathbb{IF}}_{22,k}]\left(1+O(n^{-1})\right).
$$

**Proposition A.1.** *Under Condition SW, as $n \to \infty$ and $M \to \infty$,*

$$\frac{\widetilde{\mathsf{var}}[\widehat{\mathbb{IF}}_{22,k}]}{\mathsf{var}_\theta[\widehat{\mathbb{IF}}_{22,k}]} = 1 + o_{\mathsf{P}_\theta}(1).$$

*Proof.* Let the $\widetilde{\mathsf{var}}[\widehat{\mathbb{IF}}_{22,k}]$ be the limit (in $\mathsf{P}_{\boldsymbol{W}}$-probability) of $\widehat{\mathsf{var}}[\widehat{\mathbb{IF}}_{22,k}]$ as $M \to \infty$ given all the observed data. Using the moments of multinomial distributions (Newcomer et al., 2008), together with the explanation in Comment A.1, it is easy to show that $\mathsf{E}_\theta\left[\frac{\widetilde{\mathsf{var}}[\widehat{\mathbb{IF}}_{22,k}]}{\mathsf{var}_\theta[\widehat{\mathbb{IF}}_{22,k}]}\right] = 1 + o(1)$ and $\mathsf{var}_\theta\left[\frac{\widetilde{\mathsf{var}}[\widehat{\mathbb{IF}}_{22,k}]}{\mathsf{var}_\theta[\widehat{\mathbb{IF}}_{22,k}]}\right] = o(1)$. Combining the above two claims, we have $\frac{\widetilde{\mathsf{var}}[\widehat{\mathbb{IF}}_{22,k}]}{\mathsf{var}_\theta[\widehat{\mathbb{IF}}_{22,k}]} = 1 + o_{\mathsf{P}_\theta}(1)$. ∎

Following computations similar in spirit to but more tedious than the computations above, we propose the following estimator of $\mathsf{var}_\theta[\widehat{\mathbb{IF}}_{33,k}]$:

$$\begin{aligned}
\widehat{\mathsf{var}}[\widehat{\mathbb{IF}}_{33,k}] \coloneqq & \underbrace{\frac{1}{M-1}\sum_{m=1}^{M}\left(\widehat{\mathbb{IF}}_{33,k}^{(m)} - \frac{1}{M}\sum_{m=1}^{M}\widehat{\mathbb{IF}}_{33,k}^{(m)}\right)^2}_{S_1} \\
& - \underbrace{\frac{1}{M-1}\sum_{m=1}^{M}\left(\widehat{\mathbb{IF}}_{33,k}^{(m),c(b,p)} - \frac{1}{M}\sum_{m=1}^{M}\widehat{\mathbb{IF}}_{33,k}^{(m),c(b,p)}\right)^2}_{S_2} \\
& - \underbrace{\frac{1}{M-1}\sum_{m=1}^{M}\left(\widehat{\mathbb{IF}}_{33,k}^{(m),c(b,\Sigma)} - \frac{1}{M}\sum_{m=1}^{M}\widehat{\mathbb{IF}}_{33,k}^{(m),c(b,\Sigma)}\right)^2}_{S_3} \\
& - \underbrace{\frac{1}{M-1}\sum_{m=1}^{M}\left(\widehat{\mathbb{IF}}_{33,k}^{(m),c(p,\Sigma)} - \frac{1}{M}\sum_{m=1}^{M}\widehat{\mathbb{IF}}_{33,k}^{(m),c(p,\Sigma)}\right)^2}_{S_4}
\end{aligned} \tag{A.15}$$

where

$$\begin{aligned}
\widehat{\mathbb{IF}}_{33,k}^{(m)} &= \frac{1}{n(n-1)(n-2)}\sum_{1\leqslant i_1\neq i_2\neq i_3\leqslant n} W_{i_1}^{(m)}W_{i_2}^{(m)}W_{i_3}^{(m)}\widehat{\mathsf{IF}}_{33,k,\bar{i}_3} \\
\widehat{\mathbb{IF}}_{33,k}^{(m),c(b,p)} &= \frac{1}{n(n-1)(n-2)}\sum_{1\leqslant i_1\neq i_2\neq i_3\leqslant n} (W_{i_1}^{(m)}-1)W_{i_2}^{(m)}(W_{i_3}^{(m)}-1)\widehat{\mathsf{IF}}_{33,k,\bar{i}_3} \\
\widehat{\mathbb{IF}}_{33,k}^{(m),c(b,\Sigma)} &= \frac{1}{n(n-1)(n-2)}\sum_{1\leqslant i_1\neq i_2\neq i_3\leqslant n} (W_{i_1}^{(m)}-1)(W_{i_2}^{(m)}-1)W_{i_3}^{(m)}\widehat{\mathsf{IF}}_{33,k,\bar{i}_3} \\
\widehat{\mathbb{IF}}_{33,k}^{(m),c(p,\Sigma)} &= \frac{1}{n(n-1)(n-2)}\sum_{1\leqslant i_1\neq i_2\neq i_3\leqslant n} W_{i_1}^{(m)}(W_{i_2}^{(m)}-1)(W_{i_3}^{(m)}-1)\widehat{\mathsf{IF}}_{33,k,\bar{i}_3}.
\end{aligned}$$

We then have:

**Proposition A.2.** *Under Condition SW, as $n \to \infty$ and $M \to \infty$,*

$$\frac{\widehat{\mathsf{var}}[\widehat{\mathbb{IF}}_{33,k}]}{\mathsf{var}_\theta[\widehat{\mathbb{IF}}_{33,k}]} = 1 + o_{\mathsf{P}_\theta}(1).$$

Finally, we consider estimating $\mathsf{var}_\theta[\widehat{\mathbb{IF}}_{22\to 33,k}]$ through nonparametric bootstrap. It is not difficult to see that the following estimator has the desired property $\frac{\widehat{\mathsf{var}}[\widehat{\mathbb{IF}}_{22\to 33,k}]}{\mathsf{var}_\theta[\widehat{\mathbb{IF}}_{22\to 33,k}]} = 1 + o_{\mathsf{P}_\theta}(1)$ as $M \to \infty$:

$$\widehat{\mathsf{var}}[\widehat{\mathbb{IF}}_{22\to 33,k}] = \widehat{\mathsf{var}}[\widehat{\mathbb{IF}}_{22,k}] + \widehat{\mathsf{var}}[\widehat{\mathbb{IF}}_{33,k}] + 2\widehat{\mathsf{cov}}[\widehat{\mathbb{IF}}_{22,k}, \widehat{\mathbb{IF}}_{33,k}]$$

where

$$\begin{aligned}
\widehat{\mathsf{cov}}[\widehat{\mathbb{IF}}_{22,k}, \widehat{\mathbb{IF}}_{33,k}] = & \frac{1}{M-1}\sum_{m=1}^{M}\left(\widehat{\mathbb{IF}}_{22,k}^{(m)} - \frac{1}{M}\sum_{m'=1}^{M}\widehat{\mathbb{IF}}_{22,k}^{(m')}\right)\left(\widehat{\mathbb{IF}}_{33,k}^{(m)} - \frac{1}{M}\sum_{m'=1}^{M}\widehat{\mathbb{IF}}_{33,k}^{(m')}\right) \\
& - \frac{2}{M-1}\sum_{m=1}^{M-1}\left(\widehat{\mathbb{IF}}_{22,k}^{(m),c} - \frac{1}{M}\sum_{m'=1}^{M}\widehat{\mathbb{IF}}_{22,k}^{(m'),c}\right)\left(\widehat{\mathbb{IF}}_{33,k}^{(m),c(b,p)} - \frac{1}{M}\sum_{m'=1}^{M}\widehat{\mathbb{IF}}_{33,k}^{(m'),c(b,p)}\right).
\end{aligned}$$

It is not difficult to generalize the above arguments to construct bootstrapped estimators of $\mathsf{var}_\theta[\widehat{\mathbb{IF}}_{mm,k}]$ for general $m \geqslant 2$. We plan to report the general construction elsewhere.

A.4. **Proof of Theorem 3.2.**

*Proof.* The rejection probability of $\widehat{\chi}_{2,k}(\Sigma_k^{-1}; \varsigma_k, \delta)$ is computed as follows:

$$\begin{aligned}
& \lim_{n\to\infty} \mathsf{P}_\theta\left(\frac{|\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})|}{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]} - \varsigma_k \frac{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]}{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]} > \delta\right) \\
= & \lim_{n\to\infty}\left\{\begin{array}{l} \mathsf{P}_\theta\left(\frac{\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]} > \varsigma_k + \delta\frac{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]}\right) \\ + \mathsf{P}_\theta\left(\frac{\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]} < -\varsigma_k - \delta\frac{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]}\right) \end{array}\right\} \\
= & \lim_{n\to\infty} \mathsf{P}_\theta\left(\frac{\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1}) - \mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]} > \varsigma_k - \frac{\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]} + \delta\frac{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]}\right) \\
& + \lim_{n\to\infty} \mathsf{P}_\theta\left(\frac{\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1}) - \mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]} < -\varsigma_k - \frac{\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]} - \delta\frac{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]}\right) \\
= & \lim_{n\to\infty} \mathsf{P}_\theta\left(\frac{\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})) - \mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]}(1 + o_{\mathsf{P}_\theta}(1)) > \varsigma_k - (\gamma - \delta)\frac{\mathsf{s.e.}_\theta[\widehat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]}(1 + o_{\mathsf{P}_\theta}(1))\right) \\
& + \lim_{n\to\infty} \mathsf{P}_\theta\left(\frac{\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1}) - \mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})}(1 + o_{\mathsf{P}_\theta}(1)) < -\varsigma_k - (\gamma + \delta)\frac{\mathsf{s.e.}_\theta[\widehat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]}(1 + o_{\mathsf{P}_\theta}(1))\right) \\
= & 1 - \Phi\left(\varsigma_k - (\gamma - \delta)\frac{\mathsf{s.e.}_\theta[\widehat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]}\right) + \Phi\left(-\varsigma_k - (\gamma + \delta)\frac{\mathsf{s.e.}_\theta[\widehat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]}\right) \\
= & 2 - \Phi\left(\varsigma_k - (\gamma - \delta)\frac{\mathsf{s.e.}_\theta[\widehat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]}\right) - \Phi\left(\varsigma_k + (\gamma + \delta)\frac{\mathsf{s.e.}_\theta[\widehat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]}\right).
\end{aligned}$$

∎

A.5. **Proof of Proposition 4.2.** As discussed in Section 4, since $\Sigma_k^{-1}$ is generally unknown, $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ needs to be replaced by $\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})$. We consider the following statistic used in the test $\widehat{\chi}_{3,k}(\widehat{\Sigma}_k^{-1};\varsigma_k,\delta)$ after standardization:

(A.16)

$$
\begin{aligned}
&\frac{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})]}\left(\frac{\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})}{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]}-\delta\right)\\
&=\left(\frac{\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})-\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)}{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]}\frac{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})]}\right)+\frac{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})]}\left(\frac{\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)}{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]}-\delta\right)\\
&=\left(\frac{\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})-\mathsf{E}_\theta[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})]}{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]}\frac{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})]}\right)+\frac{\mathsf{EB}_{\theta,3,k}(\widehat{\Sigma}_k^{-1})}{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]}\frac{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})]}\\
&\quad+\frac{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})]}\left(\frac{\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)}{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]}-\delta\right)\\
&=\left\{\left(\frac{\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})-\mathsf{E}_\theta[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})]}\right)+\left(\frac{\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)+\mathsf{EB}_{\theta,3,k}(\widehat{\Sigma}_k^{-1})}{\mathsf{s.e.}_\theta[\widehat{\psi}_1]}-\delta\right)\frac{\mathsf{s.e.}_\theta[\widehat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})]}\right\}(1+o_{\mathsf{P}_\theta}(1))\\
&=\left\{\left(\frac{\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})-\mathsf{E}_\theta[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})]}\right)+\left(\gamma+\frac{\mathsf{EB}_{\theta,3,k}(\widehat{\Sigma}_k^{-1})}{\mathsf{s.e.}_\theta[\widehat{\psi}_1]}-\delta\right)\frac{\mathsf{s.e.}_\theta[\widehat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})]}\right\}(1+o_{\mathsf{P}_\theta}(1))\\
&=\left\{\underbrace{\left(\frac{\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})-\mathsf{E}_\theta[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})]}\right)-(\delta-\gamma)\frac{\mathsf{s.e.}_\theta[\widehat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})]}}_{A}+\underbrace{\frac{\mathsf{EB}_{\theta,3,k}(\widehat{\Sigma}_k^{-1})}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})]}}_{B}\right\}(1+o_{\mathsf{P}_\theta}(1)).
\end{aligned}
$$

The effect of estimating $\Sigma_k^{-1}$ on the asymptotic validity of the test $\widehat{\chi}_{3,k}(\widehat{\Sigma}_k^{-1};\varsigma_k,\delta)$ of $\mathsf{H}_{0,k}(\delta)$ thus depends on the orders of terms A and B. A has variance 1 and mean $-(\delta-\gamma)\frac{\mathsf{s.e.}_\theta[\widehat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})]}$. B depends on the estimation bias due to estimating $\Sigma_k^{-1}$ by $\widehat{\Sigma}_k^{-1}$. Hence if we have:

(1) $\frac{\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})-\mathsf{E}_\theta[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})]}$ is asymptotically $N(0,1)$ conditional on the training sample;
(2) $\frac{\mathsf{s.e.}_\theta[\widehat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})]}\Big/\frac{\mathsf{s.e.}_\theta[\widehat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})]}\to 1$;
(3) $B=o(1)$ under $\mathsf{H}_{0,k}(\delta)$ or fixed alternatives to $\mathsf{H}_{0,k}(\delta)$;

(3′) $B\ll-(\delta-\gamma)\frac{\mathsf{s.e.}_\theta[\widehat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})]}$ under diverging alternatives to $\mathsf{H}_{0,k}(\delta)$ i.e. $\delta-\gamma=c$ for some $c\to\infty$ (at any rate).

$\widehat{\chi}_{3,k}(\widehat{\Sigma}_k^{-1};\varsigma_k,\delta)$ is an asymptotically level $2-2\Phi(\varsigma_k)$ two-sided test for $\mathsf{H}_{0,k}(\delta)$ and rejects the null with probability approaching 1 under diverging alternatives to $\mathsf{H}_{0,k}(\delta)$.

According to Proposition 4.1, both (1) and (2) hold for $\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})$. In terms of (3) and (3′), under the conditions of Proposition 4.2:

- Under $\mathsf{H}_{0,k}(\delta)$ or fixed alternatives to $\mathsf{H}_{0,k}(\delta)$, $\mathsf{EB}_{\theta,3,k}(\widehat{\Sigma}_k^{-1})\lesssim\frac{k\log(k)}{n^{3/2}}\ll\frac{\sqrt{k}}{n}+\frac{1}{\sqrt{n}}$ and hence $B=o(1)$. Thus (3) is satisfied.

- Under diverging alternatives to $\mathsf{H}_{0,k}(\delta)$ i.e. $\gamma - \delta = c \to \infty$, $(\gamma - \delta)\mathsf{s.e.}_\theta(\widehat{\psi}_1) \asymp \mathbb{L}_{\theta,2,\widehat{b},k}\mathbb{L}_{\theta,2,\widehat{p},k}$,

$$B \lesssim \frac{\mathbb{L}_{\theta,2,\widehat{b},k}\mathbb{L}_{\theta,2,\widehat{p},k}}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})]} \frac{k\log(k)}{n} \ll \frac{\mathbb{L}_{\theta,2,\widehat{b},k}\mathbb{L}_{\theta,2,\widehat{p},k}}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})]}$$
$$\asymp -(\delta - \gamma)\frac{\mathsf{s.e.}_\theta[\widehat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to 33,k}(\widehat{\Sigma}_k^{-1})]}.$$

  Thus (3′) is satisfied.

In summary, $\widehat{\chi}_{3,k}(\widehat{\Sigma}_k^{-1}; \varsigma_k, \delta)$ is an asymptotically valid level $2 - 2\Phi(\varsigma_k)$ two-sided test for $\mathsf{H}_{0,k}(\delta)$ and rejects the null with probability approaching 1 under diverging alternatives to $\mathsf{H}_{0,k}(\delta)$.

## Appendix B. Supplementary information for simulation studies

The functions $h_f$, $h_b$ and $h_p$ appeared in Section 7 are of the following forms:

$$h_f(x; s_f) \propto 1 + \exp\left\{\frac{1}{2}\sum_{j\in\mathcal{J},\ell\in\mathbb{Z}} 2^{-j(s_f+0.25)}\omega_{j,\ell}(x)\right\}, \tag{B.1}$$

$$h_b(x; s_b) = \sum_{j\in\mathcal{J},\ell\in\mathbb{Z}} 2^{-j(s_b+0.25)}\omega_{j,\ell}(x), \tag{B.2}$$

$$h_p(x; s_p) = \operatorname{expit}\left\{-2\sum_{j\in\mathcal{J},\ell\in\mathbb{Z}} 2^{-j(s_p+0.25)}\omega_{j,\ell}(x)\right\} \tag{B.3}$$

where $\mathcal{J} = \{0, 3, 6, 9, 10, 16\}$ and $\omega_{j,\ell}(\cdot)$ is the D12 (or equivalently db6) father wavelets function dilated at resolution $j$, shifted by $\ell$ (Daubechies, 1992; Härdle et al., 1998; Mallat, 1999). Härdle et al. (1998)[Theorem 9.6] indeed implies that $h_f(\cdot; s_f) \in$ Hölder$(s_f)$, $h_b(\cdot; s_b) \in$ Hölder$(s_b)$ and $h_p(\cdot; s_p) \in$ Hölder$(s_p)$. We fix $s_f = 0.1$. In simulation setup I we choose $s_b = s_p = 0.25$ whereas in simulation setup II we choose $s_b = s_p = 0.6$.

| $j$ | $\tau_{b,j}$ | $\tau_{p,j}$ |
|---|---|---|
| 1 | -0.2819 | 0.09789 |
| 2 | 0.4876 | 0.08800 |
| 3 | -0.1515 | -0.4823 |
| 4 | -0.1190 | 0.4588 |

Table 5. Coefficients used in constructing $b$ and $\pi$ in Section 7.

In Table 5, we provide the numerical values for $(\tau_{b,j}, \tau_{p,j})_{j=1}^8$ used in generating the simulation experiments in Section 7.

B.1. **Generating correlated multidimensional covariates $X$ with fixed non-smooth marginal densities.** In the simulation study conducted in Section 7, one key step of generating the simulated datasets is to draw correlated multidimensional covariates $X \in [0,1]^d$ with fixed non-smooth marginal densities. First, we fix the marginal densities of $X$ in each dimension proportional to $h_f(\cdot)$ (equation (B.1)). Then we make $2K$ independent draws of $\widetilde{X}_{i,j}$, $i = 1, \ldots, 2K$, from $h_f$ for every $j = 1, \ldots, d$ so $\widetilde{X} = (\widetilde{X}_{1,\cdot}, \ldots, \widetilde{X}_{2K,\cdot})^\top \in [0,1]^{2K\times d}$. Next, to create correlations between different dimensions, we follow the strategy proposed in Baker (2008). First we group every two consecutive draws:

$$(\widetilde{X}_{1,\cdot}, \widetilde{X}_{2,\cdot})^\top, (\widetilde{X}_{3,\cdot}, \widetilde{X}_{4,\cdot})^\top, \ldots, (\widetilde{X}_{2K-1,\cdot}, \widetilde{X}_{2K,\cdot})^\top.$$

Then for each pair $(\widetilde{X}_{2i-1,\cdot}, \widetilde{X}_{2i,\cdot})^\top$ for $i = 1, \ldots, K$, we form the following $d$-dimensional random vectors

$$U_i \coloneqq (\mathsf{max}(\widetilde{X}_{2i-1,1}, \widetilde{X}_{2i,1}), \ldots, \mathsf{max}(\widetilde{X}_{2i-1,d}, \widetilde{X}_{2i,d}))^\top,$$
$$V_i \coloneqq (\mathsf{min}(\widetilde{X}_{2i-1,1}, \widetilde{X}_{2i,1}), \ldots, \mathsf{min}(\widetilde{X}_{2i-1,d}, \widetilde{X}_{2i,d}))^\top.$$

Lastly, we construct $K$ independent $d$-dimensional vectors $X$ by the following rule: for each $i = 1, \ldots, K$, we draw a Bernoulli random variable $B_i$ with probability 1 / 2, and if $B_i = 0$, $X_{i,\cdot} = U_i$, otherwise $X_{i,\cdot} = V_i$. Following the above strategy, we conserve the marginal density of $X_{\cdot,j}$ as that of $\widetilde{X}_{\cdot,j}$ but create dependence between different dimensions.