Can we falsify the justification of the validity of published Wald confidence intervals of doubly robust functionals, without assumptions?

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

ABSTRACT. In this article we develop a feasible version of the assumption-lean tests in Liu, Mukherjee, and Robins (2020a) that can *falsify an analyst's justification* for the validity of a reported nominal $(1 - \alpha)$ Wald confidence interval (CI) centered at a double machine learning (DML) estimator (Chernozhukov et al., 2018) for any member of the class of doubly robust (DR) functionals studied by Rotnitzky, Smucler, and Robins (2021). The class of DR functionals is broad and of central importance in economics and biostatistics. 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, Newey, and Singh (2022a) and the class of functionals studied by Robins et al. (2008). The present state-of-the-art estimators for DR functionals ψ are DML estimators $\widehat{\psi}_1$. The bias of $\widehat{\psi}_1$ depends on the product of the rates at which two nuisance functions b and p are estimated. Most commonly an analyst justifies the validity of her Wald CIs by proving that, under her 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 $\widehat{\psi}_1$ is $o(n^{-1/2})$. Thus if the hypothesis H_0: the CS upper bound is $o(n^{-1/2})$ is rejected by our test (with an extremely small p-value), we will have falsified the analyst's justification for the validity of her Wald CIs. In this work, we exhibit a valid assumption-lean falsification test of H_0, without relying on complexity-reducing assumptions on b, p, or their estimates $\widehat{b}, \widehat{p}$. To the best of our knowledge, such an assumption-lean falsification test is the first of its kind. Simulation experiments are conducted to demonstrate how the proposed assumption-lean test can be used in practice. An unavoidable limitation of our methodology is that no assumption-lean test of H_0, including ours, can be a consistent test. Thus failure of our test to reject is not meaningful evidence in favor of H_0.

Key words: Econometrics and Machine Learning, Assumption-lean Statistics, Doubly Robust Functionals, Higher-Order U-Statistics, Higher-Order Influence Functions

---

1. Institute of Natural Sciences, MOE-LSC, School of Mathematical Sciences, CMA-Shanghai, SJTU-Yale Joint Center for Biostatistics and Data Science, Shanghai Jiao Tong University; Shanghai Artificial Intelligence Laboratory 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.

1. Introduction

Suppose a data analyst has constructed and published a nominal $1-\alpha$ large sample Wald confidence interval (CI) for a mean-square continuous linear functional ψ of a conditional expectation $b(x) = \mathsf{E}[Y|X = x]$ with the Wald CI centered by a doubly robust Double Machine Learning (DML) estimator $\widehat{\psi}_1$ of ψ (Chernozhukov, Newey, and Singh, 2022a). The estimator $\widehat{\psi}_1$ will depend on estimates $\widehat{b}(x)$ and $\widehat{p}(x)$ of two functions of x: $b(x)$ itself and the function $p(x)$ occurring in the Riesz representer of the linear functional. If queried, most data analysts would justify the validity of the Wald CI (i.e. that the actual asymptotic coverage is at least $1 - \alpha$) by (i) making restrictive assumptions on both the complexities of b and p (e.g. in terms of smoothness or sparsity) and the class of algorithms used in their estimation, followed by (ii) an appeal to theorems that guarantee the asymptotic validity of the Wald CI under these restrictions. However, these assumptions may be incorrect. Indeed, we conjecture that in many studies with high dimensional X, the actual coverage of the aforementioned nominal $1-\alpha$ interval is less, and often much less, than $1-\alpha$. To investigate such a conjecture we would like to construct an assumption-lean (i.e. essentially assumption-free) empirical test, with non-negligible power against certain alternatives, of the *null hypothesis* that the true asymptotic coverage of ψ of the above nominal $1 - \alpha$ Wald CI is greater than or equal to $1 - \alpha$ under repeated sampling. Henceforth we will say a large sample Wald CI is valid if and only if the above null hypothesis is true. Unfortunately, as discussed below, such a test can be shown not to exist. However, progress can often still be made as follows.

Chernozhukov et al. (2018) showed that a sufficient (but not necessary) condition for a Wald CI centered at a doubly robust DML estimator to be valid is that the (weighted) $L_2(\mathsf{P})$ rate of convergence $n^{-\kappa_b}$ of $\widehat{b}$ to b multiplied by the rate of convergence $n^{-\kappa_p}$ of $\widehat{p}$ to p is $o_p(n^{-1/2})$ or, equivalently, $\kappa_b + \kappa_p > 1/2$. The main technical result of the our paper is that we exhibit an assumption-lean empirical test, with power against certain alternatives, of the null hypothesis that $\kappa_b + \kappa_p > 1/2$. One might naturally wonder how can we make use of our test, as the condition $\kappa_b + \kappa_p > 1/2$ is not necessary for the validity of the nominal Wald CI?

The central goal of our paper is to answer this question. The answer is that our test can falsify an analyst's justification for their claim that their nominal Wald CI centered at a *standard* DML estimator (for its definition, see later in the Introduction) is valid. If falsified, the analyst should then retract her claim that her Wald CI is valid. For a standard DML estimator, the only justification for the validity of Wald CI centered at the estimator is to make complexity reducing assumptions on b and p so as to guarantee that $\kappa_b + \kappa_p > 1/2$; otherwise Chernozhukov et al. (2018) would have offered an alternative justification[1]. As a concrete example, when b and p are fitted by certain (weighted) penalized empirical squared L_2-risk minimization (see (2.8)) using deep neural networks (Farrell, Liang, and Misra, 2021; Chen et al., 2020;

---

[1] If our test rejects the hypothesis that $\kappa_b + \kappa_p > 1/2$, it also rejects the hypothesis that any assumed collection of restrictions on complexity of and fitting algorithms for b and p that imply $\kappa_b + \kappa_p > 1/2$ are all true.

Xu, Liu, and Liu, 2022), $L_2(\mathsf{P})$-convergence rates can often be established under complexity-reducing assumptions such as Hölder smoothness (Schmidt-Hieber, 2020).

Our results will apply to the more general class of mixed bias functionals or Doubly Robust (DR) functionals studied in Rotnitzky, Smucler, and Robins (2021). 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, Newey, and Singh (2022a) and the class of functionals studied by Robins et al. (2008). Many DR functionals are of substantive scientific and economic interest, see the examples after Definition 2.1. Our unified treatment of the entire DR functional class requires that we use rather abstract notation. In order to prevent such abstract notation from making the reader miss the forest for the trees, we shall complete the introduction by using a familiar functional to motivate both our goals and our methodology. Furthermore, detailed regularity conditions will be suppressed in the Introduction to facilitate the exposition.

The motivation for our paper is best described by the following inferential quandary faced many (perhaps dozens) of times daily by data analysts employed by large tech companies. The quandary arises when the analyst needs to estimate a population average effect $\mathsf{E}[Y(a=1)] - \mathsf{E}[Y(a=0)]$ of a dichotomous treatment A on a response Y from observational data. Here $Y(a)$ is the counterfactual outcome under treatment level a. Often, data on a very high-dimensional vector X (with dimension d) of pre-treatment covariates are available and deemed sufficient for ignorability $Y(a) \perp\!\!\!\perp A | X$ to hold, thereby identifying[2] $\psi^a \equiv -\mathsf{E}[Y(a)]$ as $-\mathsf{E}[b_a(X)]$ with $b_a(x) = \mathsf{E}[Y|X = x, A = a]$. Chernozhukov et al. (2018) argued persuasively that in an effort to obtain valid inference (i.e. confidence intervals) for ψ^a [and thus for the average treatment effect (ATE) $-\psi^{a=1} + \psi^{a=0}$] with very high dimensional X, one should use so-called Double Machine Learning (DML) estimators $\widehat{\psi}^a_{\mathsf{cf},1}$ to center nominal $1 - \alpha$ large sample Wald CI $\widehat{\psi}^a_{\mathsf{cf},1} \pm z_{\alpha/2}\widehat{\mathsf{s.e.}}(\widehat{\psi}^a_{\mathsf{cf},1})$ where $z_{\alpha/2}$ is the $\alpha/2$ upper-quantile of a standard normal random variable and $\widehat{\mathsf{s.e.}}(\widehat{\psi}^a_{\mathsf{cf},1})$ is the estimator of the standard error $\mathsf{s.e.}(\widehat{\psi}^a_{\mathsf{cf},1})$ of $\widehat{\psi}^a_{\mathsf{cf},1}$ given in Proposition 2.2 below. As a result, DML estimators rapidly became the standard in high tech.

For DR functionals, DML estimators combine the benefits of cross-fitting (cf), double robustness, and machine learning of nuisance parameters (Chernozhukov et al., 2018; Chernozhukov et al., 2022). Henceforth, for notational convenience, we remove the a index by restricting to the case $a = 1$ so, for example, $\psi^{a=1}$ becomes ψ and b_a becomes b. Then, to compute a DML estimator, the data is randomly divided into two (or more) samples – the estimation sample of size n and the training or nuisance/training 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 take $c = 0.5$. Estimators $\widehat{b}(\cdot)$ and $\widehat{p}(\cdot)$ of $b(\cdot)$ and inverse propensity score $p(\cdot) := 1/\mathsf{E}[A|X = \cdot]$, where p^{-1} is assumed to be strictly bounded between $(0, 1)$, are computed from the training sample

---

[2]The minus sign plays no essential role and is added for notational convenience that will be made clear in Definition 2.1.

data using modern black-box highly nonlinear machine learning algorithms, often deep neural networks. In semiparametric statistics literature, b and p are referred to as nuisance parameters/functions. Let P_n denote a sample average over the estimation sample. Then the doubly robust one step estimator $\widehat{\psi}_1 = \widehat{\psi} + \mathsf{P}_n[\widehat{\mathsf{IF}}_{1,\psi}] = \mathsf{P}_n[-\widehat{b}(X) - A\widehat{p}(X)(Y - \widehat{b}(X))]$ is constructed by adding to an initial plug-in estimator $\widehat{\psi} = \mathsf{P}_n[\widehat{b}(X)]$ the sample average of an estimate of the first order influence function $\mathsf{IF}_{1,\psi} = -b(X) - Ap(X)(Y - b(X)) - \psi$ of ψ[3]. The cross-fit DML estimator $\widehat{\psi}_{\mathsf{cf},1}$ is the arithmetic average of $\widehat{\psi}_1$ and $\overline{\overline{\widehat{\psi}}}_1$, where $\overline{\overline{\widehat{\psi}}}_1$ is computed like $\widehat{\psi}_1$ but with the roles of training and estimation samples switched. We define a standard DML estimator to be a DML estimator where $\widehat{b}$ and $\widehat{p}$ are separately estimated, each using data from the entire training sample; see Section 2.2. This allows us to distinguish standard DML estimators from, for instance, "DCDR estimators" of Newey and Robins (2018) that estimate b and p from separate non-overlapping subsamples of the training sample.

The analyst's inferential quandary is how to justify the claim that $\widehat{\psi}_{\mathsf{cf},1} \pm z_{\alpha/2}\widehat{\mathsf{s.e.}}(\widehat{\psi}_{\mathsf{cf},1})$ is a valid large sample $1 - \alpha$ Wald CI. For it to be valid the following are necessary: (i) $\widehat{\psi}_{\mathsf{cf},1}$ is asymptotically normal, (ii) $\widehat{\mathsf{s.e.}}(\widehat{\psi}_{\mathsf{cf},1})/\mathsf{s.e.}(\widehat{\psi}_{\mathsf{cf},1})$ converges to 1 in probability and (iii) the (asymptotic) bias of $\widehat{\psi}_{\mathsf{cf},1}$ is of smaller order than $\mathsf{s.e.}(\widehat{\psi}_{\mathsf{cf},1})$. Since $\widehat{\psi}_{\mathsf{cf},1}$ is the average of $\widehat{\psi}_1$ and its "twin" $\overline{\overline{\widehat{\psi}}}_1$, it must be the case that $\widehat{\psi}_1 \pm z_{\alpha/2}\widehat{\mathsf{s.e.}}(\widehat{\psi}_1)$ is also a valid large sample $1 - \alpha$ Wald CI for ψ, therefore satisfying (i) - (iii) with $\widehat{\psi}_1$ substituted for $\widehat{\psi}_{\mathsf{cf},1}$. Since $\mathsf{s.e.}(\widehat{\psi}_1)$ is order $n^{-1/2}$, it is necessary that its bias is $o_p(n^{-1/2})$ to satisfy (iii) above. As discussed in Newey and Robins (2018), by far the most difficult of the three assumptions to satisfy and thus to justify is (iii). As mentioned above, for a standard DML estimator, if $\kappa_b + \kappa_p > 1/2$[4] holds, then (iii) holds. Specifically, that $\kappa_b + \kappa_p > 1/2$ implies the bias of $\widehat{\psi}_1$ is $o_p(n^{-1/2})$ is a consequence of applying Cauchy-Schwarz (CS) inequality to upper bound the bias, as we show next. The exact conditional bias of $\widehat{\psi}_1$ given the training sample, denoted as $\mathsf{Bias}(\widehat{\psi}_1)$, is $\mathsf{E}[A(\widehat{b}(X) - b(X))(\widehat{p}(X) - p(X))] = \int p^{-1}(x)(\widehat{b}(x) - b(x))(\widehat{p}(x) - p(x))\mathrm{d}F(x)$, where F denotes the distribution of X. Here and henceforth, unless stated otherwise, all expectations will be understood to be conditional on the training sample, although that fact is suppressed in the notation for sake of brevity. It then follows from the CS inequality that

(1.1) $\qquad |\mathsf{Bias}(\widehat{\psi}_1)| \leqslant \{\mathsf{E}[p^{-1}(X)(\widehat{b}(X) - b(X))^2]\}^{1/2}\{\mathsf{E}[p^{-1}(X)(\widehat{p}(X) - p(X))^2]\}^{1/2}.$

We refer to the RHS of the above display as the (conditional on the training sample) Cauchy Schwarz bias, denoted as $\mathsf{CSBias}(\widehat{\psi}_1)$, of $\widehat{\psi}_1$. More generally, for any pair of positive functions $w = (w_b, w_p)$ of x strictly bounded from above and below, define

$$\mathsf{CSBias}^w(\widehat{\psi}_1) \equiv \{\mathsf{E}[w_b(X)(\widehat{b}(X) - b(X))^2]\}^{1/2}\{\mathsf{E}[w_p(X)(\widehat{p}(X) - p(X))^2]\}^{1/2}.$$

---

[3]In the econometrics literature, $\widehat{\psi}$ is referred to as a first stage estimator, $\widehat{\psi}_1$ as the second stage estimator, and $\mathsf{P}_n[-A\widehat{p}(X)(Y - \widehat{b}(X))]$ is the debiasing term that makes $\widehat{\psi}_1$ a doubly robust estimator satisfying Neyman orthogonality (Chernozhukov et al., 2018; Chernozhukov et al., 2022).

[4]This condition is referred to as "rate double robustness" in Smucler, Rotnitzky, and Robins (2019)

Then applying Hölder inequality, as in (2.9) later in our paper, we have that $\mathsf{CSBias}^w(\widehat{\psi}_1)$ and $\mathsf{CSBias}(\widehat{\psi}_1)$ are equal up to a multiplicative positive constant, except that when $w_b = w_p = p^{-1}$, the equality is exact. Hence $\mathsf{CSBias}^w(\widehat{\psi}_1) = o_p(n^{-1/2})$, if and only if $\mathsf{CSBias}(\widehat{\psi}_1) = o_p(n^{-1/2})$, if and only if $\kappa_b + \kappa_p > 1/2$ (here we refer to $n^{-\kappa_b}$ and $n^{-\kappa_p}$ as the p^{-1}-weighted $L_2(\mathsf{P})$ convergence rates of $\widehat{b}$ and $\widehat{p}$ to b and p). Thus $\mathsf{CSBias}^w(\widehat{\psi}_1) = o_p(n^{-1/2})$ also implies $\mathsf{Bias}(\widehat{\psi}_1) = o_p(n^{-1/2})$.

Main technical contributions. If we can empirically reject the null hypothesis $\mathsf{NH}_{0,\mathsf{CS}} : \mathsf{CSBias}(\widehat{\psi}_1) = o_p(n^{-1/2})$, then $\kappa_b + \kappa_p \leqslant 1/2$, and we falsify the justification of the validity of a Wald CI centered at the standard DML estimator $\widehat{\psi}_1$. The main technical contribution of this paper is to construct a test of $\mathsf{NH}_{0,\mathsf{CS}}$.

The reader might rightfully complain at this point that finite sample tests of asymptotic hypotheses such as $\mathsf{NH}_{0,\mathsf{CS}}$ (or equivalently $\kappa_b + \kappa_p > 1/2$) that concern rates of convergence cannot be constructed. To overcome this problem, we pair each asymptotic null hypothesis of interest with a natural non-asymptotic null hypothesis and then, by convention, formally declare the asymptotic hypothesis (not) rejected if the paired non-asymptotic null hypothesis is (not) rejected. We say the asymptotic null hypothesis has been **operationalized** by its paired non-asymptotic null hypothesis. In particular, we pair the asymptotic null hypothesis $\mathsf{NH}_{0,\mathsf{CS}}$ with the following non-asymptotic null hypothesis

$$(1.2) \qquad\qquad \mathsf{H}_{0,\mathsf{CS}}(\delta) : \mathsf{CSBias}(\widehat{\psi}_1) \leqslant \delta\,\mathsf{s.e.}(\widehat{\psi}_1)$$

and construct an $\alpha^\dagger$-level falsification test of $\mathsf{H}_{0,\mathsf{CS}}(\delta)$, where $\delta > 0$ is chosen by the analyst. With such operationalized pairing

$$(1.3) \qquad \mathsf{NH}_{0,\mathsf{CS}} : \mathsf{CSBias}(\widehat{\psi}_1) = o_p(n^{-1/2}) \sim \mathsf{H}_{0,\mathsf{CS}}(\delta) : \mathsf{CSBias}(\widehat{\psi}_1) < \mathsf{s.e.}(\widehat{\psi}_1)\delta,$$

we will, by convention, declare $\mathsf{NH}_{0,\mathsf{CS}}$ (not) rejected if $\mathsf{H}_{0,\mathsf{CS}}(\delta)$ is (not) rejected. The larger δ that one chooses, the stronger evidence that rejection of $\mathsf{H}_{0,\mathsf{CS}}(\delta)$ has against the asymptotic null hypothesis $\mathsf{NH}_{0,\mathsf{CS}}$, but the less power one has to reject $\mathsf{H}_{0,\mathsf{CS}}(\delta)$. Suppose the analyst insists that her justification is $\mathsf{CSBias}^w(\widehat{\psi}_1) = o(n^{-1/2})$ as defined in (2.11), then, by the aforementioned equivalence between $\mathsf{CSBias}(\widehat{\psi}_1)$ and $\mathsf{CSBias}^w(\widehat{\psi}_1)$, we can use the same operationalized pairing (1.3) to empirically falsify $\mathsf{CSBias}_\theta^w(\widehat{\psi}_1) = o(n^{-1/2})$.

Since we want our test to be assumption-lean, we make essentially no assumptions on the nuisance functions b and p, their estimates $\widehat{b}$ and $\widehat{p}$, or the algorithms used to construct $\widehat{b}$ and $\widehat{p}$ from the training sample. To avoid relying on assumptions on $\widehat{b}$ and $\widehat{p}$, the falsification test and its properties are established by conditioning on the training sample, so any randomness from the training sample is treated as fixed and thus statements such as $\mathsf{CSBias}(\widehat{\psi}_1) = o_p(n^{-1/2})$ becomes $\mathsf{CSBias}(\widehat{\psi}_1) = o(n^{-1/2})$. By arguing as in Robins and Ritov (1997) or Ritov et al. (2014), in absence of complexity reducing assumptions on b and p, there is no uniformly consistent estimator of either $\mathsf{Bias}(\widehat{\psi}_1)$ or $\mathsf{CSBias}(\widehat{\psi}_1)$. However, we will exhibit a

5

functional, denoted as $\mathsf{Bias}_k(\widehat{\psi}_1)$, that is uniformly consistently estimable using a third-order U-statistic [derived using the theory of Higher-Order Influence Functions (HOIFs) (Robins et al., 2008; Liu et al., 2017)] such that one can reject the "k-projected" null hypothesis

$$(1.4) \qquad\qquad \mathsf{H}_{0,k}(\delta) : |\mathsf{Bias}_k(\widehat{\psi}_1)| \leqslant \delta \, \mathsf{s.e.}(\widehat{\psi}_1)$$

with nontrivial power. Here $\mathsf{Bias}_k(\widehat{\psi}_1)$ is defined as

$$\mathsf{Bias}_k(\widehat{\psi}_1) := \mathsf{E}[\Pi[p^{-1/2}(b-\widehat{b})|p^{-1/2}\bar{\mathsf{z}}_k](X)\Pi[p^{-1/2}(p-\widehat{p})|p^{-1/2}\bar{\mathsf{z}}_k](X)],$$

where for any $h \in L_2(\mathsf{P}_F)$, $\Pi[h|p^{-1/2}\bar{\mathsf{z}}_k]$ denotes the population projection of h onto the linear span of a (user-selected) k-dimensional basis functions $p^{-1/2}\bar{\mathsf{z}}_k = p^{-1/2}(\mathsf{z}_1,\cdots,\mathsf{z}_k)^\top$. We now show that $|\mathsf{Bias}_k(\widehat{\psi}_1)|$ is a lower bound for $\mathsf{CSBias}(\widehat{\psi}_1)$ as follows:

(1.5)
$$|\mathsf{Bias}_k(\widehat{\psi}_1)| \leqslant \underbrace{\{\mathsf{E}[\Pi[p^{-1/2}(b-\widehat{b})|p^{-1/2}\bar{\mathsf{z}}_k](X)^2]\}^{1/2}\{\mathsf{E}[\Pi[p^{-1/2}(p-\widehat{p})|p^{-1/2}\bar{\mathsf{z}}_k](X)^2]\}^{1/2}}_{=:\mathsf{CSBias}_k(\widehat{\psi}_1)} \leqslant \mathsf{CSBias}(\widehat{\psi}_1)$$

where the first inequality again follows from CS inequality and the second inequality follows from the fact that projection contracts norms[5]. Hence if $\mathsf{H}_{0,k}(\delta)$ is false, then $\mathsf{H}_{0,\mathsf{CS}}(\delta)$ is false and if $\mathsf{H}_{0,\mathsf{CS}}(\delta)$ is true, then $\mathsf{H}_{0,k}(\delta)$ is true; but the converses of the above two clauses do not necessarily hold. In fact, no test, ours included, can be a consistent test of $\mathsf{H}_{0,\mathsf{CS}}(\delta)$ (that is, no test can have power against all alternatives to $\mathsf{H}_{0,\mathsf{CS}}(\delta)$) unless one makes further possibly incorrect complexity-reducing assumptions on the nuisance functions of b and p and their estimates $\widehat{b}$ and $\widehat{p}$. This again follows from the argument in Robins and Ritov (1997) or Ritov et al. (2014). But we also provide an intuitive explanation in Section 2.3.

To further illustrate our approach, we can rewrite $\mathsf{Bias}_k(\widehat{\psi}_1)$ as follows (see Appendix A.3):

$$\mathsf{Bias}_k(\widehat{\psi}_1) = \mathsf{E}[A(Y-\widehat{b}(X))\bar{\mathsf{z}}_k(X)^\top]\Sigma_k^{-1}\mathsf{E}[\bar{\mathsf{z}}_k(X)(A\widehat{p}(X)-1)]$$

where $\Sigma_k \equiv \mathsf{E}[A\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top]$. When Σ_k is known, $\mathsf{Bias}_k(\widehat{\psi}_1)$ can be unbiasedly estimated by the following second-order U-statistic:

$$\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1}) \equiv \frac{1}{n(n-1)}\sum_{1\leqslant i_1\neq i_2\leqslant n} A_{i_1}(Y_{i_1}-\widehat{b}(X_{i_1}))\bar{\mathsf{z}}_k(X_{i_1})^\top\Sigma_k^{-1}\bar{\mathsf{z}}_k(X_{i_2})(A_{i_2}\widehat{p}(X_{i_2})-1),$$

which, by definition in Robins et al. (2008), is the second-order influence function of $\mathsf{Bias}_k(\widehat{\psi}_1)$. Since Σ_k is generally unknown, we estimate Σ_k from the training sample and use a third-order U-statistic to further correct the estimation bias to protect the level of our test; see Section 3 for more details.

---

[5]The reason why we do not focus on a different k-projected null hypothesis

$$(1.6) \qquad\qquad \mathsf{H}_{0,k,\mathsf{CS}}(\delta) : \mathsf{CSBias}_k(\widehat{\psi}_1) \leqslant \delta \, \mathsf{s.e.}(\widehat{\psi}_1).$$

will be explained in Remark 3.5 of Section 3. In principle, though, one can also consider testing $\mathsf{H}_{0,k,\mathsf{CS}}(\delta)$.

A preliminary version of the above idea has appeared in Liu, Mukherjee, and Robins (2020a), but the results therein are mainly for (1) the case when the distribution of the potentially high-dimensional covariates X is known, or the so-called *semisupervised* setting, and (2) one special case of DR functionals, the expected conditional covariance. In this article, we advance the literature in the following regards.

- First, we extend the results in Liu, Mukherjee, and Robins (2020a) to the more realistic case where Σ_k is unknown. In particular, we will show that a test statistic based on a third-order U-statistic (which is the third-order influence function of $\mathsf{Bias}_k(\widehat{\psi}_1)$) can be used to falsify the justification (1.1) with desired statistical guarantees without knowing the distribution of X.

- Second, the results in this paper require much weaker assumptions[6] on the basis functions $\bar{\mathsf{z}}_k$ than those in Liu, Mukherjee, and Robins (2020a), allowing the proposed methodology to be used more broadly in substantive studies.

- Third, we extend the results of Liu, Mukherjee, and Robins (2020a) for expected conditional covariance to the entire DR functional class, which requires that we derive the HOIFs of DR functionals in rather abstract notation. This could be of independent interest, considering the role of HOIFs in constructing rate-optimal estimators for smooth/differentiable functionals or their generalizations (Tchetgen Tchetgen et al., 2008; Kennedy, Balakrishnan, and Wasserman, 2022).

Literature overview. To the best of our knowledge, the assumption-lean falsification test as constructed in this paper is new in the literature. Thus we first mention a subset of the fast-growing literature in statistics and econometrics on estimating and drawing statistical inference for (certain members of) DR functionals using standard DML estimators; due to space limitation, see Farrell (2015), Chernozhukov et al. (2018), Smucler, Rotnitzky, and Robins (2019), Bradic, Wager, and Zhu (2019), and Chernozhukov et al. (2022) and references therein.

Some recent works also consider further refinement of standard DML estimators (Newey and Robins, 2018; Kline, Saggio, and Sølvsten, 2020; McGrath and Mukherjee, 2022; Bradic et al., 2019; Kennedy, 2020). The main distinction of these nonstandard DML estimators from the standard ones is to estimate b and p from separate subsamples of the training sample. Some of these nonstandard DML estimators are able to achieve semiparametric efficiency (Newey, 1990; Newey and McFadden, 1994) under weaker assumptions than standard DML estimators without resorting to CS inequality. Yet these nonstandard DML estimators often rely on carefully crafted nuisance function estimators $\widehat{b}$ and $\widehat{p}$ with special algebraic structures for the bias analysis to bypass the CS inequality step as in (1.1). When complex machine learning methods are used, it is unclear if these nonstandard estimators still work well both in theory and in practice. As a result, this article mainly focuses upon the standard DML estimators but these

---

[6]Note that both the assumptions in this paper and in Liu, Mukherjee, and Robins (2020a) are easily checkable by the analysts; see Conditions S and W.

nonstandard DML estimators will also be considered briefly, only in Section 4. In Remark 2.1, we also discuss several other works related to the methodology developed in our paper.

Plan. The remainder of the paper is arranged as follows. In Section 2, we describe the mathematical setup formally and review the definition and properties of DR functionals recently characterized in Rotnitzky, Smucler, and Robins (2021). In Section 2.1, we review the statistical properties of their standard DML estimators, based upon which we motivate and formally define our approach.

The main result of this paper, Section 3, is to construct a valid $\alpha^\dagger$-level falsification test of $\mathsf{H}_{0,\mathsf{CS}}(\delta)$ based on a third-order U-statistic, denoted as $\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})$ (see (3.5)), which is the third-order influence function of $\mathrm{Bias}_k(\widehat{\psi}_1)$ but with Σ_k replaced by $\widehat{\Sigma}_k$, following the notation used in Robins et al. (2008); also see Appendix A.6 for derivations.

In Section 4, we study if the proposed falsification test could be also meaningful for nonstandard DML estimators, whose bias could be $o(n^{-1/2})$ even when $\kappa_b + \kappa_p \leqslant 1/2$. We first argue that the k-projected null hypothesis $\mathsf{H}_{0,k}(\delta)$ is a natural hypothesis to falsify. Then we construct a valid $\alpha^\dagger$-level test of $\mathsf{H}_{0,k}(\delta)$ by modifying the falsification test in Section 3 using higher order U-statistics. Since higher order U-statistics are computational costly, we also propose an early-stopping strategy that takes the analyst's computational budget into account. In Section 5 we present results of simulation studies to evaluate the finite sample performance of our methods. Section 6 concludes with a discussion of some open problems. Many of the technical details are deferred to the online Appendix.

2. Formal setup and an illustration of our approach

The formal setup is as follows. We observe N i.i.d. copies of the data vector $O = (W, X)$ drawn from some unknown probability distribution P_θ belonging to a locally nonparametric model

$$\mathcal{M} = \{\mathsf{P}_\theta; \theta = (b, p, \theta \setminus \{b, p\}) \in \Theta = \mathcal{B} \times \mathcal{P} \times \Theta \setminus \{\mathcal{B}, \mathcal{P}\}\}$$

parameterized by the parameter $\theta = (b, p, \theta \setminus \{b, p\})$ with $b, p, \theta \setminus \{b, p\}$ variation independent. To stress the dependence on θ, from here on, we will attach θ to many symbols that have appeared, such as $\psi(\theta)$, E_θ, $\mathrm{Bias}_\theta(\widehat{\psi}_1)$, $\mathrm{CSBias}_\theta(\widehat{\psi}_1)$, $\mathrm{Bias}_{\theta,k}(\widehat{\psi}_1)$, etc. X is a d-dimensional random vector with compact support (with distribution function F) whose density f is bounded away from 0 and ∞ on its support, and d is allowed to increase with N. Here the maps $b : x \mapsto b(x) \in \mathcal{B}$ and $p : x \mapsto p(x) \in \mathcal{P}$ have range contained in $\mathbb{R}$. $\mathcal{M}$ is locally nonparametric in the sense that the tangent space for the model at each $\theta \in \Theta$ is equal to $L_2(\mathsf{P}_\theta)$, e.g. when b, p belong to Hölder balls with certain smoothness.

To avoid extraneous technical issues, we assume that observed data O is bounded with probability 1 (see Remark 3.2 for further discussion). We consider functionals (i.e. parameters) $\psi : \theta \mapsto \psi(\theta)$ that possess a (first order) influence function[7] $\mathsf{IF}_{1,\psi}(\theta) = \mathsf{if}_{1,\psi}(O; \theta)$ (and thus a positive and finite semiparametric

---

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

variance bound (**ai2012semiparametric**; Newey, 1990; Ichimura and Newey, 2022)) and are contained in the mixed bias or doubly robust class of functionals of Rotnitzky, Smucler, and Robins (2021) defined as follows. Under the locally nonparametric models defined above, the influence function $\mathsf{IF}_{1,\psi}(\theta)$ is unique.

Definition 2.1 (Definition of a mixed bias or doubly robust functional (DR functional) (Definition 1 of Rotnitzky, Smucler, and Robins (2021))). *$\psi(\theta)$ is a doubly robust 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 θ, θ'*

$$(2.1) \qquad \psi(\theta) - \psi(\theta') + \mathsf{E}_\theta\left[\mathsf{IF}_{1,\psi}(\theta')\right] = \mathsf{E}_\theta\left[S_{bp}(b(X) - b'(X))(p(X) - p'(X))\right]$$

where $S_{bp} \equiv s_{bp}(O)$ with $o \mapsto s_{bp}(o)$ a known function that does not depend on θ or θ' satisfying either $\mathsf{P}_\theta(S_{bp} \geqslant 0) = 1$ or $\mathsf{P}_\theta(S_{bp} \leqslant 0) = 1$. b and p are called nuisance functions in the semiparametric statistics literature, a terminology we also adopt in this paper. Finally, we let

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

and assume λ to be strictly bounded from above and below.

To understand the implication of DR functionals, let P_n be the empirical expectation operator over the estimation sample and suppose θ' were an estimate of θ from the training sample (regarded 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 (Scharfstein, Rotnitzky, and Robins, 1999; Robins and Rotnitzky, 2001; Bang and Robins, 2005). That is, by (2.1), it is unbiased for $\psi(\theta)$ under P_θ if either $b = b'$ or $p = p'$. Because of this fact, we will use the term DR functional in this paper, as is done in much of the current literature, instead of the "mixed bias" terminology employed in Rotnitzky, Smucler, and Robins (2021). 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 $\psi(\theta) = -\psi^\dagger(\theta)$.

Let $W = (Y, A)$. Below are some examples of DR functionals that are of substantive interest in economics and statistics.

(1) The counterfactual mean of Y when $\{0, 1\}$-valued A is set to 1 (under strong ignorability), $\mathsf{E}_\theta[Y(a = 1)]$ is identified from the distribution of $W = (A, AY, X)$ by the DR functional $\psi^\dagger(\theta) = \mathsf{E}_\theta[b(X)]$ with $b(x) = \mathsf{E}_\theta[Y|X = x, A = 1]$, $p(x) = 1/\mathsf{E}_\theta[A|X = x]$, $S_{bp} = -A$, and $\lambda(x) = -p(x)^{-1}$. 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[8]. Average treatment effect, one of the most intensively studied causal parameters, is thus a difference of two DR functionals.

(2) The expected conditional covariance

$$\psi(\theta) = \mathsf{E}_\theta[(Y - b(X))(A - p(X))]$$

with $b(x) = \mathsf{E}_\theta[Y|X = x], p(x) = \mathsf{E}_\theta[A|X = x]$ is a DR functional (equivalently, a MB functional) with $S_{bp} = \lambda(X) \equiv 1$. When Y and $b(X)$ are replaced by A and $p(X)$, $\psi(\theta)$ reduces to the expected conditional variance $\phi(\theta) = \mathsf{E}_\theta[(A - p(X))^2]$.

Before proceeding further, we collect some frequently used notation. We use $\mathsf{E}_\theta[\cdot], \mathsf{var}_\theta[\cdot]$ and etc. to denote the expectation, variance, and etc. with respect to P_θ. The data is randomly divided into an 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. 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, $\|V\| \equiv (V^{\otimes 2})^{1/2} = (V^\top V)^{1/2}$ its ℓ_2 norm and $\|V\|_\infty$ its $L_\infty(\mathsf{P}_\theta)$ norm. For any matrix M, $\|M\|$ will be reserved for its operator norm. Given an integer k, and a random vector $\bar{z}_k = \bar{z}_k(X)$, $\Pi_\theta[\cdot|\bar{z}_k]$ denotes the population linear projection operator onto the linear space spanned by $\bar{z}_k$ conditioning on the training sample, and $\Pi_\theta^\perp[\cdot|\bar{z}_k] = (\mathsf{I} - \Pi_\theta)[\cdot|\bar{z}_k]$ is the projection onto the ortho-complement of $\bar{z}_k$ in the Hilbert space $L_2(\mathsf{P}_F)$ where P_F denotes the marginal law of X under θ. That is, for a random variable V,

(2.2) $\quad \Pi_\theta[V|\bar{z}_k](x) = \bar{z}_k^\top(x)\{\mathsf{E}_\theta[\bar{z}_k(X)\bar{z}_k(X)^\top]\}^{-1}\mathsf{E}_\theta[\bar{z}_k(X)V], \Pi_\theta^\perp[V|\bar{z}_k](x) = V - \Pi_\theta[V|\bar{z}_k](x).$

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 P possibly depending on the sample size n, $X_n = O_\mathsf{P}(a_n)$ denotes that X_n/a_n is bounded in 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 ϵ.

2.1. Influence functions of DR functionals.

In this section, we generalize our discussion in the Introduction on $\psi(\theta) = -\mathsf{E}_\theta[Y(1)]$ under ignorability to the DR functionals and provide a more detailed exposition. We first review the influence functions of DR functionals (see Definition 2.1) established in Rotnitzky, Smucler, and Robins (2021). Although

---

[8]Chernozhukov, Newey, and Singh (2022a) used an alternative decomposition under which they included our A, X in their X. See Example 1 of Rotnitzky, Smucler, and Robins (2021).

the definition of DR functional is quite abstract, Rotnitzky, Smucler, and Robins (2021) derived the (non-parametric) influence functions of DR functionals by directly leveraging the form of the product bias appeared in Definition 2.1. We summarize their results below for the sake of completeness.

Proposition 2.1 (A summary of Theorems 1 and 2 of Rotnitzky, Smucler, and Robins (2021)). *Suppose (1) $\psi(\theta)$ for $\theta = (b, p, \theta \setminus \{b, p\}) \in \Theta \equiv \mathcal{B} \times \mathcal{P} \times \Theta \setminus \{\mathcal{B}, \mathcal{P}\}$ is a DR functional (equivalently MB functional) according to Definition 2.1; (2) the regularity Condition A.1 in Appendix A.1 (Condition 1 of Rotnitzky, Smucler, and Robins (2021)) holds; (3) $b(X)$, $p(X)$, $\lambda(X)b(X)$ and $\lambda(X)p(X)$ belong to $L_2(\mathsf{P}_F)$ for all $\theta \in \Theta$.*

Then there exists a statistic S_0 and linear 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 θ satisfying the following:

- *The influence function of $\psi(\theta)$ is given by:*

(2.3)
$$\mathsf{IF}_{1,\psi}(\theta) = \mathcal{H}(b, p) - \psi(\theta)$$

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

- *Furthermore, the following first-order doubly robust moment conditions hold:*

(2.4)
$$\begin{cases} \mathsf{E}_\theta[S_{bp}h(X)p(X) + m_1(O, h)] = 0 \text{ for all } h \in \mathcal{B}, \\ \mathsf{E}_\theta[S_{bp}b(X)h(X) + m_2(O, h)] = 0 \text{ for all } h \in \mathcal{P}. \end{cases}$$

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

- *Further suppose the maps $h \mapsto \mathsf{E}_\theta[m_1(O, h)]$ and $h \mapsto \mathsf{E}_\theta[m_2(O, h)]$ for $h \in L_2(\mathsf{P}_F)$ are continuous and linear with Riesz representers $\mathcal{R}_1(X) \equiv \mathcal{R}_1(X; \theta)$ and $\mathcal{R}_2(X) \equiv \mathcal{R}_2(X; \theta)$[9] respectively. Then $b(X) = -\mathcal{R}_2(X)/\lambda(X)$ and $p(X) = -\mathcal{R}_1(X)/\lambda(X)$; moreover, b and p are the (global) minimizers of the following minimization problems*

(2.5)
$$b(\cdot) = \arg\min_{h \in L_2(\mathsf{P}_F)} \mathsf{E}_\theta\left[S_{bp}\frac{h(X)^2}{2} + m_2(O, h)\right],$$
$$p(\cdot) = \arg\min_{h \in L_2(\mathsf{P}_F)} \mathsf{E}_\theta\left[S_{bp}\frac{h(X)^2}{2} + m_1(O, h)\right].$$

We first give some examples of the influence functions $\mathsf{IF}_{1,\psi}(\theta)$ and Riesz representers of DR functionals. Similar moment conditions such as (2.4) can be traced back to **newey1994asymptotic; newey1997convergence**; Newey, Hsieh, and Robins (1998), Newey, Hsieh, and Robins (2004), and Chernozhukov, Newey, and Singh (2022a).

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

---

[9]The Riesz representer $\mathcal{R}(X)$ of a continuous linear functional $h \mapsto \mathsf{E}[m(O, h)]$ for $h \in L_2(\mathsf{P}_F)$ is, by definition, the function of X satisfying $\mathsf{E}[m(O, h)] = \mathsf{E}[\mathcal{R}(X)h(X)]$; see Chernozhukov, Newey, and Singh (2022b), Chernozhukov, Newey, and Singh (2022a), and Rotnitzky, Smucler, and Robins (2021).

(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

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

(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

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

We now briefly comment on the three parts of Proposition 2.1 as they are all important for future development of the paper. Eq. (2.3) exhibits the general formula of influence functions of DR functionals, which are the basic building blocks for standard DML estimators and most nonstandard DML estimators. Also, our framework heavily relies on higher order influence functions, which are derived from the (first-order) influence functions. Eq. (2.4) and (2.5) together suggest natural loss functions that could be used to fit the nuisance functions b and p from data. To see why, we first make the following important observation:

Lemma 2.1. *The minimization problems given in (2.5), with the function class $L_2(\mathsf{P}_F)$ replaced by some $\mathcal{F} \subset L_2(\mathsf{P}_F)$*

(2.6)
$$\widetilde{b}(\cdot) = \arg\min_{h \in \mathcal{F}} \mathsf{E}_\theta\left[S_{bp}\frac{h(X)^2}{2} + m_2(O,h)\right],$$
$$\widetilde{p}(\cdot) = \arg\min_{h \in \mathcal{F}} \mathsf{E}_\theta\left[S_{bp}\frac{h(X)^2}{2} + m_1(O,h)\right]$$

are equivalent to:

(2.7)
$$\widetilde{b}(\cdot) = \arg\min_{h \in \mathcal{F}} \mathsf{E}_\theta\left[\lambda(X)(b(X) - h(X))^2\right],$$
$$\widetilde{p}(\cdot) = \arg\min_{h \in \mathcal{F}} \mathsf{E}_\theta\left[\lambda(X)(p(X) - h(X))^2\right].$$

The proof of Lemma 2.1 can be found in Appendix A.2. Based on Lemma 2.1, one can often establish rates of convergence of $\widehat{b}, \widehat{p}$ to b, p in $L_2(\mathsf{P}_\theta)$ norm by solving the following minimization problem from

the training sample:

$$
(2.8) \quad \begin{aligned}
\widehat{b} &= \arg\min_{h \in \mathcal{F}} \mathsf{P}_{n_{\mathrm{tr}}} \left[S_{bp} \frac{h^2}{2} + m_2(O, h) \right], \\
\widehat{p} &= \arg\min_{h \in \mathcal{F}} \mathsf{P}_{n_{\mathrm{tr}}} \left[S_{bp} \frac{h^2}{2} + m_1(O, h) \right].
\end{aligned}
$$

where $\mathcal{F}$ is the set of functions computable by the machine learning algorithm. This is because the convergence properties of $\widehat{b}$ and $\widehat{p}$ are often established by excess risk bound that connects the empirical loss (2.8) to the expected λ-weighted $L_2(\mathsf{P}_\theta)$-loss (2.6) under certain complexity reducing assumptions. In Appendix A.2, we also point out several other possible choices, including unweighted $L_2(\mathsf{P}_\theta)$-loss, cross-entropy loss, and adversarial loss minimization strategies. Importantly, given positive weight functions w_b, w_p over X, that are strictly bounded from above and below, any w_b- (or w_p-) weighted $L_2(\mathsf{P}_\theta)$-loss and λ-weighted $L_2(\mathsf{P}_\theta)$-loss are equivalent up to constants under our assumption on λ in Definition 2.1:

$$
(2.9) \quad \inf_x \frac{\lambda(x)}{w_b(x)} \mathsf{E}_\theta[w_b(X)\{\widehat{b}(X) - b(X)\}^2] \leqslant \mathsf{E}_\theta[\lambda(X)\{\widehat{b}(X) - b(X)\}^2] \leqslant \sup_x \frac{\lambda(x)}{w_b(x)} \mathsf{E}_\theta[w_b(X)\{\widehat{b}(X) - b(X)\}^2]
$$

and similarly for $\mathsf{E}_\theta[\lambda(X)\{\widehat{p}(X) - p(X)\}^2]$. For DR functionals, $\mathsf{CSBias}_\theta(\widehat{\psi}_1)$, the Cauchy-Schwarz bias of $\widehat{\psi}_1$, is defined as

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

which is an upper bound of $\mathsf{Bias}_\theta(\widehat{\psi}_1)$ to be defined in (2.12) by CS inequality. Hence we have

$$
(2.11) \quad \mathsf{CSBias}_\theta(\widehat{\psi}_1) \asymp \mathsf{CSBias}_\theta^w(\widehat{\psi}_1) \equiv \{\mathsf{E}_\theta[w_b(X)\{\widehat{b}(X) - b(X)\}^2]\mathsf{E}_\theta[w_p(X)\{\widehat{p}(X) - p(X)\}^2]\}^{1/2},
$$

that is, they are equivalent up to a multiplicative positive constant.

2.2. Standard DML estimators.

The following algorithm defines the standard DML estimators $\widehat{\psi}_1$ and $\widehat{\psi}_{\mathrm{cf},1}$ of a general DR functional $\psi(\theta)$ satisfying the regularity conditions given in Proposition 2.1.

(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 $i = 1, \dots, n$ corresponds to the estimation sample.

(ii) Nuisance estimators $\widehat{b}$ and $\widehat{p}$ are separately constructed from the entire training sample data using machine learning, such as deep neural networks (Farrell, Liang, and Misra, 2021; Chen et al., 2020) and define $\widehat{\theta} := (\widehat{b}, \widehat{p}, \theta \setminus \{b, p\})$.

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

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

13

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.

The next proposition provides asymptotic properties of standard DML estimators $\widehat{\psi}_1$ (and $\widehat{\psi}_{\mathsf{cf},1}$) of DR functionals. Its proof is straightforward and can be found in Chernozhukov et al. (2018), Smucler, Rotnitzky, and Robins (2019) or Farrell, Liang, and Misra (2021, Theorem 3).

Proposition 2.2. *Under the conditions of Proposition 2.1, conditional on the training sample, if a) the bias of $\widehat{\psi}_1$*

$$(2.12) \qquad \mathsf{Bias}_\theta(\widehat{\psi}_1) := \mathsf{E}_\theta[S_{bp}(b(X) - \widehat{b}(X))(p(X) - \widehat{p}(X))] \equiv \mathsf{E}_\theta[\lambda(X)(b(X) - \widehat{b}(X))(p(X) - \widehat{p}(X))]$$

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) *$\widehat{\psi}_1 - \psi(\theta) = n^{-1}\sum_{i=1}^n \mathsf{IF}_{1,\psi,i}(\theta) + o(n^{-1/2})$ and $\widehat{\psi}_{\mathsf{cf},1} - \psi(\theta) = N^{-1}\sum_{i=1}^N \mathsf{IF}_{1,\psi,i}(\theta) + o_p(N^{-1/2})$.*
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 $\mathsf{var}_\theta[\mathsf{IF}_{1,\psi}(\theta)]$.

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

$$(2.13) \qquad \mathsf{CI}_\alpha(\widehat{\psi}_1) := \widehat{\psi}_1 \pm z_{\alpha/2}\widehat{\mathsf{s.e.}}[\widehat{\psi}_1], \ \mathsf{CI}_\alpha(\widehat{\psi}_{\mathsf{cf},1}) := \widehat{\psi}_{\mathsf{cf},1} \pm z_{\alpha/2}\widehat{\mathsf{s.e.}}[\widehat{\psi}_{\mathsf{cf},1}]$$

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

$$(2.14) \qquad \widehat{\mathsf{var}}[\widehat{\psi}_1] = \frac{1}{n^2}\sum_{i=1}^n\left(\mathsf{IF}_{1,\psi,i}(\widehat{\theta}) - \frac{1}{n}\sum_{i=1}^n\mathsf{IF}_{1,\psi,i}(\widehat{\theta})\right)^2$$

and $\widehat{\mathsf{s.e.}}[\widehat{\psi}_{\mathsf{cf},1}] = \frac{1}{2}\{\widehat{\mathsf{var}}[\widehat{\psi}_1] + \widehat{\mathsf{var}}[\overline{\widehat{\psi}}_1]\}^{1/2}$. The interval $\mathsf{CI}_\alpha(\widehat{\psi}_1)$, unlike $\mathsf{CI}_\alpha(\widehat{\psi}_{\mathsf{cf},1})$, is also an asymptotically valid $(1 - alpha)$ CI conditional on the training sample.

2.3. **A prelude to our approach.**

As briefly described in Section 1, our main goal is to develop assumption-lean empirical methods to falsify $\mathsf{NH}_{0,\mathsf{CS}} : \mathsf{CSBias}_\theta(\widehat{\psi}_1) = o(n^{-1/2})$ by falsifying its operationalized pair $\mathsf{H}_{0,\mathsf{CS}}(\delta) : \mathsf{CSBias}_\theta(\widehat{\psi}_1) < \mathsf{s.e.}_\theta(\widehat{\psi}_1)\delta$.

However, as we argued in the Introduction, $\mathsf{CSBias}_\theta(\widehat{\psi}_1)$ and $\mathsf{Bias}_\theta(\widehat{\psi}_1)$ are generally not uniformly consistently estimable without complexity reducing assumptions on b and p or their estimates $\widehat{b}$ and $\widehat{p}$. This is evident from the following decomposition of $\mathsf{Bias}_\theta(\widehat{\psi}_1)$ because we do not have control over $\Pi_{\bar\theta}^\perp[\lambda^{1/2}(\widehat{b} - b)|\lambda^{1/2}\bar{\mathsf{z}}_k]$ and $\Pi_{\bar\theta}^\perp[\lambda^{1/2}(\widehat{p} - p)|\lambda^{1/2}\bar{\mathsf{z}}_k]$ without restrictive complexity-reducing assumptions on b and p or on $\widehat{b}$ and $\widehat{p}$:

$$(2.15) \qquad \mathsf{Bias}_\theta(\widehat{\psi}_1) = \mathsf{Bias}_{\theta,k}(\widehat{\psi}_1) + \mathsf{TB}_{\theta,k}(\widehat{\psi}_1)$$

14

where $\mathsf{TB}_{\theta,k}(\widehat{\psi}_1) \coloneqq \mathsf{E}_\theta[\Pi_\theta^\perp[\lambda^{1/2}(\widehat{b}-b)|\lambda^{1/2}\bar{\mathsf{z}}_k](X)\Pi_\theta^\perp[\lambda^{1/2}(\widehat{p}-p)|\lambda^{1/2}\bar{\mathsf{z}}_k](X)]$, which was referred to as the truncation bias in Robins et al. (2008) and Liu, Mukherjee, and Robins (2020a) (also see Appendix A.6), and

$$(2.16) \quad \begin{aligned} \mathsf{Bias}_{\theta,k}(\widehat{\psi}_1) &\coloneqq \mathsf{E}_\theta[\Pi_\theta[\lambda^{1/2}(\widehat{b}-b)|\lambda^{1/2}\bar{\mathsf{z}}_k](X)\Pi_\theta[\lambda^{1/2}(\widehat{p}-p)|\lambda^{1/2}\bar{\mathsf{z}}_k](X)] \\ &\equiv \mathsf{E}_\theta[\lambda(X)(\widehat{b}(X)-b(X))\bar{\mathsf{z}}_k(X)]^\top \Sigma_k^{-1} \mathsf{E}_\theta[\bar{\mathsf{z}}_k(X)\lambda(X)(\widehat{p}(X)-p(X))]. \end{aligned}$$

Here $\bar{\mathsf{z}}_k$ is a vector of k-dimensional vector of basis functions chosen by the analyst satisfying mild regularity conditions (see Condition W in Section 3) and $\Sigma_k \coloneqq \mathsf{E}_\theta[\lambda(X)\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top] \equiv \mathsf{E}_\theta[S_{bp}\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top]$[10] is the population Gram matrix of $\lambda^{1/2}\bar{\mathsf{z}}_k$.

But fortunately, as can be shown in the same fashion as in (1.1), the quantity $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$ is a lower bound of $\mathsf{CSBias}_\theta(\widehat{\psi}_1)$. The test statistic of $\mathsf{H}_{0,\mathsf{CS}}(\delta)$, later introduced in Section 3, is based on estimators of $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$, derived using the theory of higher-order influence functions (HOIFs) (which are higher-order U-statistics) developed in a series of papers by some of the authors (Robins et al., 2008; Robins et al., 2017; Liu et al., 2017). Due to space limitation, we refer the interested readers to Robins et al. (2008) or van der Vaart (2014) for a more comprehensive review. *En route* to constructing these HOIF estimators and associated tests, we require access to only the study data and the functions $\widehat{b}$ and $\widehat{p}$ obtained by analysts. However, our tests are constructed without: i) refitting, modifying, or even having knowledge of the machine learning 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 nuisance functions b or p. The key to achieve i) is by conditioning on the training sample data. For a related discussion, see Remark 3.4.

Formally, our proposed approach begins by noticing that $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$ can be unbiasedly estimated by the following infeasible second order U-statistic

$$(2.17) \quad \begin{aligned} \widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1}) &\coloneqq \frac{1}{n(n-1)} \sum_{1 \leqslant i_1 \neq i_2 \leqslant n} \left[\mathcal{E}_{\widehat{b},m_2}(\bar{\mathsf{z}}_k)(O)\right]_{i_1}^\top \Sigma_k^{-1} \left[\mathcal{E}_{\widehat{p},m_1}(\bar{\mathsf{z}}_k)(O)\right]_{i_2}, \text{ where} \\ \mathcal{E}_{\widehat{b},m_2}(\bar{\mathsf{z}}_k)(O) &\coloneqq S_{bp}\widehat{b}(X)\bar{\mathsf{z}}_k(X) + m_2(O,\bar{\mathsf{z}}_k), \mathcal{E}_{\widehat{p},m_1}(\bar{\mathsf{z}}_k)(O) \coloneqq S_{bp}\widehat{p}(X)\bar{\mathsf{z}}_k(X) + m_1(O,\bar{\mathsf{z}}_k) \end{aligned}$$

with standard error of order $\frac{\sqrt{k}}{n} \vee \frac{1}{\sqrt{n}}$. In fact, $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ is the second-order influence function of $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$[11]; see Appendix A.6 for more details. $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ is infeasible because in general Σ_k^{-1} is unknown. The unbiasedness follows directly from equation (2.4) in Proposition 2.1; the proof of the standard error bound of $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ is deferred to Appendix A.3, and can also be found in Liu, Mukherjee, and Robins (2020a), which mostly focused on the infeasible case by assuming Σ_k^{-1} to be known. Since Σ_k^{-1} is generally unknown, we propose to estimate Σ_k^{-1} by $\widehat{\Sigma}_k^{-1}$ from the training sample data, where

---

[10]To relate to the discussion in Section 1, $S_{bp} = A$ for the functional $\psi(\theta) = -\mathsf{E}_\theta[Y(1)]$.

[11]This is why we adopt the notation $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ by following the convention in Robins et al. (2008).

$\widehat{\Sigma}_k := \mathsf{P}_{n_{\mathrm{tr}}}[S_{bp}\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top]$ is simply the empirical Gram matrix estimator. In particular, throughout this paper, we choose $1 \ll k \ll n$. The reason for this choice is discussed in detail in the following remark.

Remark 2.1 (On the choice of k). *We shall always choose k to be much less than the sample size $n = N/2$ of the estimation sample, for the standard error of $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ of order 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 $|\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)|/\mathsf{s.e.}_\theta(\widehat{\psi}_1) \geqslant \delta$, when the sample size n is sufficiently large.*

Moreover, we also need $k \ll n_{\mathrm{tr}} = n$, the sample size of the training sample, to ensure $\widehat{\Sigma}_k^{-1}$ to be a operator-norm consistent estimator of Σ_k^{-1}, without imposing, possibly incorrect, additional smoothness assumption on the distribution of (S_{bp}, X) or sparsity assumptions on Σ_k or Σ_k^{-1} required for accurate estimation when $k \geqslant n$.

Finally, we take $k \gg 1$ to ensure the asymptotic normality of the U-statistic $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$ (after appropriate scaling) (Bhattacharya and Ghosh, 1992) to determine the rejection region of our falsification test to be studied in the next section.

One might view the choice of $k \ll n$ a disadvantage. On the surface, it seems to exclude high-dimensional statistical models. However, this is not the case: (1) The nuisance functions are allowed to be estimated by any procedure, be it high-dimensional or not, from the training sample; (2) $k \ll n$ ensures that our testing procedure can be a valid test without any complexity-reducing assumptions on the nuisance functions, including the distribution of X, that is required to evaluate Σ_k^{-1}. We cannot let $k \asymp n$ because, as we will see, our test needs the bias due to estimating Σ_k by $\widehat{\Sigma}_k$ to be sufficiently small and recent developments along this direction (Cattaneo, Jansson, and Newey, 2018; Kline, Saggio, and Sølvsten, 2020; Yadlowsky, 2022; Jiang et al., 2022; Jochmans, 2022) cannot be used because of the more stringent assumptions on $\bar{\mathsf{z}}_k$, such as sub-Gaussian tail conditions. The jackknife or bootstrap bias correction methods developed by Cattaneo, Jansson, and Ma (2019) and Cattaneo and Jansson (2018) could potentially be applied to our problem but since they only allow $k = O(\sqrt{n})$, the chance of rejecting $\mathsf{H}_{0,\mathsf{CS}}(\delta)$ could be diminished. It will be an interesting problem to study if their approach can be extended to allow $\sqrt{n} \ll k \ll n$, so as to be applied to our problem. We list this as an open problem in Section 6.

3. Main results: A valid falsification test of null hypothesis $\mathsf{H}_{0,\mathsf{CS}}(\delta)$

Having introduced $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$ as an estimator of $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$, we are now prepared to construct a valid nominal level-$\alpha^\dagger$ falsification test of the non-asymptotic null hypothesis paired with $\mathsf{NH}_{0,\mathsf{CS}}$:

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

3.1. **What constitutes a valid test of $H_{0,CS}(\delta)$?**

Before proceeding to the main result, we note that the infeasible falsification test of $H_{0,CS}(\delta)$ proposed in Liu, Mukherjee, and Robins (2020a) can be generalized to the class of DR functionals as follows[12]:

$$(3.1) \qquad \widehat{\chi}_{2,k}(\Sigma_k^{-1}; \varsigma_k, \delta) = \mathbb{1}\left\{ \frac{|\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})|}{\widehat{\text{s.e.}}[\widehat{\psi}_1]} - \varsigma_k \frac{\widehat{\text{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]}{\widehat{\text{s.e.}}[\widehat{\psi}_1]} \geqslant \delta \right\}$$

where $\widehat{\text{s.e.}}[\widehat{\psi}_1]$ (see (2.14)) and $\widehat{\text{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]$ (see (A.19) in Appendix A.8) are estimators of the standard errors $\text{s.e.}_\theta[\widehat{\psi}_1]$ and $\text{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]$. Here one choose the cut-off $\varsigma_k = z_{\alpha^\dagger/2}$ by normal approximation. The asymptotic validity of $\widehat{\chi}_{2,k}(\Sigma_k^{-1}; z_{\alpha^\dagger/2}, \delta)$ relies on the unbiasedness of $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$, but the unbiasedness can be relaxed as in the proposition below.

Proposition 3.1. *Given any estimator $\widehat{\text{Bias}}_{\theta,k}(\widehat{\psi}_1)$ of $\text{Bias}_{\theta,k}(\widehat{\psi}_1)$ that, conditioning on the training sample data, satisfies:*

(1) Under $H_{0,CS}(\delta)$: $\mathsf{E}_\theta[\widehat{\text{Bias}}_{\theta,k}(\widehat{\psi}_1) - \text{Bias}_{\theta,k}(\widehat{\psi}_1)] = o\left(\frac{\sqrt{k}}{n}\right)$;

(2) $\text{s.e.}_\theta[\widehat{\text{Bias}}_{\theta,k}(\widehat{\psi}_1)] \asymp \text{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]$;

(3) Under $H_{0,CS}(\delta)$: $\frac{\widehat{\text{Bias}}_{\theta,k}(\widehat{\psi}_1) - \text{Bias}_{\theta,k}(\widehat{\psi}_1)}{\text{s.e.}_\theta[\widehat{\text{Bias}}_{\theta,k}(\widehat{\psi}_1)]} \xrightarrow{d} N(0,1)$,

Then, again conditioning on the training sample data, the test

$$(3.2) \qquad \widehat{\chi}_k(\varsigma_k, \delta) = \mathbb{1}\left\{ \frac{|\widehat{\text{Bias}}_{\theta,k}(\widehat{\psi}_1)|}{\widehat{\text{s.e.}}[\widehat{\psi}_1]} - \varsigma_k \frac{\widehat{\text{s.e.}}[\widehat{\text{Bias}}_{\theta,k}(\widehat{\psi}_1)]}{\widehat{\text{s.e.}}[\widehat{\psi}_1]} \geqslant \delta \right\},$$

where $\widehat{\text{s.e.}}[\widehat{\text{Bias}}_{\theta,k}(\widehat{\psi}_1)]$ is a consistent estimator of $\text{s.e.}_\theta[\widehat{\text{Bias}}_{\theta,k}(\widehat{\psi}_1)]$, rejects $H_{0,CS}(\delta)$ with probability less than $\alpha^\dagger$ as $n \to \infty$ with $\varsigma_k = z_{\alpha^\dagger/2}$ under $H_{0,CS}(\delta)$. That is, we say that $\widehat{\chi}_k(z_{\alpha^\dagger/2}, \delta)$ is an asymptotically valid test of $H_{0,CS}(\delta)$.

The above statement also holds if one replaces $H_{0,CS}(\delta)$ by the k-projected null hypothesis $H_{0,k}(\delta)$.

We did not give a formal proof as the argument for Proposition 3.1 is quite simple. Since $\text{s.e.}_\theta[\widehat{\text{Bias}}_{\theta,k}(\widehat{\psi}_1)]$ can be as small as of order $\frac{\sqrt{k}}{n}$, one needs the bias $\mathsf{E}_\theta[\widehat{\text{Bias}}_{\theta,k}(\widehat{\psi}_1) - \text{Bias}_{\theta,k}(\widehat{\psi}_1)]$ to be dominated by this order under the null $H_{0,CS}(\delta)$ or the null $H_{0,k}(\delta)$ to protect the level of the test. The final clause of Proposition 3.1 will be most relevant to Section 4.

3.2. **Bias-reduced estimator of $\text{Bias}_{\theta,k}(\widehat{\psi}_1)$.**

The most natural estimator of $\widehat{\text{Bias}}_{\theta,k}(\widehat{\psi}_1)$ is $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$ and the corresponding test is

$$(3.3) \qquad \widehat{\chi}_{2,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^\dagger/2}, \delta) = \mathbb{1}\left\{ \frac{|\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})|}{\widehat{\text{s.e.}}[\widehat{\psi}_1]} - z_{\alpha^\dagger/2} \frac{\widehat{\text{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})]}{\widehat{\text{s.e.}}[\widehat{\psi}_1]} \geqslant \delta \right\}.$$

---

[12]We also provide relevant results in Appendix A.3.

However, we *cannot* prove that the above test is asymptotically valid for $\mathsf{H}_{0,\mathsf{CS}}(\delta)$, because Theorem 3.1 below will show the bias of $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$ for estimating $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$ is

$$(3.4) \qquad \mathsf{EB}_{\theta,k,2} := \mathsf{E}_\theta[\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1}) - \mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)] = O\left(\frac{\sqrt{k}}{n}\sqrt{\log k}\right)$$

under $\mathsf{H}_{0,\mathsf{CS}}(\delta)$. This obtained upper bound $O\left(\frac{\sqrt{k\log k}}{n}\right)$ of the bias due to estimating Σ_k^{-1} exceeds that which is needed to protect the level of the test $\widehat{\chi}_{2,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^\dagger/2}, \delta)$, as given in Condition (1) of Proposition 3.1.

Fortunately, Theorem 3.1 will also show that the following third-order U-statistic estimator of $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$

$$(3.5)$$
$$\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1}) := \widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1}) + \widehat{\mathbb{IF}}_{33,k}(\widehat{\Sigma}_k^{-1}), \text{ where}$$

$$\widehat{\mathbb{IF}}_{33,k}(\widehat{\Sigma}_k^{-1}) := \frac{(n-2)!}{n!} \sum_{1\leqslant i_1 \neq i_2 \neq i_3 \leqslant n} \left[\mathcal{E}_{\widehat{b},m_2}(\bar{z}_k)(O)\right]_{i_1}^\top \widehat{\Sigma}_k^{-1} \left[S_{bp}\bar{z}_k(X)\bar{z}_k(X)^\top - \widehat{\Sigma}_k\right]_{i_3} \widehat{\Sigma}_k^{-1} \left[\mathcal{E}_{\widehat{p},m_1}(\bar{z}_k)(O)\right]_{i_2}$$

can reduce the upper bound of the bias due to estimating Σ_k^{-1} from $O\left(\frac{\sqrt{k\log k}}{n}\right)$ in (3.4) to

$$(3.6) \qquad \mathsf{EB}_{\theta,k,3} := \mathsf{E}_\theta[\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1}) - \mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)] = O\left(\frac{\sqrt{k}}{n}\sqrt{\frac{k\log k}{n}}\right) = o\left(\frac{\sqrt{k}}{n}\right)$$

as long as we choose k such that $k\log k = o(n)$.

Remark 3.1. $\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})$ *can be viewed as a de-biased version of* $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$, *and the bias is partially corrected by* $\widehat{\mathbb{IF}}_{33,k}(\widehat{\Sigma}_k^{-1})$. *As will be shown in Appendix A.6,* $\widehat{\mathbb{IF}}_{22\to33,k}(\Sigma_k^{-1})$ *is the third-order influence function of* $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$ *and* $\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})$ *is simply its feasible version.*

To avoid clutter in the remainder of this paper, we introduce the following additional notation for various L_q-type norms of functions or their weighted $L_2(\mathsf{P}_\theta)$-projections, for $q = 2, 4$:

$$\mathbb{L}_{\theta,2,\widehat{b}} := \left\{\mathsf{E}_\theta[\lambda(X)(\widehat{b}(X) - b(X))^2]\right\}^{1/2}, \mathbb{L}_{\theta,q,\widehat{b},k} := \left\{\mathsf{E}_\theta[\{\Pi_\theta[\lambda^{1/2}(\widehat{b} - b)|\lambda^{1/2}\bar{z}_k](X)\}^q]\right\}^{1/q},$$

$$\mathbb{L}_{\theta,2,\widehat{p}} := \left\{\mathsf{E}_\theta[\lambda(X)(\widehat{p}(X) - p(X))^2]\right\}^{1/2}, \mathbb{L}_{\theta,q,\widehat{p},k} := \left\{\mathsf{E}_\theta[\{\Pi_\theta[\lambda^{1/2}(\widehat{p} - p)|\lambda^{1/2}\bar{z}_k](X)\}^q]\right\}^{1/q},$$

$$\mathbb{L}_{\theta,\infty,\widehat{b},k} := \|\Pi_\theta[\lambda^{1/2}(\widehat{b} - b)|\lambda^{1/2}\bar{z}_k]\|_\infty, \mathbb{L}_{\theta,\infty,\widehat{p},k} := \|\Pi_\theta[\lambda^{1/2}(\widehat{p} - p)|\lambda^{1/2}\bar{z}_k]\|_\infty, \mathbb{L}_{\theta,2,\widehat{\Sigma},k} := \|\widehat{\Sigma}_k - \Sigma_k\|,$$

which will also be useful for the later development of this paper. Note that we define $\mathsf{CSBias}_{\theta,k}(\widehat{\psi}_1) \equiv \mathbb{L}_{\theta,2,\widehat{b},k}\mathbb{L}_{\theta,2,\widehat{p},k}$, generalizing the definition of $\mathsf{CSBias}_{\theta,k}(\widehat{\psi}_1)$ given in (1.5) to the whole class of DR functionals.

We also need to further impose the following weak regularity conditions (Condition W).

Condition W.

(1) *All the eigenvalues of Σ_k are bounded away from 0 and ∞.*

(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 P_θ-probability 1, and $\lambda(X)$ are bounded away from 0 and ∞ with P_θ-probability 1.*

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

Note that Condition W should also be compared to the following slightly stronger Condition S in Liu, Mukherjee, and Robins (2020a), which is the same as Condition W except (3) shall be replaced by the following (3′):

Condition S.

(3′) *$\|\bar{\mathsf{z}}_k^\top \bar{\mathsf{z}}_k\|_\infty \leqslant Bk$ for some constant $B > 0$; in addition, $\mathbb{L}_{\theta,\infty,\widehat{b},k} < \infty$ and $\mathbb{L}_{\theta,\infty,\widehat{p},k} < \infty$.*

Remark 3.2. *Condition S(3′) is stronger than Condition W(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 or Legendre polynomial series of X when X is compactly supported (Belloni et al., 2015). We also refer interested readers to Kennedy, Balakrishnan, and Wasserman (2020, Section 6) and Liu, Mukherjee, and Robins (2020b, Section 2.1) for further motivation on why we decide to relax Condition S by Condition W. However, as we will see, when Σ_k is unknown and need to be estimated from the training sample, violation of Condition S(3′) but not Condition W(3), will incur a loss in the power of the test (the magnitude depending on the particular series used; see Appendix A.7), but the validity of the test (i.e. level) is still guaranteed. Nevertheless, such loss in power happens (or equivalently, Condition W holds but Condition S fails to hold), only if the L_∞-norms of the projections $\Pi_\theta[\lambda^{1/2}(\widehat{b}-b)|\lambda^{1/2}\bar{\mathsf{z}}_k]$ and $\Pi_\theta[\lambda^{1/2}(\widehat{p}-p)|\lambda^{1/2}\bar{\mathsf{z}}_k]$ are not bounded even though those of $\lambda^{1/2}(\widehat{b}-b)$ and $\lambda^{1/2}(\widehat{p}-p)$ are bounded.*

Finally, we summarize the 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 Theorem 3.1 below:

Theorem 3.1. *Under the conditions of Proposition 2.1 and Condition W, with $k, n \to \infty$ but $k\log k = o(n)$, conditioning on the training sample, the following hold on the event that $\widehat{\Sigma}_k$ is invertible:*

(1) *The bias and variance of $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$ as an estimator of $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$ are of the following order:*

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

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

where $\mathsf{EB}_{\theta,2,k}$ is defined in equation (3.4).

19

(2) *The bias and variance of $\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1}) := \widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1}) + \widehat{\mathbb{IF}}_{33,k}(\widehat{\Sigma}_k^{-1})$ are of the following order:*

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

$$\mathsf{var}_\theta\left[\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})\right] \lesssim \frac{1}{n}\left\{\frac{k}{n} + \mathbb{L}_{\theta,2,\widehat{b},k}^2 + \mathbb{L}_{\theta,2,\widehat{p},k}^2 + k\mathbb{L}_{\theta,2,\widehat{b},k}^2\mathbb{L}_{\theta,2,\widehat{p},k}^2\right\}$$

where $\mathsf{EB}_{\theta,3,k}$ is defined in equation (3.4).

If, however, Condition W is strengthened to Condition S

$$\mathsf{var}_\theta\left[\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})\right] \lesssim \frac{1}{n}\left\{\frac{k}{n} + \mathbb{L}_{\theta,2,\widehat{b},k}^2 + \mathbb{L}_{\theta,2,\widehat{p},k}^2\right\}.$$

(3) *Under $\mathsf{H}_{0,\mathsf{CS}}(\delta)$:*

$$\frac{\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1}) - \mathsf{E}_\theta\left[\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})\right]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})]} \text{ and } \frac{\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1}) - \mathsf{E}_\theta\left[\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})\right]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})]} \xrightarrow{d} N(0,1).$$

When we estimate $\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$ and $\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})$ respectively by their consistent bootstrap estimators $\widehat{\mathsf{s.e.}}.[\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$ and $\widehat{\mathsf{s.e.}}.[\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})$ defined in (A.19) and (A.22), with Σ_k replaced by $\widehat{\Sigma}_k$, in Appendix A.8, we also have

$$\frac{\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1}) - \mathsf{E}_\theta\left[\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})\right]}{\widehat{\mathsf{s.e.}}.[\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})]} \text{ and } \frac{\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1}) - \mathsf{E}_\theta\left[\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})\right]}{\widehat{\mathsf{s.e.}}.[\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})]} \xrightarrow{d} N(0,1).$$

Remark 3.3.

- *The bias and variance bounds under Condition S have been proved in Liu et al. (2017). The change in the proof under Condition W is given in Lemma A.3 in Appendix A.7. The normal approximation under $\mathsf{H}_{0,\mathsf{CS}}(\delta)$ follows from Slutsky theorem and asymptotic normality of U-statistics with diverging (with n) kernels given in Theorem 1 of Bhattacharya and Ghosh (1992).*

- *The generic bias bounds in Theorem 3.1 imply those stated in (3.4) and (3.6) under $\mathsf{H}_{0,\mathsf{CS}}(\delta)$, as*

$$\mathbb{L}_{\theta,2,\widehat{b},k}\mathbb{L}_{\theta,2,\widehat{p},k} \leqslant \mathbb{L}_{\theta,2,\widehat{b}}\mathbb{L}_{\theta,2,\widehat{p}} = \mathsf{CSBias}_\theta(\widehat{\psi}_1) \lesssim n^{-1/2}$$

when $\mathsf{H}_{0,\mathsf{CS}}(\delta)$ is true.

3.3. **The proposed assumption-lean falsification test of $\mathsf{H}_{0,\mathsf{CS}}(\delta)$.**

All the previous discussions in this section culminate in (1) the following test of $\mathsf{H}_{0,\mathsf{CS}}(\delta)$:

$$(3.7) \qquad \widehat{\chi}_{3,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^\dagger/2}, \delta) := \mathbb{1}\left\{\frac{|\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})|}{\widehat{\mathsf{s.e.}}.[\widehat{\psi}_1]} - z_{\alpha^\dagger/2}\frac{\widehat{\mathsf{s.e.}}.[\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})]}{\widehat{\mathsf{s.e.}}.[\widehat{\psi}_1]} \geqslant \delta\right\},$$

and (2) the following theorem showing its asymptotically validity:

Theorem 3.2. *Assume all the conditions in Theorem 3.1. Under* $\mathsf{H}_{0,\mathsf{CS}}(\delta) : \frac{\mathsf{CSBias}_\theta(\widehat{\psi}_1)}{\mathsf{s.e.}_\theta(\widehat{\psi}_1)} \leqslant \delta,$

$$(3.8) \qquad\qquad \lim_{n\to\infty} \mathbb{P}_\theta\left(\widehat{\chi}_{3,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^\dagger/2}, \delta) = 1\right) \leqslant \alpha^\dagger.$$

That is, $\widehat{\chi}_{3,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^\dagger/2}, \delta)$ *is an asymptotically level-*$\alpha^\dagger$ *test of* $\mathsf{H}_{0,\mathsf{CS}}(\delta)$.

The proof of Theorem 3.2 is a direct consequence of Theorem 3.1 and is deferred to Appendix A.9. Since $\widehat{\chi}_{3,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^\dagger/2}, \delta)$ is not a consistent test, we say $\widehat{\chi}_{3,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^\dagger/2}, \delta)$ can only falsify the null hypothesis $\mathsf{H}_{0,\mathsf{CS}}(\delta)$.

Remark 3.4. *Because our results are conditioning on the training sample, the test* $\widehat{\chi}_{3,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^\dagger/2}, \delta)$ *only relies on 'asymptopia' to ensure that its rejection probability can be closely approximated by its Gaussian limiting distribution. If Berry-Esseen bound or tail inequalities with constants estimable from the data were available for* $\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})$, *we could in principle eliminate our dependence on asymptopia. At the sample size used in our simulation (see Section 5), the normal qqplots of* $\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})$ *in Figure 2 suggest that normal approximation looks fine between* $(-2, 2)$ *under* $\mathsf{H}_{0,\mathsf{CS}}(\delta)$.

Remark 3.5. *Now we return to the issue raised in footnote 5 in the Introduction. Another natural choice for the* k-*projected null hypothesis is*

$$\mathsf{H}_{0,k,\mathsf{CS}}(\delta) : \mathsf{CSBias}_{\theta,k}(\widehat{\psi}_1)^2 \equiv \mathsf{E}_\theta\left[\Pi_\theta[\lambda^{1/2}(\widehat{b} - b)|\lambda^{1/2}\bar{\mathsf{z}}_k](X)^2\right]\mathsf{E}_\theta\left[\Pi_\theta[\lambda^{1/2}(\widehat{p} - p)|\lambda^{1/2}\bar{\mathsf{z}}_k](X)^2\right] \leqslant \delta^2\mathsf{var}_\theta(\widehat{\psi}_1)$$

because rejection of $\mathsf{H}_{0,k,\mathsf{CS}}(\delta)$ *implies rejection of* $\mathsf{H}_{0,\mathsf{CS}}(\delta)$. $\mathsf{CSBias}_{\theta,k}(\widehat{\psi}_1)^2$ *can be unbiasedly estimated by the following infeasible fourth-order U-statistic:*

$$\widehat{\mathbb{IF}}_{44,\mathsf{CS},k}(\Sigma_k^{-1}) := \frac{(n-4)!}{n!} \sum_{1\leqslant i_1 \neq i_2 \neq i_3 \neq i_4 \leqslant n} [\mathcal{E}_{\widehat{b},m_2}(\bar{\mathsf{z}}_k)(O)]_{i_1}^\top \Sigma_k^{-1} [\mathcal{E}_{\widehat{b},m_2}(\bar{\mathsf{z}}_k)(O)]_{i_2} [\mathcal{E}_{\widehat{p},m_1}(\bar{\mathsf{z}}_k)(O)]_{i_3}^\top \Sigma_k^{-1} [\mathcal{E}_{\widehat{p},m_1}(\bar{\mathsf{z}}_k)(O)]_{i_4}.$$

Using $\widehat{\mathbb{IF}}_{44,\mathsf{CS},k}(\Sigma_k^{-1})$ *as a starting point, one can indeed follow the previous development in Section 3 to construct a feasible test of* $\mathsf{H}_{0,k,\mathsf{CS}}(\delta)$. *We decide not to further pursue this alternative strategy because it is much more complex to analyze and we leave it to future investigation.*

4. What if nonstandard DML estimators are used in practice?

4.1. **Motivation for testing** $\mathsf{H}_{0,k}(\delta)$.

Several recent works (Newey and Robins, 2018; Bradic et al., 2019; Kline, Saggio, and Sølvsten, 2020; Kennedy, 2020; McGrath and Mukherjee, 2022) have exhibited nonstandard DML estimators, denoted also as $\widehat{\psi}_1$ in this section (to avoid introducing new notation at this point), that improve upon standard DML estimators, with $\mathsf{Bias}_\theta(\widehat{\psi}_1) = o(n^{-1/2})$, even though $\mathsf{CSBias}_\theta(\widehat{\psi}_1) \gg n^{-1/2}$, i.e. $\kappa_b + \kappa_p \leqslant 1/2$ [recall that $n^{-\kappa_b}$ and $n^{-\kappa_p}$ are rates of convergence of $\widehat{b}$ and $\widehat{p}$ to b and p in (weighted) $L_2(\mathsf{P}_\theta)$ norm].

However, to obtain these results, they all (1) use very special nuisance function estimators with $\widehat{b}$ and $\widehat{p}$ computed from separate non-overlapping subsamples of the training sample, and (2) assume very specific complexity-reducing assumptions such as Hölder smoothness (Newey and Robins, 2018; McGrath and Mukherjee, 2022) or (approximate) sparsity (Bradic et al., 2019). The problem for our methodology when applied to such estimators is that, if (1) and (2) above were true, our test $\widehat{\chi}_{3,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^\dagger/2}, \delta)$ might still reject because $\kappa_b + \kappa_p \leqslant 1/2$, even though $\mathsf{Bias}_\theta(\widehat{\psi}_1) = o(n^{-1/2})$ holds.

In this section, we study one possible approach to extend our methodology to apply to nonstandard DML estimators. We will consider an approach in which we simply wish to test the null hypothesis $\mathsf{H}_{0,k}(\delta)$. This can be viewed as testing if the bias of $\widehat{\psi}_1$ is large along the *direction* of the basis/dictionary $\bar{z}_k$ chosen by the analyst. If the analyst has a good grasp of the direction along which $\mathsf{Bias}_\theta(\widehat{\psi}_1)$ may be large, say when b, p belong to Hölder balls with certain smoothness, then she can test $\mathsf{H}_{0,k}(\delta)$ with $\bar{z}_k$ chosen to be wavelet or B-spline series. If otherwise, we then suggest the analyst tests $\mathsf{H}_{0,k}(\delta)$ with a variety of choices of $\bar{z}_k$, as a form of sensitivity analyses (Robins, 2003). If $\mathsf{H}_{0,k}(\delta)$ is rejected with some $\bar{z}_k$ (with a very small p-value), then $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1) \gtrsim n^{-1/2}$ under the operationalized pairing between $\mathsf{H}_{0,k}(\delta)$ and $\mathsf{NH}_{0,k} : \mathsf{Bias}_{\theta,k}(\widehat{\psi}_1) = o(n^{-1/2})$. Recall from Section 2.3 that $\mathsf{Bias}_\theta(\widehat{\psi}_1)$ can be decomposed as follows:

$$\mathsf{Bias}_\theta(\widehat{\psi}_1) \equiv \mathsf{Bias}_{\theta,k}(\widehat{\psi}_1) + \mathsf{TB}_{\theta,k}(\widehat{\psi}_1).$$

Furthermore, $\mathsf{TB}_{\theta,k}(\widehat{\psi}_1)$ cannot be uniformly consistently estimated. Even so, $\mathsf{Bias}_\theta(\widehat{\psi}_1)$ will not be $o(n^{-1/2})$, and therefore its associated Wald CI centered at $\widehat{\psi}_1$ will not be valid, unless $\mathsf{TB}_{\theta,k}(\widehat{\psi}_1)$ and $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$ happen to have leading order terms of the same magnitude but opposite sign. As it seems quite fortuitous for such a cancellation to occur, an analyst might agree to retract her claim that the Wald CI is valid. Hence in this section we focus on $\mathsf{H}_{0,k}(\delta)$ and investigate how to test this particular null hypothesis.

4.2. **Issues with $\widehat{\chi}_{3,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^\dagger/2}, \delta)$ as a test of $\mathsf{H}_{0,k}(\delta)$ and a rescue by HOIFs.**

In this section, we begin by showing that $\widehat{\chi}_{3,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^\dagger/2}, \delta)$ is not necessarily an $\alpha^\dagger$-level test of $\mathsf{H}_{0,k}(\delta)$ for nonstandard DML estimators (neither for standard DML estimators, which does not contradict the claim about the validity of $\widehat{\chi}_{3,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^\dagger/2}, \delta)$ in Theorem 3.2 because it is shown to be valid under $\mathsf{H}_{0,\mathsf{CS}}(\delta)$, a different null hypothesis from $\mathsf{H}_{0,k}(\delta)$). Recall from Theorem 3.1 that the bias $\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})$ for $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$ was upper bounded by:

$$\mathsf{EB}_{3,\theta,k} = \mathsf{E}_\theta[\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1}) - \mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)] \lesssim \mathbb{L}_{\theta,2,\widehat{b},k}\mathbb{L}_{\theta,2,\widehat{p},k}\frac{k\log k}{n}.$$

Under $\mathsf{H}_{0,k}(\theta)$, $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1) \lesssim n^{-1/2}$, but because $\mathbb{L}_{\theta,2,\widehat{b},k}\mathbb{L}_{\theta,2,\widehat{p},k} \equiv \mathsf{CSBias}_{\theta,k}(\widehat{\psi}_1) \geqslant \mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$ by CS inequality, we cannot ensure $\mathbb{L}_{\theta,2,\widehat{b},k}\mathbb{L}_{\theta,2,\widehat{p},k} \lesssim n^{-1/2}$. Hence one cannot guarantee $\mathsf{EB}_{3,\theta,k} = o\left(\frac{\sqrt{k}}{n}\right)$, as required in Proposition 3.1 to ensure the validity of $\widehat{\chi}_{3,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^\dagger/2}, \delta)$ as a test of $\mathsf{H}_{0,k}(\delta)$, without making further unverifiable assumptions on $b, p, \widehat{b}, \widehat{p}$ or $\bar{z}_k$. A natural solution would be to further reduce the bias

of $\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})$ due to estimating Σ_k^{-1} by $\widehat{\Sigma}_k^{-1}$, which motivates the following m-th order U-statistic estimator of $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$:

$$\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1}) := \sum_{j=2}^{m} \widehat{\mathbb{IF}}_{jj,k}(\widehat{\Sigma}_k^{-1}), \text{ where}$$

$$\widehat{\mathbb{IF}}_{jj,k}(\widehat{\Sigma}_k^{-1}) := (-1)^j \frac{(n-j)!}{n!} \left[\mathcal{E}_{\widehat{b},m_2}(\bar{z}_k)(O) \right]_{i_1}^{\top} \left\{ \prod_{s=3}^{j} \widehat{\Sigma}_k^{-1} \left(\left[S_{bp} \bar{z}_k(X) \bar{z}_k(X)^{\top} \right]_{i_s} - \widehat{\Sigma}_k \right) \right\} \widehat{\Sigma}_k^{-1} \left[\mathcal{E}_{\widehat{p},m_1}(\bar{z}_k)(O) \right]_{i_2}.$$

$\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})$ can further reduce the bias [and $\widehat{\mathbb{IF}}_{22\to mm,k}(\Sigma_k^{-1})$ is the m-th order influence function of $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$ (see HOIF-related theory in Appendix A.6)]. In particular, we have the following theorem on the statistical properties of $\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})$, similar to Theorem 3.1. Again, the bias and variance bounds under Condition S are proved in Liu et al. (2017). Under Condition W, see the proof of Lemma A.3 in Appendix A.7. Recall from Section 3 that Condition W is weaker than Condition S.

Theorem 4.1. *Under the conditions as in Theorem 3.1 and $k\log k \vee (km^2) = o(n)$:*

(1) *The bias and variance of $\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})$ as an estimator of $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$ are of the following order:*

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

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

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

If, however, Condition W is strengthened to Condition S,

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

(2) *Under $\mathsf{H}_{0,k}(\delta)$ and Condition S,*

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

When we estimate $\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]$ by its consistent bootstrap estimator $\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]$ (e.g. by (A.22) in Appendix A.8 with Σ_k replaced by $\widehat{\Sigma}_k$), we also have

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

Remark 4.1. *First, note that under $\mathsf{H}_{0,k}(\delta)$ instead of $\mathsf{H}_{0,\mathsf{CS}}(\delta)$ as in the previous section, we only have asymptotic normality of $\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})$ under Condition S. Under Condition W, if we were able to show $\mathbb{L}_{\theta,4,\widehat{b},k}^2 \mathbb{L}_{\theta,4,\widehat{p},k}^2 = O(1)$ when $\mathbb{L}_{\theta,2,\widehat{b}}, \mathbb{L}_{\theta,2,\widehat{p}}, \mathbb{L}_{\theta,\infty,\widehat{b}}, \mathbb{L}_{\theta,\infty,\widehat{p}}$ are all $O(1)$, then asymptotic normality would also hold. This remains an open question. For more detailed discussion, see Appendix A.7.*

Further note the following trade-off between the asymptotic statistical properties and computational cost:

(1) *As m grows, the bias of $\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})$ decays to 0 at a rate monotonically increasing with m.*

(2) *But larger m incurs higher computational cost to compute $\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})$.*

We conjecture that such a trade-off is unavoidable but a rigorous proof is beyond the scope of this paper.

Finally, the extra condition $km^2 = o(n)$ imposes restrictions on both k and m and is a result of the variance bound of $\widehat{\mathbb{IF}}_{mm,k}(\widehat{\Sigma}_k^{-1})$ shown in Lemma A.3. For example, if one chooses $k = n/(\log n)^2$, then m is at most $o(\log n)$.

4.3. **The proposed assumption-lean test of $\mathsf{H}_{0,k}(\delta)$ and early-stopping.**

Based on the above discussion, we propose the following nominal $\alpha^\dagger$-level test of $\mathsf{H}_{0,k}(\delta)$: for any fixed $m \geqslant 3$,

$$(4.1) \qquad \widehat{\chi}_{m,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^\dagger/2}, \delta) := \mathbb{1}\left\{ \frac{|\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})|}{\widehat{\text{s.e.}}[\widehat{\psi}_1]} - z_{\alpha^\dagger/2} \frac{\widehat{\text{s.e.}}[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]}{\widehat{\text{s.e.}}[\widehat{\psi}_1]} > \delta \right\}.$$

Theorem 4.1 immediately implies the following:

Theorem 4.2. *Under the conditions in Theorem 4.1 with Condition S, $k\log(k) \vee (km^2) = o(n)$ together with the additional restriction*

$$(4.2) \qquad \mathsf{Bias}_{\theta,k}(\widehat{\psi}_1) \neq o\left(\mathbb{L}_{\theta,2,\widehat{b},k} \mathbb{L}_{\theta,2,\widehat{p},k} \left(\frac{k\log(k)}{n} \right)^{\frac{m-1}{2}} \right)$$

for any given $\delta > 0$, suppose that $\frac{|\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)|}{\text{s.e.}_\theta[\widehat{\psi}_1]} = \gamma$ for some (sequence) $\gamma = \gamma(n)$ (where $\gamma(n)$ can diverge with n), then the rejection probability of $\widehat{\chi}_{m,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^\dagger/2}, \delta)$ converges to

$$(4.3)$$
$$2 - \Phi\left(z_{\alpha^\dagger/2} - \lim_{n\to\infty}(\gamma - \delta) \frac{\text{s.e.}_\theta[\widehat{\psi}_1]}{\text{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]} \right) - \Phi\left(z_{\alpha^\dagger/2} + \lim_{n\to\infty}(\gamma + \delta) \frac{\text{s.e.}_\theta[\widehat{\psi}_1]}{\text{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]} \right)$$

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

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

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

(2′) *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 $c = c(n) \to \infty$, $\widehat{\chi}_{m,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^\dagger/2}, \delta)$ has rejection probability converging to 1, as $n \to \infty$.*

Remark 4.2. *Note that the greater m is, the less restrictive (4.2) becomes. In fact we can let $m \equiv m(n) \to \infty$ as $n \to \infty$, by which (4.2) converges to a null requirement.*

However, as stated in Remark 4.1, although increasing m relaxes the restriction (4.2), it incurs higher computational cost. In practice, one might want m to be not too large. In view of such computational concern in practice, we propose the following "early-stopping" test that terminates at the smallest m at which it fails to reject $\mathsf{H}_{0,k}(\delta)$ and reports "failure to reject $\mathsf{H}_{0,k}(\delta)$".

Finally, notice that we only state Theorem 4.2 under the stronger Condition S. If weakened to Condition W, even under $\mathsf{H}_{0,k}(\delta)$, $k\mathbb{L}^2_{\theta,2,\widehat{b},k}\mathbb{L}^2_{\theta,2,\widehat{p},k}$ may still diverge, and the asymptotic normality of $\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})$ cannot be justified unless we were able to show $\mathbb{L}^2_{\theta,4,\widehat{b},k}\mathbb{L}^2_{\theta,4,\widehat{p},k}$ to be bounded, as discussed in Remark 4.1 and Appendix A.7. Thus we need to change the cutoff from $z_{\alpha^\dagger/2}$ in (4.1) to $(\alpha^\dagger)^{-1/2}$ and justify the level of the test using Chebyshev inequality instead.

Formally, for a fixed integer $M > 0$ that is determined by the analyst's computational budget, define

(4.4)
$$\widehat{\chi}^{es}_{M,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^\dagger/2}, \delta) := \mathbb{1}\left\{ \frac{|\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})|}{\widehat{\text{s.e.}}[\widehat{\psi}_1]} - z_{\alpha^\dagger/2}\frac{\widehat{\text{s.e.}}[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]}{\widehat{\text{s.e.}}[\widehat{\psi}_1]} > \delta : \forall\, m = 2, \ldots, M \right\}.$$

If $\widehat{\chi}^{es}_{M,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^\dagger/2}, \delta)$ 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 is again an asymptotically valid $\alpha^\dagger$-level test, under slightly stronger assumptions than those in Theorem 4.2:

Proposition 4.1. *Assume all the conditions of Theorem 4.2, but with (4.2) replaced by*

$$\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1) \neq o\left(\mathbb{L}_{\theta,2,\widehat{b},k}\mathbb{L}_{\theta,2,\widehat{p},k}\left(\frac{k\log(k)}{n} \right)^{\frac{M-1}{2}} \right).$$

Under $\mathsf{H}_{0,k}(\delta)$:

$$\mathsf{P}_\theta\left(\widehat{\chi}^{es}_{M,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^\dagger/2}, \delta) = 1 \right) \leqslant \alpha^\dagger$$

that is, $\widehat{\chi}^{es}_{M,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^\dagger/2}, \delta)$ is an asymptotically valid level-$\alpha^\dagger$ test of $\mathsf{H}_{0,k}(\delta)$.

The power of the above "early-stopping" procedure is more challenging to characterize, hence out of the scope of this paper.

5. Monte Carlo experiments

In the Monte Carlo (MC) experiments, we focus on the parameter (up to a minus sign) that was extensively discussed in the Introduction, $\psi(\theta) \equiv \mathsf{E}_\theta[Y(a = 1)] \equiv \mathsf{E}_\theta[b(X)]$ under ignorability. Recall that $p(X) = 1/\mathsf{E}_\theta[A|X]$ 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, $\Sigma_k = \mathsf{E}_\theta[A\bar{z}_k(X)\bar{z}_k(X)^\top]$ and $\widehat{\Sigma}_k = n^{-1}\sum_{i\in\mathsf{tr}} A_i\bar{z}_k(X_i)\bar{z}_k(X_i)^\top$. We choose $\psi(\theta) \equiv 0$.

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_j of each X_j is supported on $[0,1]$ with $f_j \in$ 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 at resolutions $\ell \in (6,7,8)$ to form the basis functions

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

with the corresponding $k' \in \{2^6 = 64, 2^7 = 128, 2^8 = 256\}$ and $k \in \{64 \cdot 4 = 256, 128 \cdot 4 = 512, 256 \cdot 4 = 1024\}$. To compute the oracle statistics and tests, we evaluate Σ_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). In particular, the smoothing parameters were selected by generalized cross validation, the default setup in gam function from R package mgcv. We choose db6 father wavelets to construct HOIF estimators when analyzing the simulated data.

5.1. **Finite sample performance of tests for** $\mathsf{H}_{0,k}(\delta)$**.** In this section, we consider testing the null hypothesis $\mathsf{H}_{0,k}(\delta)$. 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 $\widehat{\chi}_{2,k}(\Sigma_k^{-1})$ where $\widehat{\psi}_{2,k}(\Sigma_k^{-1}) = \widehat{\psi}_1 - \widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$, together with the statistics

and tests relying on $\widehat{\Sigma}_k^{-1}$: $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$, $\widehat{\mathbb{IF}}_{33,k}(\widehat{\Sigma}_k^{-1})$, $\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})$, $\widehat{\chi}_{3,k}(\widehat{\Sigma}_k^{-1})$, $\widehat{\psi}_{2,k}(\widehat{\Sigma}_k^{-1})$, $\widehat{\psi}_{3,k}(\widehat{\Sigma}_k^{-1})$, and $\widehat{\chi}_{2,k}(\widehat{\Sigma}_k^{-1})$, where $\widehat{\psi}_{2,k}(\widehat{\Sigma}_k^{-1}) = \widehat{\psi}_1 - \widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$ and $\widehat{\psi}_{2,k}(\widehat{\Sigma}_k^{-1}) = \widehat{\psi}_1 - \widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})$

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 $\frac{\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})]}$ 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 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 $\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})$, as described in Appendix A.8. 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 $\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})$ 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 $\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})$ 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 $\mathsf{H}_{0,k}(\delta)$. 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 -34.26×10^{-3} and 8.77×10^{-3}. Since the ratio between the bias and standard error is around 4, we expect the associated 90% Wald CI $\widehat{\psi}_1 \pm z_{0.05}\widehat{\mathsf{s.e.}}(\widehat{\psi}_1)$ 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 CI 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 -8.88×10^{-3} and 8.10×10^{-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., 2009). Here the ratio between the bias and standard error is around 1 and we expect the associated 90% Wald CI $\widehat{\psi}_1 \pm z_{0.05}\widehat{\mathsf{s.e.}}(\widehat{\psi}_1)$ is slightly undercovered. 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 CI of $\widehat{\psi}_1$ is 83% (note that $\widehat{\psi}_{2,k} = \widehat{\psi}_1$ when $k = 0$).

- 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 $\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})$ 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 -18.76×10^{-3} to -25.34×10^{-3}), but there is no obvious difference in the MC averages of $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ between $k = 512$ and $k = 1024$ (-25.34×10^{-3} and -25.43×10^{-3}, about 75% of the total bias). The MC average of the estimated standard error increases with k, from 2.45×10^{-3} at $k = 256$ to 3.43×10^{-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 linearly 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 $\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})$ 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 -4.54×10^{-3} to -4.94×10^{-3}). The MC average of the estimated standard error increases with k, from 1.51×10^{-3} at $k = 256$ to 2.43×10^{-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 $\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})$ 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 CIs 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 Σ_k is unknown.

  In the third columns of Table 4, we display the MC coverage probabilities of the 90% two-sided CIs 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 Σ_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).

k	256	512	1024
$\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$	2.437	3.011	3.247
	2.268 (0.218)	2.716 (0.268)	2.841 (0.322)
$\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$	2.456	3.075	3.546
	2.271 (0.218)	2.778 (0.278)	3.032 (0.378)
$\widehat{\mathbb{IF}}_{33,k}(\widehat{\Sigma}_k^{-1})$	0.495	0.814	1.518
	0.603 (0.0598)	1.126 (0.126)	1.998 (0.285)
$\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})$	2.282	2.745	2.743
	2.251 (0.227)	2.738 (0.290)	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 $\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})$.

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 $\mathsf{H}_0(\delta)$ that the ratio between the bias of $\widehat{\psi}_1$ and its standard error is lower than δ. All the rejection rates in Table 2 are 100% when $\delta = 3/4$, regardless of whether we are using the oracle test or not. When $\delta = 2$, we obtained nontrivial rejection rates (data not shown). 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 $1/2$.

Remark 5.1. *Here we choose $k \lesssim n/(\log n)^2$, which is on the order of 1000 when $n = 50,000$. Larger k could easily lead to numerical instability due to inverting a large-dimensional sample Gram matrix (also see Liu, Mukherjee, and Robins (2020a)). In a technical report by one of the authors (Liu and Li, 2023), a new class of empirical HOIF estimators is proposed that overcomes the above issues by replacing $\widehat{\Sigma}_k$ computed from the training sample by that computed from the estimation sample. The theoretical results are somewhat more complicated to state than the estimators used in this paper. We will thus only refer interested readers to Liu and Li (2023). We also plan to report the finite-sample and real-world data performance of these new estimators and the corresponding test statistics in a separate manuscript.*

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	$\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1}) \times 10^{-3}$	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 $\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})$ (lower panel), together with the coverage probabilities of two-sided 90% Wald CIs (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 $\widehat{\chi}_{2,k}^{(1)}(\Sigma_k^{-1}; z_{0.10/2}, \delta = 3/4)$ (upper panel) and $\widehat{\chi}_{3,k}^{(1)}(\widehat{\Sigma}_k^{-1}; z_{0.10/2}, \delta = 3/4)$ (lower panel). In the upper panel, for $k = 0$, $\widehat{\psi}_{2,k=0} \equiv \widehat{\psi}_1$.

k	256	512	1024
$\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$	1.166 1.209 (0.156)	1.367 1.408 (0.191)	1.673 1.584 (0.267)
$\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})$	1.187 1.219 (0.159)	1.433 1.439 (0.198)	1.828 1.697 (0.285)
$\widehat{\mathbb{IF}}_{33,k}(\widehat{\Sigma}_k^{-1})$	0.251 0.331 (0.0451)	0.417 0.629 (0.0997)	0.854 1.251 (0.247)
$\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})$	1.109 1.186 (0.156)	1.256 1.346 (0.191)	1.488 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 $\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})$.

6. Concluding remarks

In this paper, we developed a valid assumption-lean third-order U-statistic based test that can empirically falsify the justification for a Wald CI centered at a standard DML estimator $\widehat{\psi}_1$ covering the underlying DR functional $\psi(\theta)$ at nominal coverage (see Section 3). When nonstandard DML estimators are used, we also develop a test that is based on higher-order U-statistics, which are in fact HOIFs of Bias$_{\theta,k}(\widehat{\psi}_1)$. We mention a few interesting future directions to end our manuscript. First, it is important to extend the proposed approach to allow endogeneity (Angrist, Imbens, and Rubin, 1996; Newey and

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	$\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1}) \times 10^{-3}$	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 $\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})$ (lower panel), together with the coverage probabilities of two-sided 90% Wald CIs (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 $\widehat{\chi}_{2,k}^{(1)}(\Sigma_k^{-1}; z_{0.10/2}, \delta = 3/4)$ (upper panel) and $\widehat{\chi}_{3,k}^{(1)}(\widehat{\Sigma}_k^{-1}; z_{0.10/2}, \delta = 3/4)$ (lower panel). In the upper panel, for $k = 0$, $\widehat{\psi}_{2,k=0} \equiv \widehat{\psi}_1$.

Powell, 2003; Newey, 2013; Chen and Christensen, 2013; Breunig and Chen, 2021; Chen and Christensen, 2018; Chen and Qiu, 2016), when some auxiliary variables are available for point-identifying the causal effects. Machine learning, including deep learning, has also been applied to these problems in recent years (Chen, Chen, and Tamer, 2023; Kompa et al., 2022). In Appendix C, we document the HOIFs for the running example $\psi(\theta) = -\mathsf{E}_\theta[Y(a = 1)]$ in Section 1 under the so-call "proximal causal learning" framework (Tchetgen Tchetgen et al., 2020)[13]. This framework has been shown to be closely related to other approaches dealing with endogeneity, e.g. synthetic controls (Abadie, Diamond, and Hainmueller, 2010; Shi et al., 2021) or instrumental variables (Breunig and Chen, 2021). Based on the form of these HOIFs, it is straightforward to generalize our method to scenarios under endogeneity. Second, extending our framework to heterogeneous treatment effect (Chernozhukov et al., 2017; Kennedy, Balakrishnan, and Wasserman, 2022) could also be of practical interest, given its significance in personalized decision making. Third, another interesting direction is to explore if it is possible to use other bias correction strategies to construct the assumption-lean falsification test, such as the (iterative) bootstrap approach investigated in Cattaneo and Jansson (2018), Cattaneo, Jansson, and Ma (2019), and Koltchinskii (2022). Finally, we are also considering to directly learn a data-driven representation $\widetilde{z}_k$ via the penultimate layer

---

[13]Based on personal communication, Eric Tchetgen Tchetgen and colleagues have also independently derived HOIFs of this parameter under the proximal causal learning framework in an unpublished manuscript.

(Ansuini et al., 2019) or distillation (Ha et al., 2021) of a deep neural network trained to predict the residuals of the nuisance function estimators, hoping to increase the chance of rejection when the bias indeed exceeds $n^{-1/2}$ in real-world settings.

REFERENCES

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller (2010). "Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program". In: *Journal of the American Statistical Association* 105.490, pp. 493–505.

Angrist, Joshua D, Guido W Imbens, and Donald B Rubin (1996). "Identification of causal effects using instrumental variables". In: *Journal of the American Statistical Association* 91.434, pp. 444–455.

Ansuini, Alessio et al. (2019). "Intrinsic dimension of data representations in deep neural networks". In: *Advances in Neural Information Processing Systems*. Vol. 32, pp. 6111–6122.

Bang, Heejung and James M Robins (2005). "Doubly robust estimation in missing data and causal inference models". In: *Biometrics* 61.4, pp. 962–973.

Belloni, Alexandre et al. (2015). "Some new asymptotic theory for least squares series: Pointwise and uniform results". In: *Journal of Econometrics* 186.2, pp. 345–366.

Bhattacharya, Rabi N and Jayanta K Ghosh (1992). "A class of U-statistics and asymptotic normality of the number of k-clusters". In: *Journal of Multivariate Analysis* 43.2, pp. 300–330.

Bradic, Jelena, Stefan Wager, and Yinchu Zhu (2019). "Sparsity double robust inference of average treatment effects". In: *arXiv preprint arXiv:1905.00744*.

Bradic, Jelena et al. (2019). "Minimax Semiparametric Learning With Approximate Sparsity". In: *arXiv preprint arXiv:1912.12213*.

Breunig, Christoph and Xiaohong Chen (2021). "Adaptive Estimation of Quadratic Functionals in Nonparametric Instrumental Variable Models". In: *arXiv preprint arXiv:2101.12282*.

Cattaneo, Matias D and Michael Jansson (2018). "Kernel-Based Semiparametric Estimators: Small Bandwidth Asymptotics and Bootstrap Consistency". In: *Econometrica* 86.3, pp. 955–995.

Cattaneo, Matias D, Michael Jansson, and Xinwei Ma (2019). "Two-step estimation and inference with possibly many included covariates". In: *The Review of Economic Studies* 86.3, pp. 1095–1122.

Cattaneo, Matias D, Michael Jansson, and Whitney K Newey (2018). "Inference in linear regression models with many covariates and heteroscedasticity". In: *Journal of the American Statistical Association* 113.523, pp. 1350–1361.

Chen, Jiafeng, Xiaohong Chen, and Elie Tamer (2023). "Efficient estimation in NPIV models: A comparison of various neural networks-based estimators". In: *Journal of Econometrics*.

Chen, Xiaohong and Timothy Christensen (2013). "Optimal uniform convergence rates for sieve nonparametric instrumental variables regression". In: *arXiv preprint arXiv:1311.0412*.

Chen, Xiaohong and Timothy M Christensen (2018). "Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric IV regression". In: *Quantitative Economics* 9.1, pp. 39–84.

Chen, Xiaohong and Yin Jia Jeff Qiu (2016). "Methods for nonparametric and semiparametric regressions with endogeneity: A gentle guide". In: *Annual Review of Economics* 8, pp. 259–290.

Chen, Xiaohong et al. (2020). "Casual inference of General Treatment Effects using Neural Networks with A Diverging Number of Confounders". In: *arXiv preprint arXiv:2009.07055*.

Chernozhukov, Victor, Whitney K Newey, and Rahul Singh (2022a). "Automatic debiased machine learning of causal and structural effects". In: *Econometrica* 90.3, pp. 967–1027.

– (2022b). "Debiased machine learning of global and local parameters using regularized Riesz representers". In: *The Econometrics Journal* 25.3, pp. 576–601.

Chernozhukov, Victor et al. (2017). "Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments". In: *arXiv preprint arXiv:1712.04802*.

Chernozhukov, Victor et al. (2018). "Double/debiased machine learning for treatment and structural parameters". In: *The Econometrics Journal* 21.1, pp. C1–C68.

Chernozhukov, Victor et al. (2022). "Locally robust semiparametric estimation". In: *Econometrica* 90.4, pp. 1501–1535.

Farrell, Max H (2015). "Robust inference on average treatment effects with possibly more covariates than observations". In: *Journal of Econometrics* 189.1, pp. 1–23.

Farrell, Max H, Tengyuan Liang, and Sanjog Misra (2021). "Deep neural networks for estimation and inference". In: *Econometrica* 89.1, pp. 181–213.

Ha, Wooseok et al. (2021). "Adaptive wavelet distillation from neural networks through interpretations". In: *Advances in Neural Information Processing Systems* 34, pp. 20669–20682.

Ichimura, Hidehiko and Whitney K Newey (2022). "The influence function of semiparametric estimators". In: *Quantitative Economics* 13.1, pp. 29–61.

Jiang, Kuanhao et al. (2022). "A New Central Limit Theorem for the Augmented IPW Estimator: Variance Inflation, Cross-Fit Covariance and Beyond". In: *arXiv preprint arXiv:2205.10198*.

Jochmans, Koen (2022). "Heteroscedasticity-robust inference in linear regression models with many covariates". In: *Journal of the American Statistical Association* 117.538, pp. 887–896.

Kennedy, Edward H (2020). "Towards optimal doubly robust estimation of heterogeneous causal effects". In: *arXiv preprint arXiv:2004.14497*.

Kennedy, Edward H, Sivaraman Balakrishnan, and Larry Wasserman (2020). "Discussion of "On Nearly Assumption-Free Tests of Nominal Confidence Interval Coverage for Causal Parameters Estimated by Machine Learning"". In: *Statistical Science* 35.3, pp. 540–544.

– (2022). "Minimax rates for heterogeneous causal effect estimation". In: *arXiv preprint arXiv:2203.00837*.

Kline, Patrick, Raffaele Saggio, and Mikkel Sølvsten (2020). "Leave-out estimation of variance components". In: *Econometrica* 88.5, pp. 1859–1898.

Koltchinskii, Vladimir (2022). "Estimation of smooth functionals of covariance operators: Jackknife bias reduction and bounds in terms of effective rank". In: *arXiv preprint arXiv:2205.10280*.

Kompa, Benjamin et al. (2022). "Deep learning methods for proximal inference via maximum moment restriction". In: *Advances in Neural Information Processing Systems* 36.

Liu, Lin and Chang Li (2023). "New $\sqrt{n}$-consistent, numerically stable empirical higher-order influence function estimators". In: *arXiv preprint arXiv:2302.08097*.

Liu, Lin, Rajarshi Mukherjee, and James M Robins (2020a). "On nearly assumption-free tests of nominal confidence interval coverage for causal parameters estimated by machine learning". In: *Statistical Science* 35.3, pp. 518–539.

– (2020b). "Rejoinder: On nearly assumption-free tests of nominal confidence interval coverage for causal parameters estimated by machine learning". In: *Statistical Science* 35.3, pp. 545–554.

Liu, Lin et al. (2017). "Semiparametric efficient empirical higher order influence function estimators". In: *arXiv preprint arXiv:1705.07577*.

McGrath, Sean and Rajarshi Mukherjee (2022). "On Undersmoothing and Sample Splitting for Estimating a Doubly Robust Functional". In: *arXiv preprint arXiv:2212.14857*.

Newey, Whitney K (1990). "Semiparametric efficiency bounds". In: *Journal of Applied Econometrics* 5.2, pp. 99–135.

– (2013). "Nonparametric instrumental variables estimation". In: *American Economic Review* 103.3, pp. 550–556.

Newey, Whitney K, Fushing Hsieh, and James Robins (1998). *Undersmoothing and bias corrected functional estimation*. Tech. rep.

Newey, Whitney K, Fushing Hsieh, and James M Robins (2004). "Twicing kernels and a small bias property of semiparametric estimators". In: *Econometrica* 72.3, pp. 947–962.

Newey, Whitney K and Daniel McFadden (1994). "Large sample estimation and hypothesis testing". In: *Handbook of Econometrics* 4, pp. 2111–2245.

Newey, Whitney K and James L Powell (2003). "Instrumental variable estimation of nonparametric models". In: *Econometrica* 71.5, pp. 1565–1578.

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

Ritov, Ya'acov et al. (2014). "The Bayesian analysis of complex, high-dimensional models: Can it be CODA?" In: *Statistical Science* 29.4, pp. 619–639.

Robins, James et al. (2008). "Higher order influence functions and minimax estimation of nonlinear functionals". In: *Probability and Statistics: Essays in Honor of David A. Freedman*. Institute of Mathematical Statistics, pp. 335–421.

Robins, James et al. (2009). "Semiparametric minimax rates". In: *Electronic Journal of Statistics* 3, pp. 1305–1321.

Robins, James M (2003). "General methodological considerations". In: *Journal of Econometrics* 112.1, pp. 89–106.

Robins, James M and Ya'acov Ritov (1997). "Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models". In: *Statistics in Medicine* 16.3, pp. 285–319.

Robins, James M and Andrea Rotnitzky (2001). "Comments on "Inference for semiparametric models: some questions and an answer"". In: *Statistica Sinica* 11.4, pp. 920–936.

Robins, James M et al. (2017). "Minimax estimation of a functional on a structured high-dimensional model". In: *The Annals of Statistics* 45.5, pp. 1951–1987.

Rotnitzky, Andrea, Ezequiel Smucler, and James M Robins (2021). "Characterization of parameters with a mixed bias property". In: *Biometrika* 108.1, pp. 231–238.

Scharfstein, Daniel O, Andrea Rotnitzky, and James M Robins (1999). "Rejoinder". In: *Journal of the American Statistical Association* 94.448, pp. 1135–1146.

Schmidt-Hieber, Johannes (2020). "Nonparametric regression using deep neural networks with ReLU activation function". In: *Annals of Statistics* 48.4, pp. 1875–1897.

Shi, Xu et al. (2021). "Theory for identification and inference with synthetic controls: a proximal causal inference framework". In: *arXiv preprint arXiv:2108.13935*.

Smucler, Ezequiel, Andrea Rotnitzky, and James M Robins (2019). "A unifying approach for doubly-robust ℓ_1 regularized estimation of causal contrasts". In: *arXiv preprint arXiv:1904.03737*.

Tchetgen Tchetgen, Eric et al. (2008). "Minimax estimation of the integral of a power of a density". In: *Statistics & Probability Letters* 78.18, pp. 3307–3311.

Tchetgen Tchetgen, Eric J et al. (2020). "An Introduction to Proximal Causal Learning". In: *arXiv preprint arXiv:2009.10982*.

van der Vaart, Aad (2014). "Higher Order Tangent Spaces and Influence Functions". In: *Statistical Science* 29.4, pp. 679–686.

Xu, Siqi, Lin Liu, and Zhonghua Liu (2022). "DeepMed: Semiparametric causal mediation analysis with debiased deep learning". In: *Advances in Neural Information Processing Systems* 36.

Yadlowsky, Steve (2022). "Causal Inference in High Dimensions–Without Sparsity". In: *arXiv preprint arXiv:2203.12538*.

APPENDIX A. DERIVATIONS AND PROOFS

A.1. **More detailed review of doubly robust functionals (DR functionals).** All the theoretical results in this paper rely on the following regularity condition (Condition 1 in Rotnitzky, Smucler, and Robins (2021)):

Condition A.1. *There exists a dense set $\mathcal{H}_b$ of $L_2(\mathsf{P}_F)$ such that $\mathcal{F}_b \cap \mathcal{B} \neq \emptyset$, and for each $\theta \in \Theta$ and $h \in \mathcal{H}_b$, there exists $\varepsilon(\theta, h) > 0$ such that $b + th \in \mathcal{B}$ if $|t| < \varepsilon(\theta, h)$. The same holds replacing b with p and $\mathcal{B}$ with $\mathcal{P}$. Furthermore $\mathsf{E}_\theta[|S_{bp}p(X)h(X)|] < \infty$ for any $h \in \mathcal{H}_b$ and $\mathsf{E}_\theta[|S_{bp}b(X)h(X)|] < \infty$ for any $h \in \mathcal{H}_p$. Moreover for all $\theta \in \Theta$, $\mathsf{E}_\theta[|S_{bp}b'(X)p'(X)|] < \infty$ for any $b' \in \mathcal{B}$ and $p' \in \mathcal{P}$.*

A.2. **More details on training strategies for nuisance functions.** In this section, we discuss training strategies for nuisance functions in further details. First, we prove Lemma 2.1.

Proof of Lemma 2.1. The following algebra makes this statement explicit:

$$\frac{1}{2}\mathsf{E}_\theta[\lambda(X)(b(X) - h(X))^2] = \frac{1}{2}\mathsf{E}_\theta[S_{bp}(b(X) - h(X))^2]$$

$$= \mathsf{E}_\theta[S_{bp}\frac{h(X)^2}{2} - S_{bp}b(X)h(X)] + \frac{1}{2}\mathsf{E}_\theta[S_{bp}b(X)^2]$$

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

where in the third line we use equation (2.4) 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,

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

∎

For cross-entropy loss commonly employed in machine learning (in particular deep learning) when the response is binary, it is also approximately solving an $L_2(\mathsf{P}_\theta)$ loss minimization: here we denote $\pi = p^{-1}$ and $\widetilde{\pi}$ is the one being fitted in the minimization problem, by using Taylor expansion

$$\mathsf{E}_\theta[A\log\widetilde{\pi}(X) + (1 - A)\log(1 - \pi(X))]$$

$$= \mathsf{E}_\theta[\pi(X)\log\widetilde{\pi}(X) + (1 - \pi(X))\log(1 - \widetilde{\pi}(X))]$$

$$= \mathsf{E}_\theta\left[\pi(X)\left\{\log\pi(X) + \frac{\widetilde{\pi}(X) - \pi(X)}{\pi(X)} - \frac{(\widetilde{\pi}(X) - \pi(X))^2}{\pi(X)^2}\right\} + (1 - \pi(X))\left\{\log(1 - \pi(X)) - \frac{\widetilde{\pi}(X) - \pi(X)}{1 - \pi(X)} + \frac{(\widetilde{\pi}(X) - \pi(X))^2}{(1 - \pi(X))^2}\right\}\right]$$

$$+ \text{remainders}$$

$$= \text{constant} + \mathsf{E}_\theta\left[(\widetilde{\pi}(X) - \pi(X))^2\left(\frac{1}{\pi(X)} + \frac{1}{1 - \pi(X)}\right)\right].$$

We also want to point out the connection to recent works on using adversarial loss (Chernozhukov et al., 2020; Ghassami et al., 2022) to train the nuisance estimators. Adversarial loss is essentially trying to leverage the moment conditions in equation (2.4). As shown in Theorem 1 of Ghassami et al. (2022), the solution to the population adversarial loss with L_2-regularization is equivalent to the solution to the loss functions in equation (2.5) and

they establish the convergence rates exactly in the λ-weighted L_2-loss. Therefore, at least in principle, our testing framework is also applicable when nuisance functions are trained using adversarial loss according to Ghassami et al. (2022) and Chernozhukov et al. (2020) with an additional L_2-regularization term.

A.3. **The infeasible procedure when** $\Sigma_k^{-1} := \{\mathsf{E}_\theta[\lambda(X)\bar{z}_k(X)\bar{z}_k(X)^\top]\}^{-1}$ **is known.** It is pedagogically useful to first consider an infeasible procedure that would only be implementable if Σ_k^{-1} were known. We will focus on such situation in this section.

A.3.1. *Truncated parameters and their unbiased estimators.* Though there does not exist uniformly (in θ) 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 δ of its standard error 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 A.1 (The truncated parameter $\widetilde{\psi}_k(\theta)$, truncation bias $\mathsf{TB}_{\theta,k}$, 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 $\bar{z}_k(x)$ in $L_2(\mathsf{P}_F)$:*

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

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

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

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

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

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

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

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

where

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

Now define

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

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

(A.2)
$$\mathsf{TB}_{\theta,k} := \widetilde{\psi}_k(\theta) - \psi(\theta) = \mathsf{E}_\theta \left[S_{bp} \left(\widetilde{b}_{k,\theta}(X) - b(X) \right) \left(\widetilde{p}_{k,\theta}(X) - p(X) \right) \right].$$

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

(A.3)
$$\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1) := \mathsf{E}_\theta \left[\widehat{\psi}_1 - \widetilde{\psi}_k(\theta) \right].$$

By definition, $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1) = \mathsf{Bias}_\theta(\widehat{\psi}_1) - \mathsf{TB}_{\theta,k}$. Combining equation (A.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, $\mathsf{TB}_{\theta,k}$ is not consistently estimable without imposing smoothness/sparsity assumptions on the nuisance function spaces $\mathcal{B}$ and $\mathcal{P}$. But $\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$ can be unbiasedly estimated by the following oracle (i.e. Σ_k^{-1} known) second-order U-statistic:

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

where

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

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

(A.4)
$$\mathcal{E}_{b',m_2}(h)(O) := S_{bp} b'(X) h(X) + m_2(O,h)$$

and

(A.5)
$$\mathcal{E}_{p',m_1}(h)(O) := S_{bp} p'(X) h(X) + m_1(O,h).$$

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

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

(1) Example 1(1):
$$\begin{cases} \mathcal{E}_{\widehat{b},m_2}(\bar{z}_k)(O) = A\widehat{b}(X)\bar{z}_k(X) - AY\bar{z}_k(X), \\ \mathcal{E}_{\widehat{p},m_1}(\bar{z}_k)(O) = A\widehat{p}(X)\bar{z}_k(X) - \bar{z}_k(X). \end{cases}$$

(2) Example 1(2):
$$\begin{cases} \mathcal{E}_{\widehat{b},m_2}(\bar{z}_k)(O) = \widehat{b}(X)\bar{z}_k(X) - Y\bar{z}_k(X), \\ \mathcal{E}_{\widehat{p},m_1}(\bar{z}_k)(O) = \widehat{p}(X)\bar{z}_k(X) - A\bar{z}_k(X). \end{cases}$$

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

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

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

Lemma A.1. *Define* $\beta_{\widehat{b},k} := \mathsf{E}_\theta \left[\mathcal{E}_{\widehat{b},m_2}(\bar{\mathsf{z}}_k)(O) \right]^\top \Sigma_k^{-1}$ *and* $\beta_{\widehat{p},k} := \mathsf{E}_\theta \left[\mathcal{E}_{\widehat{p},m_1}(\bar{\mathsf{z}}_k)(O) \right]^\top \Sigma_k^{-1}$. *Under the conditions of Proposition [2.1](#), we have*

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

(A.6)
$$= \mathsf{E}_\theta \left[\Pi_\theta \left[\Delta_{1,\widehat{b}} | \bar{v}_k \right] (X) \Pi_\theta \left[\Delta_{2,\widehat{p}} | \bar{v}_k \right] (X) \right] = \beta_{\widehat{b},k}^\top \Sigma_k \beta_{\widehat{p},k}$$

and

(A.7) $\quad \mathsf{TB}_{\theta,k} := \widetilde{\psi}_k(\theta) - \psi(\theta) = \mathsf{Bias}_\theta(\widehat{\psi}_1) - \mathsf{Bias}_{\theta,k}(\widehat{\psi}_1) = \mathsf{E}_\theta \left[\Pi_\theta^\perp \left[\Delta_{1,\widehat{b}} | \bar{v}_k \right] (X) \Pi_\theta^\perp \left[\Delta_{2,\widehat{p}} | \bar{v}_k \right] (X) \right].$

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

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

(A.8)
$$\mathsf{Bias}_\theta(\widehat{\psi}_{2,k}(\Sigma_k^{-1})) \equiv \mathsf{E}_\theta \left[\widehat{\psi}_{2,k}(\Sigma_k^{-1}) - \psi(\theta) \right] = \mathsf{TB}_{\theta,k},$$

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

Remark A.1. *The proof of the above lemma is deferred to Appendix [A.4.1](#). 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:*

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

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

A.4. **Projection, truncation bias, and the motivation of higher-order influence functions.**

A.4.1. *Proof of Lemma [A.1](#).*

Proof. First, we show the first part of equation ([A.6](#)).

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

$$= \mathsf{E}_\theta \left[\begin{array}{l} S_{bp} \left\{ \widehat{b}(X) - b(X) - \mathsf{E}_\theta \left[(S_{bp}\widehat{b}(X)\bar{\mathsf{z}}_k(X) + m_2(O, \bar{\mathsf{z}}_k)) \right]^\top \Sigma_k^{-1} \bar{\mathsf{z}}_k(X) \right\} \\ \times \left\{ \widehat{p}(X) - p(X) - \mathsf{E}_\theta \left[(S_{bp}\widehat{p}(X)\bar{\mathsf{z}}_k(X) + m_1(O, \bar{\mathsf{z}}_k)) \right]^\top \Sigma_k^{-1} \bar{\mathsf{z}}_k(X) \right\} \end{array} \right]$$

$$= \mathsf{E}_\theta \left[S_{bp}(\widehat{b}(X) - b(X))(\widehat{p}(X) - p(X)) \right]$$

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

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

$$+ \mathsf{E}_\theta \left[S_{bp}(\widehat{p}(X) - p(X))\bar{\mathsf{z}}_k(X) \right]^\top \Sigma_k^{-1} \mathsf{E}_\theta \left[S_{bp}(\widehat{b}(X) - b(X))\bar{\mathsf{z}}_k(X) \right]$$

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

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

where in the third equality we apply Proposition 2.1(3). 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 Proposition 2.1(3),

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

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

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 (A.6):

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

$$= \mathsf{E}_\theta \left[\lambda(X)(\widehat{b}(X) - b(X))\bar{\mathsf{z}}_k(X) \right]^\top \Sigma_k^{-1} \mathsf{E}_\theta \left[\lambda(X)(\widehat{p}(X) - p(X))\bar{\mathsf{z}}_k(X) \right]$$

$$= \mathsf{E}_\theta \left[\lambda(X)^{1/2}(\widehat{b}(X) - b(X))\bar{v}_k(X) \right]^\top \Sigma_k^{-1} \mathsf{E}_\theta \left[\lambda(X)^{1/2}(\widehat{p}(X) - p(X))\bar{v}_k(X) \right]$$

$$= \mathsf{E}_\theta \left[\Delta_{1,\widehat{b}}\bar{v}_k(X) \right]^\top \Sigma_k^{-1} \mathsf{E}_\theta \left[\Delta_{2,\widehat{p}}\bar{v}_k(X) \right]$$

$$= \mathsf{E}_\theta \left[\Pi_\theta \left[\Delta_{1,\widehat{b}} | \bar{v}_k \right] (X) \Pi_\theta \left[\Delta_{2,\widehat{p}} | \bar{v}_k \right] (X) \right].$$

where in the last line we use the definition of the projection operator $\Pi_\theta \left[\cdot | \bar{v}_k \right]$ (see equation (2.2)).

Then equation (A.7) immediately follows:

$$\mathsf{TB}_{\theta,k} = \mathsf{Bias}_\theta(\widehat{\psi}_1) - \mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)$$

$$= \mathsf{E}_\theta \left[\Delta_{1,\widehat{b}}\Delta_{2,\widehat{p}} \right] - \mathsf{E}_\theta \left[\Pi_\theta \left[\Delta_{1,\widehat{b}} | \bar{v}_k \right] (X)^\top \Pi_\theta \left[\Delta_{2,\widehat{p}} | \bar{v}_k \right] (X) \right]$$

$$= \mathsf{E}_\theta \left[\Pi_\theta^\perp \left[\Delta_{1,\widehat{b}} | \bar{v}_k \right] (X) \right]^\top \mathsf{E}_\theta \left[\Pi_\theta^\perp \left[\Delta_{2,\widehat{p}} | \bar{v}_k \right] (X) \right].$$

Finally, we prove equation (A.8):

$$\mathsf{Bias}_\theta(\widehat{\psi}_{2,k}(\Sigma_k^{-1})) = \mathsf{E}_\theta \left[\widehat{\psi}_{2,k}(\Sigma_k^{-1}) - \psi(\theta) \right]$$

$$= \mathsf{E}_\theta \left[\widehat{\psi}_1 - \psi(\theta) - \widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1}) \right]$$

$$= \mathsf{Bias}_\theta(\widehat{\psi}_1) - \mathsf{Bias}_{\theta,k}(\widehat{\psi}_1) = \mathsf{TB}_{\theta,k}$$

where the third line follows from the definition of $\mathsf{Bias}_\theta(\widehat{\psi}_1)$ and equation (A.6). ∎

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 proposition.

Proposition A.1. *Under the conditions of Proposition 2.1 and Condition W, 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.8.*

(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.7. 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.8 proves that the bootstrap standard error estimators $\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]$ satisfy $\frac{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]} = 1 + o_{\mathsf{P}_\theta}(1)$. ∎

Remark A.2. *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.*

A.5. **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 Propositions 2.2 and A.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 $\mathsf{H}_{0,k}(\delta)$ and the asymptotic level under $\mathsf{H}_{0,\mathsf{CS}}(\delta)$.

Theorem A.1. *For a DR functional $\psi(\theta)$, under the conditions in Proposition 2.1, Proposition A.1 and Condition W, 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*

(A.9) $\qquad 2 - \Phi\left(\varsigma_k - \lim_{n\to\infty}(\gamma - \delta)\frac{\mathsf{s.e.}_\theta[\widehat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]}\right) - \Phi\left(\varsigma_k + \lim_{n\to\infty}(\gamma + \delta)\frac{\mathsf{s.e.}_\theta[\widehat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]}\right)$

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* $n \to \infty$.

($2'$) *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 $n \to \infty$.*

As an immediate consequence, under $\mathsf{H}_{0,\mathsf{CS}}(\delta)$, the rejection probability of $\widehat{\chi}_{2,k}(\Sigma_k^{-1}; z_{\alpha^\dagger/2}, \delta)$ is asymptotically no greater than $\alpha^\dagger$.

Remark A.3. *In Appendix A.5.1, we prove equation (A.9). We now prove that equation (A.9) implies Theorem A.1(1)-(2) and (2').*

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

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

  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 Proposition A.1 and equation (A.9) that the rejection probability of $\widehat{\chi}_{2,k}(\Sigma_k^{-1}; \varsigma_k, \delta)$, as $n \to \infty$, is no smaller than*

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

  where $\Theta(b, p, \widehat{b}, \widehat{p}, f, \bar{\mathsf{z}}_k)$ 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 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 A.1(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 A.1(2').*

A.5.1. *Proof of Theorem A.1.*

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

$$\lim_{n\to\infty} \mathsf{P}_\theta\left(\frac{|\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})|}{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]} - \varsigma_k\frac{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]}{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]} > \delta\right)$$

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

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

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

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

$\blacksquare$

A.6. **Derivation of the higher order influence functions of $\widetilde{\psi}_k(\theta)$ in Theorem A.3.** 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.

A.6.1. *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, Smucler, and Robins, 2021) under a (locally) nonparametric model of every order with the derivation deferred to Appendix A.6. First, we need to formally define HOIFs of a generic functional $\psi(\theta)$, not necessarily in the DR functional/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 P_θ.

Definition A.2. *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 $\left(o_{i_1},\cdots,o_{i_{m-1}}\right)$ 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

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

43

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 A.2 (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.*

A.6.2. *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 A.3. *Under the conditions of Proposition 2.1, 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 $\widehat{\mathbb{IF}}_{\ell\ell\to mm,k}(\Sigma_k^{-1}) := \sum_{j=\ell}^m \widehat{\mathbb{IF}}_{jj,k}(\Sigma_k^{-1})$,*

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

and for $j > 2$,

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

Here

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

and

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

(A.10) $$= (-1)^j \left[\mathcal{E}_{1,\widehat{b}}(\bar{z}_k)(O)\right]_{i_1}^\top \left\{\prod_{s=3}^j \Sigma_k^{-1} \left(\left[S_{bp}\bar{z}_k(X)\bar{z}_k(X)^\top\right]_{i_s} - \Sigma_k\right)\right\} \Sigma_k^{-1} \left[\mathcal{E}_{2,\widehat{p}}(\bar{z}_k)(O)\right]_{i_2}.$$

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 $\bar{i}_j := \{i_1, \ldots, i_j\}$.

Furthermore, define

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

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 $\mathsf{P}_{\widehat{\theta}}$.

Proof. Without loss of generality, we assume $\mathsf{P}_\theta(S_{bp} \geqslant 0) = 1$. First, let's recall the results from Proposition 2.1 (or Rotnitzky, Smucler, and Robins (2021, 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 the definition of Riesz representers, 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 first-order influence function. Recall from Definition A.1 that $\widetilde{\psi}_k(\theta)$ is of the following form:

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

Then

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

where the second equality follows from the definitions of $\widetilde{\overline{\zeta}}_{b,k}(\theta)$ and $\widetilde{\overline{\zeta}}_{p,k}(\theta)$ in equation (A.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:

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

where

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

and by Lemma A.2

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

Then

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

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 $\mathbb{V}_m[U_{i_1,\dots,i_m}]$ maps an m-th order U-statistic kernel $U_{i_1,\dots,i_m}$ to an m-th order U-statistic:

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

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

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

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

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

$$= \frac{1}{n(n-1)} \sum_{1 \leqslant i_1 \neq i_2 \leqslant n} [S_{bp}\widetilde{p}_{k,\theta}(X)\bar{\mathsf{z}}_k(X) + m_1(O, \bar{\mathsf{z}}_k)]_{i_1}^\top \Sigma_k^{-1} \left[S_{bp}\widetilde{b}_{k,\theta}(X)\bar{\mathsf{z}}_k(X) + m_2(O, \bar{\mathsf{z}}_k) \right]_{i_2}.$$

Hence formally we have

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

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

$$\mathsf{IF}_{jj,k,\bar{i}_j}(\Sigma_k^{-1})$$

$$= (-1)^j [S_{bp}\widetilde{p}_{k,\theta}(X)\bar{\mathsf{z}}_k(X) + m_1(O, \bar{\mathsf{z}}_k)]_{i_1}^\top \prod_{s=3}^{j} \Sigma_k^{-1} \left(\left[S_{bp}\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top \right]_{i_s} - \Sigma_k \right)$$

$$\times \ \Sigma_k^{-1} \left[S_{bp}\widetilde{b}_{k,\theta}(X)\bar{\mathsf{z}}_k(X) + m_2(O, \bar{\mathsf{z}}_k) \right]_{i_2}$$

and

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

Then

$$\mathbb{V}_{j+1} \left[\mathsf{IF}_{(j+1),(j+1),k,\bar{i}_{j+1}}(\Sigma_k^{-1}) \right]$$

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

Then the derivatives w.r.t. θ's in $\widetilde{\zeta}_{b,k}(\theta), \widetilde{\zeta}_{p,k}(\theta)$ and $(j-1)$ terms of Σ_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 Σ_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.2,

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

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

(A.11)
$$- (-1)^j [S_{bp}\widetilde{p}_{k,\theta}(X)\bar{\mathsf{z}}_k(X) + m_1(O, \bar{\mathsf{z}}_k)]_{i_1}^\top$$

$$\times \sum_{s=3}^{j+1} \Sigma_k^{-1} \left\{ \left(S_{bp}\bar{\mathsf{z}}_k(X)\bar{\mathsf{z}}_k(X)^\top \right)_{i_s} - \Sigma_k \right\}$$

$$\times \Sigma_k^{-1} \left[S_{bp}\widetilde{b}_{k,\theta}(X)\bar{\mathsf{z}}_k(X) + m_2(O, \bar{\mathsf{z}}_k) \right]_{i_2}$$

which is degenerate and thus orthogonal to $\mathcal{U}_j(\theta)$. Notice that equation (A.11) is exactly the kernel of $\mathbb{IF}_{(j+1),(j+1),k}(\Sigma_k^{-1})$. Then differentiating w.r.t. θ's in $\widetilde{p}_{k,\theta}$ and $\widetilde{b}_{k,\theta}$, we further obtain terms of the following form contributing to

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

$$(-1)^j \mathsf{IF}_{1,\widetilde{\bar{\zeta}}_{p,k},i_{j+1}}(\theta)^\top \left[S_{bp}\bar{z}_k(X)\bar{z}_k(X)^\top \right]_{i_1} \left[\sum_{s=3}^{j} \Sigma_k^{-1} \left\{ \left(S_{bp}\bar{z}_k(X)\bar{z}_k(X)^\top \right)_{i_s} - \Sigma_k \right\} \right]$$

$$\times \Sigma_k^{-1} \left[S_{bp}\widetilde{b}_{k,\theta}(X)\bar{z}_k(X) + m_2(O,\bar{z}_k) \right]_{i_2}$$

(A.12)

$$+ (-1)^j \left[S_{bp}\widetilde{p}_{k,\theta}(X)\bar{z}_k(X) + m_1(O,\bar{z}_k) \right]_{i_1}^\top \left[\sum_{s=3}^{j} \Sigma_k^{-1} \left\{ \left(S_{bp}\bar{z}_k(X)\bar{z}_k(X)^\top \right)_{i_s} - \Sigma_k \right\} \right]$$

$$\times \Sigma_k^{-1} \left[S_{bp}\bar{z}_k(X)\bar{z}_k(X)^\top \right]_{i_2} \mathsf{IF}_{1,\widetilde{\bar{\zeta}}_{b,k},i_{j+1}}(\theta).$$

Applying Lemma A.2 again and projecting the above display onto $\mathcal{U}_j^{\perp,j+1}(\theta)$, we obtain that

$$\Pi_\theta \left[\text{equation (A.12)} | \mathcal{U}_j^{\perp,j+1}(\theta) \right] \equiv 2 \times \text{equation (A.11)}.$$

Recall that there are $(j-1)$ terms of equation (A.11) in total contributing to $\mathsf{IF}_{1,\mathrm{if}_{jj,k,\bar{i}_j},i_{j+1}}(\theta)$, thus eventually there are $(j+1)$ terms of equation (A.11) in total contributing to $\mathbb{V}_{j+1}\left[\Pi_\theta \left[\mathsf{IF}_{1,\mathrm{if}_{jj,k,\bar{i}_j},i_{j+1}}(\theta) | \mathcal{U}_j^{\perp,j+1}(\theta) \right] \right]$. Further recall that

$$\mathbb{V}_{j+1} \left[\mathsf{IF}_{(j+1),(j+1),k,\bar{i}_{j+1}}(\Sigma_k^{-1}) \right]$$

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

Therefore we have

$$\mathbb{IF}_{(j+1),(j+1),k}(\Sigma_k^{-1}) \equiv \mathbb{V}_{j+1} \left[\mathsf{IF}_{(j+1),(j+1),k,\bar{i}_{j+1}}(\Sigma_k^{-1}) \right]$$

$$= \frac{1}{j+1} \mathbb{V}_{j+1} \left[(j+1) \times \text{equation (A.11)} \right]$$

$$= \mathbb{V}_{j+1} \left[\text{equation (A.11)} \right].$$

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

Lemma A.2. *As defined in equation* (A.1), *the influence functions of* $\widetilde{\bar{\zeta}}_{b,k}(\theta)$ *and* $\widetilde{\bar{\zeta}}_{p,k}(\theta)$ *are*

(A.13)

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

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

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

Then

$$\mathsf{IF}_{1,\widetilde{\bar{\zeta}}_{b,k}}(\theta)$$

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

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

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

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

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

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

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

$$= - \Sigma_k^{-1} \left(S_{bp} \widetilde{b}_{k,\theta}(X) \bar{z}_k(X) + m_2(O, \bar{z}_k) \right)$$

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 Proposition 2.1, $\widetilde{b}_{k,\widehat{\theta}} = \widehat{b}$ 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 Proposition 2.1 with θ 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 $\widehat{\mathbb{IF}}_{\ell\ell \to mm,k}(\Sigma_k^{-1}) := \sum_{j=\ell}^m \widehat{\mathbb{IF}}_{jj,k}(\Sigma_k^{-1})$,

(A.14)
$$\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1}) := \frac{1}{n(n-1)} \sum_{1 \leqslant i_1 \neq i_2 \leqslant n} \widehat{\mathsf{IF}}_{22,k,\bar{i}_2}(\Sigma_k^{-1})$$

and for $j > 2$,

(A.15)
$$\widehat{\mathbb{IF}}_{jj,k}(\Sigma_k^{-1}) := \frac{(n-j)!}{n!} \sum_{1 \leqslant i_1 \neq \ldots \neq i_j \leqslant n} \widehat{\mathsf{IF}}_{jj,k,\bar{i}_j}(\Sigma_k^{-1}).$$

Here

(A.16)
$$\widehat{\mathsf{IF}}_{22,k,\bar{i}_2}(\Sigma_k^{-1}) \equiv \mathsf{IF}_{22,\widetilde{\psi}_k,\bar{i}_2}(\widehat{\theta}) = \left[\mathcal{E}_{\widehat{b},m_2}(\bar{z}_k)(O) \right]_{i_1}^\top \Sigma_k^{-1} \left[\mathcal{E}_{\widehat{p},m_1}(\bar{z}_k)(O) \right]_{i_2},$$

and

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

(A.17)
$$= (-1)^j \left[\mathcal{E}_{\widehat{b},m_2}(\bar{z}_k)(O) \right]_{i_1}^\top \left\{ \prod_{s=3}^j \Sigma_k^{-1} \left(\left[S_{bp} \bar{z}_k(X) \bar{z}_k(X)^\top \right]_{i_s} - \Sigma_k \right) \right\} \Sigma_k^{-1} \left[\mathcal{E}_{\widehat{p},m_1}(\bar{z}_k)(O) \right]_{i_2}.$$

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 (A.4) and (A.5).

As a consequence of Theorem A.3, 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 $\text{Bias}_{\theta,k}(\widetilde{\psi}_1)$ is $\widehat{\mathbb{IF}}_{22 \to mm,k}(\Sigma_k^{-1})$, with the b and p in $\mathsf{IF}_{m,k}(\Sigma_k^{-1})$ and $\mathbb{IF}_{22 \to mm,k}(\Sigma_k^{-1})$ replaced by $\widehat{b}$ and $\widehat{p}$.

A.7. General formula of 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.3. *When $m \geqslant 3$, there exists a constant $C > 0$ that depends only on the spectra of Σ and $\widehat{\Sigma}$, such that*

(A.18)

$$
\mathsf{var}_\theta\left[\widehat{\mathbb{IF}}_{mm,k}(\widehat{\Sigma}_k^{-1})\right]
$$

$$
\lesssim \frac{1}{n}\left\{
\begin{array}{l}
m^{2m}\left(\frac{Ck}{n}\right)^{m-1} + \sum_{j=2}^{m-1} m^{2j}\left(\frac{Ck}{n}\right)^{j-1}\mathbb{L}_{\theta,2,\widehat{\Sigma},k}^{2(m-j-2)\vee 0}\left\{ \mathbb{L}_{\theta,2,\widehat{\Sigma},k}^{4} + \mathbb{L}_{\theta,2,\widehat{b},k}^{2}\mathbb{L}_{\theta,2,\widehat{\Sigma},k}^{2} + \mathbb{L}_{\theta,2,\widehat{p},k}^{2}\mathbb{L}_{\theta,2,\widehat{\Sigma},k}^{2} + \mathbb{L}_{\theta,4,\widehat{b},k}^{2}\mathbb{L}_{\theta,4,\widehat{p},k}^{2} \right\} \\
\quad + m^2\mathbb{L}_{\theta,2,\widehat{\Sigma},k}^{2(m-3)}\left\{\mathbb{L}_{\theta,2,\widehat{b},k}^{2}\mathbb{L}_{\theta,2,\widehat{\Sigma},k}^{2} + \mathbb{L}_{\theta,2,\widehat{p},k}^{2}\mathbb{L}_{\theta,2,\widehat{\Sigma},k}^{2} + (\mathbb{L}_{\theta,4,\widehat{b},k}^{2}\mathbb{L}_{\theta,4,\widehat{p},k}^{2} \wedge \mathbb{L}_{\theta,2,\widehat{b},k}^{2}\mathbb{L}_{\theta,\infty,\widehat{p},k}^{2} \wedge \mathbb{L}_{\theta,\infty,\widehat{b},k}^{2}\mathbb{L}_{\theta,2,\widehat{p},k}^{2})\right\}
\end{array}
\right\}
$$

$$
\stackrel{\textit{Condition } W}{\lesssim} \frac{1}{n}\left\{
\begin{array}{l}
m^{2m}\left(\frac{Ck}{n}\right)^{m-1} + \sum_{j=2}^{m-1} m^{2j}\left(\frac{Ck}{n}\right)^{j-1}\mathbb{L}_{\theta,2,\widehat{\Sigma},k}^{2(m-j-2)\vee 0}\left\{ \mathbb{L}_{\theta,2,\widehat{\Sigma},k}^{4} + \mathbb{L}_{\theta,2,\widehat{b},k}^{2}\mathbb{L}_{\theta,2,\widehat{\Sigma},k}^{2} + \mathbb{L}_{\theta,2,\widehat{p},k}^{2}\mathbb{L}_{\theta,2,\widehat{\Sigma},k}^{2} + k\mathbb{L}_{\theta,2,\widehat{b},k}^{2}\mathbb{L}_{\theta,2,\widehat{p},k}^{2} \right\} \\
\quad + m^2\mathbb{L}_{\theta,2,\widehat{\Sigma},k}^{2(m-3)}\left\{\mathbb{L}_{\theta,2,\widehat{b},k}^{2}\mathbb{L}_{\theta,2,\widehat{\Sigma},k}^{2} + \mathbb{L}_{\theta,2,\widehat{p},k}^{2}\mathbb{L}_{\theta,2,\widehat{\Sigma},k}^{2} + k\mathbb{L}_{\theta,2,\widehat{b},k}^{2}\mathbb{L}_{\theta,2,\widehat{p},k}^{2}\right\}
\end{array}
\right\}
$$

$$
\stackrel{\textit{Condition } S}{\lesssim} \frac{1}{n}\left\{
\begin{array}{l}
m^{2m}\left(\frac{Ck}{n}\right)^{m-1} + \sum_{j=2}^{m-1} m^{2j}\left(\frac{Ck}{n}\right)^{j-1}\mathbb{L}_{\theta,2,\widehat{\Sigma},k}^{2(m-j-2)\vee 0}\left\{ \begin{array}{c}\mathbb{L}_{\theta,2,\widehat{\Sigma},k}^{4} + \mathbb{L}_{\theta,2,\widehat{b},k}^{2}\mathbb{L}_{\theta,2,\widehat{\Sigma},k}^{2} + \mathbb{L}_{\theta,2,\widehat{p},k}^{2}\mathbb{L}_{\theta,2,\widehat{\Sigma},k}^{2} \\ + \left\{\mathbb{L}_{\theta,2,\widehat{b},k}^{2}\mathbb{L}_{\theta,\infty,\widehat{p},k}^{2} \wedge \mathbb{L}_{\theta,\infty,\widehat{b},k}^{2}\mathbb{L}_{\theta,2,\widehat{p},k}^{2}\right\}\end{array}\right\} \\
\quad + m^2\mathbb{L}_{\theta,2,\widehat{\Sigma},k}^{2(m-3)}\left\{\mathbb{L}_{\theta,2,\widehat{b},k}^{2}\mathbb{L}_{\theta,2,\widehat{\Sigma},k}^{2} + \mathbb{L}_{\theta,2,\widehat{p},k}^{2}\mathbb{L}_{\theta,2,\widehat{\Sigma},k}^{2} + \left\{\mathbb{L}_{\theta,2,\widehat{b},k}^{2}\mathbb{L}_{\theta,\infty,\widehat{p},k}^{2} \wedge \mathbb{L}_{\theta,\infty,\widehat{b},k}^{2}\mathbb{L}_{\theta,2,\widehat{p},k}^{2}\right\}\right\}
\end{array}
\right\}.
$$

Proof. The detailed calculations can be found in the Appendix of Liu et al. (2017). The two different versions of (A.18) are due to the application of Hölder inequality with different Hölder conjugate pairs (i.e. $(2,2)$ and $(1,\infty)$). When Condition S does not hold but Condition W holds, we have

$$
\mathbb{L}_{\theta,4,\widehat{b},k}^{2}\mathbb{L}_{\theta,4,\widehat{p},k}^{2} \leqslant k\mathbb{L}_{\theta,2,\widehat{b},k}^{2}\mathbb{L}_{\theta,2,\widehat{p},k}^{2},
$$

hence the second line inequality in (A.18) under only Condition W. But this upper bound may not be tight. For example, when $\bar{z}_k$ is Fourier or polynomials series that violates Condition S yet satisfies Condition W (Belloni et al., 2015),

$$
\mathbb{L}_{\theta,4,\widehat{b},k}^{2}\mathbb{L}_{\theta,4,\widehat{p},k}^{2} \leqslant \log k\,\mathbb{L}_{\theta,2,\widehat{b},k}^{2}\mathbb{L}_{\theta,2,\widehat{p},k}^{2}
$$

because $\mathbb{L}_{\theta,\infty,\widehat{b},k}, \mathbb{L}_{\theta,\infty,\widehat{p},k} \lesssim \log k$ when $\mathbb{L}_{\theta,\infty,\widehat{b}}, \mathbb{L}_{\theta,\infty,\widehat{p}}$ are $O(1)$. But for Legendre polynomial series which also violates Condition S yet satisfies Condition W (Belloni et al., 2015), we currently can only have the following bound

$$
\mathbb{L}_{\theta,4,\widehat{b},k}^{2}\mathbb{L}_{\theta,4,\widehat{p},k}^{2} \leqslant k\mathbb{L}_{\theta,2,\widehat{b},k}^{2}\mathbb{L}_{\theta,2,\widehat{p},k}^{2}.
$$

∎

A.8. Bootstrapping higher order influence functions. In this section, we will use higher order moments of multinomial distributions frequently. Two good references are Mosimann (1962) and Newcomer, Neerchal, and Morel (2008). For notational convenience, we suppress the dependence on Σ_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 Σ_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

$$(A.19) \qquad \widehat{\mathsf{var}}\big[\widehat{\mathbb{IF}}_{22,k}\big] := \underbrace{\frac{1}{M-1}\sum_{m=1}^{M}\left(\widehat{\widehat{\mathbb{IF}}}_{22,k}^{(m)} - \frac{1}{M}\sum_{m=1}^{M}\widehat{\widehat{\mathbb{IF}}}_{22,k}^{(m)}\right)^2}_{T_1} - \underbrace{\frac{2}{M-1}\sum_{m=1}^{M}\left(\widehat{\widehat{\mathbb{IF}}}_{22,k}^{(m),c} - \frac{1}{M}\sum_{m=1}^{M}\widehat{\widehat{\mathbb{IF}}}_{22,k}^{(m),c}\right)^2}_{T_2}$$

where

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

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

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

Remark A.4. *We now explain why* $\widehat{\mathsf{var}}\big[\widehat{\mathbb{IF}}_{22,k}\big]$ *is proposed as an estimator of* $\mathsf{var}_\theta\big[\widehat{\mathbb{IF}}_{22,k}\big]$. *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:*

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

$$= \mathsf{E}_{\boldsymbol{W}}\left[\left(\frac{1}{n(n-1)}\sum_{1\leqslant i_1\neq i_2\leqslant n} W_{i_1}^{(m)}W_{i_2}^{(m)}\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\right)^2 |\{O_i\}_{i=1}^N\right]$$

$$- \left(\mathsf{E}_{\boldsymbol{W}}\left[\frac{1}{n(n-1)}\sum_{1\leqslant i_1\neq i_2\leqslant n} W_{i_1}^{(m)}W_{i_2}^{(m)}\widehat{\mathsf{IF}}_{22,k,i_1,i_2}|\{O_i\}_{i=1}^N\right]\right)^2$$

$$= \frac{1}{n^2(n-1)^2}\sum_{1\leqslant i_1\neq i_2\leqslant n}\{\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_1,i_2} + \widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_2,i_1}\}\{\mathsf{E}_{\boldsymbol{W}}[W_{i_1}^2W_{i_2}^2] - (\mathsf{E}_{\boldsymbol{W}}[W_{i_1}W_{i_2}])^2\}$$

$$+ \frac{1}{n^2(n-1)^2}\sum_{1\leqslant i_1\neq i_2\neq i_3\leqslant n}\left\{\begin{array}{l} \widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_1,i_3}\{\mathsf{E}_{\boldsymbol{W}}[W_{i_1}^2 W_{i_2}W_{i_3}] - \mathsf{E}_{\boldsymbol{W}}[W_{i_1}W_{i_2}]\mathsf{E}_{\boldsymbol{W}}[W_{i_1}W_{i_3}]\} \\ + \widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_3,i_1}\{\mathsf{E}_{\boldsymbol{W}}[W_{i_1}^2 W_{i_2}W_{i_3}] - \mathsf{E}_{\boldsymbol{W}}[W_{i_1}W_{i_2}]\mathsf{E}_{\boldsymbol{W}}[W_{i_3}W_{i_1}]\} \\ + \widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_2,i_3}\{\mathsf{E}_{\boldsymbol{W}}[W_{i_1}W_{i_2}^2 W_{i_3}] - \mathsf{E}_{\boldsymbol{W}}[W_{i_1}W_{i_2}]\mathsf{E}_{\boldsymbol{W}}[W_{i_2}W_{i_3}]\} \\ + \widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_3,i_2}\{\mathsf{E}_{\boldsymbol{W}}[W_{i_1}W_{i_2}^2 W_{i_3}] - \mathsf{E}_{\boldsymbol{W}}[W_{i_1}W_{i_2}]\mathsf{E}_{\boldsymbol{W}}[W_{i_3}W_{i_2}]\} \end{array}\right\}$$

$$+ \frac{1}{n^2(n-1)^2}\sum_{1\leqslant i_1\neq i_2\neq i_3\neq i_4\leqslant n}\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_3,i_4}\{\mathsf{E}_{\boldsymbol{W}}[W_{i_1}W_{i_2}W_{i_3}W_{i_4}] - \mathsf{E}_{\boldsymbol{W}}[W_{i_1}W_{i_2}]\mathsf{E}_{\boldsymbol{W}}[W_{i_3}W_{i_4}]\}.$$

Plugging in the following higher order moments of $\mathsf{Multinom}(n,\underbrace{n^{-1},\dots,n^{-1}}_{(n-1)'s})$ *(Newcomer, Neerchal, and Morel, 2008):*

$$\mathsf{E}_{\boldsymbol{W}}[W_1W_2] = \frac{n-1}{n}, \mathsf{E}_{\boldsymbol{W}}[W_1W_2W_3] = \frac{(n-1)(n-2)}{n^2}, \mathsf{E}_{\boldsymbol{W}}[W_1W_2W_3W_4] = \frac{(n-1)(n-2)(n-3)}{n^3},$$

$$\mathsf{E}_{\boldsymbol{W}}[W_1^2W_2] = \frac{2(n-1)^2}{n^2}, \mathsf{E}_{\boldsymbol{W}}[W_1^2] = \frac{2n-1}{n},$$

$$\mathsf{E}_{\boldsymbol{W}}[W_1^2W_2W_3] = \frac{(n-1)(n-2)(n-3)+n(n-1)(n-2)}{n^3}$$

51

$$= \frac{(n-1)(2n^2 - 7n + 6)}{n^3},$$

$$\mathsf{E}_{\boldsymbol{W}}[W_1^2 W_2^2] = \frac{(n-1)(n-2)(n-3) + 2n(n-1)(n-2) + n^2(n-1)}{n^3}$$

$$= \frac{(n-1)(4n^2 - 9n + 6)}{n^3},$$

we have

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

$$(A.20) \qquad = \frac{1}{n^2(n-1)^2} \sum_{1 \leqslant i_1 \neq i_2 \leqslant n} \{\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_1,i_2} + \widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_2,i_1}\} \left(3 - \frac{11}{n} + \frac{14}{n^2} - \frac{6}{n^3}\right)$$

$$+ \frac{1}{n^2(n-1)^2} \sum_{1 \leqslant i_1 \neq i_2 \neq i_3 \leqslant n} \left\{ \begin{array}{l} \widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_1,i_3} + \widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_3,i_1} \\ + \widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_2,i_3} + \widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_3,i_2} \end{array} \right\} \left(1 - \frac{7}{n} + \frac{12}{n^2} - \frac{6}{n^3}\right)$$

$$+ \frac{1}{n^2(n-1)^2} \sum_{1 \leqslant i_1 \neq i_2 \neq i_3 \neq i_4 \leqslant n} \widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_3,i_4} \left(-\frac{4}{n} + \frac{10}{n^2} - \frac{6}{n^3}\right)$$

The term T_2 is the bootstrap variance estimator for second order degenerate U-statistics and it has been proposed previously in Arcones and Giné (1992) and 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:

$$\mathsf{E}_{\boldsymbol{W}}[T_2 | \{O_i\}_{i=1}^N]$$

$$= \frac{2}{n^2(n-1)^2} \sum_{1 \leqslant i_1 \neq i_2 \leqslant n} \{\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_1,i_2} + \widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_2,i_1}\}\{\mathsf{E}_{\boldsymbol{W}}[(W_{i_1}-1)^2(W_{i_2}-1)^2] - (\mathsf{E}_{\boldsymbol{W}}[(W_{i_1}-1)(W_{i_2}-1)])^2\}$$

$$+ \frac{2}{n^2(n-1)^2} \sum_{1 \leqslant i_1 \neq i_2 \neq i_3 \leqslant n} \left\{ \begin{array}{l} \widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_1,i_3}\{\mathsf{E}_{\boldsymbol{W}}[(W_{i_1}-1)^2(W_{i_2}-1)(W_{i_3}-1)] - \mathsf{E}_{\boldsymbol{W}}[(W_{i_1}-1)(W_{i_2}-1)]\mathsf{E}_{\boldsymbol{W}}[(W_{i_1}-1)(W_{i_3}-1)]\} \\ + \widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_3,i_1}\{\mathsf{E}_{\boldsymbol{W}}[(W_{i_1}-1)^2(W_{i_2}-1)(W_{i_3}-1)] - \mathsf{E}_{\boldsymbol{W}}[(W_{i_1}-1)(W_{i_2}-1)]\mathsf{E}_{\boldsymbol{W}}[(W_{i_3}-1)(W_{i_1}-1)]\} \\ + \widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_2,i_3}\{\mathsf{E}_{\boldsymbol{W}}[(W_{i_1}-1)(W_{i_2}-1)^2(W_{i_3}-1)] - \mathsf{E}_{\boldsymbol{W}}[(W_{i_1}-1)(W_{i_2}-1)]\mathsf{E}_{\boldsymbol{W}}[(W_{i_2}-1)(W_{i_3}-1)]\} \\ + \widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_3,i_2}\{\mathsf{E}_{\boldsymbol{W}}[(W_{i_1}-1)(W_{i_2}-1)^2(W_{i_3}-1)] - \mathsf{E}_{\boldsymbol{W}}[(W_{i_1}-1)(W_{i_2}-1)]\mathsf{E}_{\boldsymbol{W}}[(W_{i_3}-1)(W_{i_2}-1)]\} \end{array} \right\}$$

$$+ \frac{2}{n^2(n-1)^2} \sum_{1 \leqslant i_1 \neq i_2 \neq i_3 \neq i_4 \leqslant n} \widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_3,i_4}\{\mathsf{E}_{\boldsymbol{W}}[(W_{i_1}-1)(W_{i_2}-1)(W_{i_3}-1)(W_{i_4}-1)] - \mathsf{E}_{\boldsymbol{W}}[(W_{i_1}-1)(W_{i_2}-1)]\mathsf{E}_{\boldsymbol{W}}[(W_{i_3}-1)(W_{i_4}-1)]\}$$

$$= \frac{2}{n^2(n-1)^2} \sum_{1 \leqslant i_1 \neq i_2 \leqslant n} \{\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_1,i_2} + \widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_2,i_1}\}\{\mathsf{E}_{\boldsymbol{W}}[(W_1^2 - 2W_1 + 1)(W_2^2 - 2W_2 + 1)] - (\mathsf{E}_{\boldsymbol{W}}[(W_1-1)(W_2-1)])^2\}$$

$$+ \frac{2}{n^2(n-1)^2} \sum_{1 \leqslant i_1 \neq i_2 \neq i_3 \leqslant n} \left\{ \begin{array}{l} \widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_1,i_3} + \widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_3,i_1} \\ + \widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_2,i_3} + \widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_3,i_2} \end{array} \right\}\{\mathsf{E}_{\boldsymbol{W}}[(W_1-1)^2(W_2-1)(W_3-1)] - (\mathsf{E}_{\boldsymbol{W}}[(W_1-1)(W_2-1)])^2\}$$

$$+ \frac{2}{n^2(n-1)^2} \sum_{1 \leqslant i_1 \neq i_2 \neq i_3 \neq i_4 \leqslant n} \widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_3,i_4}\{\mathsf{E}_{\boldsymbol{W}}[(W_1-1)(W_2-1)(W_3-1)(W_4-1)] - (\mathsf{E}_{\boldsymbol{W}}[(W_1-1)(W_2-1)])^2\}$$

$$= \frac{2}{n^2(n-1)^2} \sum_{1 \leqslant i_1 \neq i_2 \leqslant n} \{\widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_1,i_2} + \widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_2,i_1}\} \left(1 - \frac{3}{n} + \frac{7}{n^2} - \frac{6}{n^3} - \frac{1}{n^2}\right)$$

$$+ \frac{2}{n^2(n-1)^2} \sum_{1 \leqslant i_1 \neq i_2 \neq i_3 \leqslant n} \left\{ \begin{array}{l} \widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_1,i_3} + \widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_3,i_1} \\ + \widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_2,i_3} + \widehat{\mathsf{IF}}_{22,k,i_1,i_2}\widehat{\mathsf{IF}}_{22,k,i_3,i_2} \end{array} \right\} \left(-\frac{1}{n} + \frac{5}{n^2} - \frac{6}{n^3} - \frac{1}{n^2}\right)$$

$$+ \frac{2}{n^2(n-1)^2} \sum_{1 \leqslant i_1 \neq i_2 \neq i_3 \neq i_4 \leqslant n} \widehat{\mathsf{IF}}_{22,k,i_1,i_2} \widehat{\mathsf{IF}}_{22,k,i_3,i_4} \left(\frac{3}{n^2} - \frac{6}{n^3} - \frac{1}{n^2} \right).$$

Hence T_2 serves as a correction term of the inflated factor 3 appeared in equation (A.20).

Combining the above computations, we have

$$\mathsf{E}_{\boldsymbol{W}}[\widehat{\mathrm{var}}[\widehat{\mathbb{IF}}_{22,k}]|\{O_i\}_{i=1}^N] = \mathsf{E}_{\boldsymbol{W}}[T_1|\{O_i\}_{i=1}^N] - \mathsf{E}_{\boldsymbol{W}}[T_2|\{O_i\}_{i=1}^N]$$

$$= \frac{1}{n^2(n-1)^2} \sum_{1 \leqslant i_1 \neq i_2 \leqslant n} \{\widehat{\mathsf{IF}}_{22,k,i_1,i_2} \widehat{\mathsf{IF}}_{22,k,i_1,i_2} + \widehat{\mathsf{IF}}_{22,k,i_1,i_2} \widehat{\mathsf{IF}}_{22,k,i_2,i_1}\} \left(1 - \frac{5}{n} + \frac{2}{n^2} + \frac{6}{n^3} \right)$$

$$+ \frac{1}{n^2(n-1)^2} \sum_{1 \leqslant i_1 \neq i_2 \neq i_3 \leqslant n} \left\{ \begin{array}{c} \widehat{\mathsf{IF}}_{22,k,i_1,i_2} \widehat{\mathsf{IF}}_{22,k,i_1,i_3} + \widehat{\mathsf{IF}}_{22,k,i_1,i_2} \widehat{\mathsf{IF}}_{22,k,i_3,i_1} \\ + \widehat{\mathsf{IF}}_{22,k,i_1,i_2} \widehat{\mathsf{IF}}_{22,k,i_2,i_3} + \widehat{\mathsf{IF}}_{22,k,i_1,i_2} \widehat{\mathsf{IF}}_{22,k,i_3,i_2} \end{array} \right\} \left(1 - \frac{5}{n} + \frac{6}{n^3} \right)$$

$$- \frac{4}{n} \frac{1}{n^2(n-1)^2} \sum_{1 \leqslant i_1 \neq i_2 \neq i_3 \neq i_4 \leqslant n} \widehat{\mathsf{IF}}_{22,k,i_1,i_2} \widehat{\mathsf{IF}}_{22,k,i_3,i_4} \left(1 - \frac{1.5}{n} - \frac{1.5}{n^2} \right).$$

Finally, it is not difficult to see that

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

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

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

Proof. Let the $\widetilde{\mathrm{var}}[\widehat{\mathbb{IF}}_{22,k}]$ be the limit (in $\mathsf{P}_{\boldsymbol{W}}$-probability) of $\widehat{\mathrm{var}}[\widehat{\mathbb{IF}}_{22,k}]$ as $M \to \infty$ given all the observed data. Using the moments of multinomial distributions (Newcomer, Neerchal, and Morel, 2008), together with the explanation in Comment A.4, it is easy to show that $\mathsf{E}_\theta \left[\frac{\widetilde{\mathrm{var}}[\widehat{\mathbb{IF}}_{22,k}]}{\mathrm{var}_\theta[\widehat{\mathbb{IF}}_{22,k}]} \right] = 1 + o(1)$ and $\mathrm{var}_\theta \left[\frac{\widetilde{\mathrm{var}}[\widehat{\mathbb{IF}}_{22,k}]}{\mathrm{var}_\theta[\widehat{\mathbb{IF}}_{22,k}]} \right] = o(1)$. Combining the above two claims, we have $\frac{\widetilde{\mathrm{var}}[\widehat{\mathbb{IF}}_{22,k}]}{\mathrm{var}_\theta[\widehat{\mathbb{IF}}_{22,k}]} = 1 + o_{\mathsf{P}_\theta}(1)$. ∎

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

$$
\widehat{\mathrm{var}}\big[\widehat{\widehat{\mathbb{IF}}}_{33,k}\big] := \underbrace{\frac{1}{M-1}\sum_{m=1}^{M}\left(\widehat{\widehat{\mathbb{IF}}}_{33,k}^{(m)} - \frac{1}{M}\sum_{m=1}^{M}\widehat{\widehat{\mathbb{IF}}}_{33,k}^{(m)}\right)^2}_{S_1}
$$

$$
-\underbrace{\frac{1}{M-1}\sum_{m=1}^{M}\left(\widehat{\widehat{\mathbb{IF}}}_{33,k}^{(m),c(b,p)} - \frac{1}{M}\sum_{m=1}^{M}\widehat{\widehat{\mathbb{IF}}}_{33,k}^{(m),c(b,p)}\right)^2}_{S_2}
$$

(A.21)

$$
-\underbrace{\frac{1}{M-1}\sum_{m=1}^{M}\left(\widehat{\widehat{\mathbb{IF}}}_{33,k}^{(m),c(b,\Sigma)} - \frac{1}{M}\sum_{m=1}^{M}\widehat{\widehat{\mathbb{IF}}}_{33,k}^{(m),c(b,\Sigma)}\right)^2}_{S_3}
$$

$$
-\underbrace{\frac{1}{M-1}\sum_{m=1}^{M}\left(\widehat{\widehat{\mathbb{IF}}}_{33,k}^{(m),c(p,\Sigma)} - \frac{1}{M}\sum_{m=1}^{M}\widehat{\widehat{\mathbb{IF}}}_{33,k}^{(m),c(p,\Sigma)}\right)^2}_{S_4}
$$

where

$$
\widehat{\widehat{\mathbb{IF}}}_{33,k}^{(m)} = \frac{1}{n(n-1)(n-2)}\sum_{1\leqslant i_1\neq i_2\neq i_3\leqslant n}W_{i_1}^{(m)}W_{i_2}^{(m)}W_{i_3}^{(m)}\widehat{\mathsf{IF}}_{33,k,\bar{i}_3}
$$

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

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

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

We then have:

Proposition A.3. *Under Condition* W, *as* $n\to\infty$ *and* $M\to\infty$,

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

Finally, we consider estimating $\mathrm{var}_\theta\big[\widehat{\widehat{\mathbb{IF}}}_{22\to33,k}\big]$ through nonparametric bootstrap. It is not difficult to see that the following estimator has the desired property $\frac{\widehat{\mathrm{var}}[\widehat{\widehat{\mathbb{IF}}}_{22\to33,k}]}{\mathrm{var}_\theta[\widehat{\widehat{\mathbb{IF}}}_{22\to33,k}]} = 1 + o_{\mathsf{P}_\theta}(1)$ as $M\to\infty$:

(A.22)
$$
\widehat{\mathrm{var}}\big[\widehat{\widehat{\mathbb{IF}}}_{22\to33,k}\big] = \widehat{\mathrm{var}}\big[\widehat{\widehat{\mathbb{IF}}}_{22,k}\big] + \widehat{\mathrm{var}}\big[\widehat{\widehat{\mathbb{IF}}}_{33,k}\big] + 2\widehat{\mathrm{cov}}\big[\widehat{\widehat{\mathbb{IF}}}_{22,k},\widehat{\widehat{\mathbb{IF}}}_{33,k}\big]
$$

where

$$
\widehat{\mathrm{cov}}\big[\widehat{\widehat{\mathbb{IF}}}_{22,k},\widehat{\widehat{\mathbb{IF}}}_{33,k}\big] = \frac{1}{M-1}\sum_{m=1}^{M}\left(\widehat{\widehat{\mathbb{IF}}}_{22,k}^{(m)} - \frac{1}{M}\sum_{m'=1}^{M}\widehat{\widehat{\mathbb{IF}}}_{22,k}^{(m')}\right)\left(\widehat{\widehat{\mathbb{IF}}}_{33,k}^{(m)} - \frac{1}{M}\sum_{m'=1}^{M}\widehat{\widehat{\mathbb{IF}}}_{33,k}^{(m')}\right)
$$

$$-\frac{2}{M-1}\sum_{m=1}^{M-1}\left(\widehat{\mathbb{IF}}_{22,k}^{(m),c}-\frac{1}{M}\sum_{m'=1}^{M}\widehat{\mathbb{IF}}_{22,k}^{(m'),c}\right)\left(\widehat{\mathbb{IF}}_{33,k}^{(m),c(b,p)}-\frac{1}{M}\sum_{m'=1}^{M}\widehat{\mathbb{IF}}_{33,k}^{(m'),c(b,p)}\right).$$

It is straightforward but tedious to generalize the above arguments to construct bootstrapped estimators of $\mathrm{var}_\theta[\widehat{\mathbb{IF}}_{mm,k}]$ for general $m\geqslant 2$. We plan to report the general construction in a separate paper. But when estimating $\mathrm{var}_\theta[\widehat{\mathbb{IF}}_{22\to mm,k}]$, it is sufficient to use $\widehat{\mathrm{var}}[\widehat{\mathbb{IF}}_{22\to 33,k}]$ as it dominates the variance of higher-order terms.

A.9. **Proof of Theorem 3.2 and Theorem 4.2.** As discussed in Sections 3 and 4.3, since Σ_k^{-1} is generally unknown, $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})$ needs to be replaced by $\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})$. We consider the following statistic used in the test $\widehat{\chi}_{m,k}(\widehat{\Sigma}_k^{-1};z_{\alpha^\dagger/2},\delta)$ after standardization:

(A.23)

$$\frac{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]}\left(\frac{\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})}{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]}-\delta\right)$$

$$=\left(\frac{\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})-\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)}{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]}\frac{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]}\right)+\frac{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]}\left(\frac{\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)}{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]}-\delta\right)$$

$$=\left(\frac{\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})-\mathsf{E}_\theta[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]}{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]}\frac{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]}\right)+\frac{\mathsf{EB}_{\theta,m,k}}{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]}\frac{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]}$$

$$+\frac{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]}{\widehat{\mathsf{s.e.}}[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]}\left(\frac{\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)}{\widehat{\mathsf{s.e.}}[\widehat{\psi}_1]}-\delta\right)$$

$$=\left\{\left(\frac{\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})-\mathsf{E}_\theta[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]}\right)+\left(\frac{\mathsf{Bias}_{\theta,k}(\widehat{\psi}_1)+\mathsf{EB}_{\theta,m,k}}{\mathsf{s.e.}_\theta[\widehat{\psi}_1]}-\delta\right)\frac{\mathsf{s.e.}_\theta[\widehat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]}\right\}(1+o_{\mathsf{P}_\theta}(1))$$

$$=\left\{\left(\frac{\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})-\mathsf{E}_\theta[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]}\right)+\left(\gamma+\frac{\mathsf{EB}_{\theta,m,k}}{\mathsf{s.e.}_\theta[\widehat{\psi}_1]}-\delta\right)\frac{\mathsf{s.e.}_\theta[\widehat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]}\right\}(1+o_{\mathsf{P}_\theta}(1))$$

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

The effect of estimating Σ_k^{-1} on the asymptotic validity of the test $\widehat{\chi}_{m,k}(\widehat{\Sigma}_k^{-1};z_{\alpha^\dagger/2},\delta)$ of $\mathsf{H}_{0,k}(\delta)$ thus depends on the orders of terms A and B. A has variance 1 and mean $-(\delta-\gamma)\frac{\mathsf{s.e.}_\theta[\widehat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]}$. B depends on the estimation bias due to estimating Σ_k^{-1} by $\widehat{\Sigma}_k^{-1}$. Hence if we have:

(1) $\frac{\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})-\mathsf{E}_\theta[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]}$ is asymptotically $N(0,1)$ conditional on the training sample;

(2) $\frac{\mathsf{s.e.}_\theta[\widehat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]}\Big/\frac{\mathsf{s.e.}_\theta[\widehat{\psi}_1]}{\mathsf{s.e.}_\theta[\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})]}\to 1$;

(3) $B=o(1)$ under $\mathsf{H}_{0,k}(\delta)$ or fixed alternatives to $\mathsf{H}_{0,k}(\delta)$;

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

$\widehat{\chi}_{m,k}(\widehat{\Sigma}_k^{-1};z_{\alpha^\dagger/2},\delta)$ is an asymptotically level $\alpha^\dagger$ 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 Theorem 4.1, both (1) and (2) hold for $\widehat{\mathbb{IF}}_{22\to mm,k}(\widehat{\Sigma}_k^{-1})$. In terms of (3) and (3'), under the conditions of Theorem 4.2:

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

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

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

Thus (3') is satisfied.

In summary, $\widehat{\chi}_{m,k}(\widehat{\Sigma}_k^{-1}; z_{\alpha^\dagger/2}, \delta)$ is an asymptotically valid level $\alpha^\dagger$ 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 5 are of the following forms:

$$
\text{(B.1)} \qquad h_f(x; s_f) \propto 1 + \exp\left\{\frac{1}{2}\sum_{j\in\mathcal{J},\ell\in\mathbb{Z}} 2^{-j(s_f+0.5)}\omega_{j,\ell}(x)\right\},
$$

$$
\text{(B.2)} \qquad h_b(x; s_b) = \sum_{j\in\mathcal{J},\ell\in\mathbb{Z}} 2^{-j(s_b+0.5)}\omega_{j,\ell}(x),
$$

$$
\text{(B.3)} \qquad h_p(x; s_p) = -2\sum_{j\in\mathcal{J},\ell\in\mathbb{Z}} 2^{-j(s_p+0.5)}\omega_{j,\ell}(x)
$$

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 ℓ. By the equivalent characterization of Besov-Triebel spaces by the corresponding wavelet coefficients in the frequency domain (Giné and Nickl, 2016, page 331), $h_f(\cdot; s_f) \in \text{Hölder}(s_f)$, $h_b(\cdot; s_b) \in \text{Hölder}(s_b)$ and $h_p(\cdot; s_p) \in \text{Hölder}(s_p)$. 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 π in Section 5.

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 5.

B.1. Generating correlated multidimensional covariates X with fixed non-smooth marginal densities. In the simulation study conducted in Section 5, 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

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

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.

Appendix C. Proximal higher-order influence functions

In this section, we further generalize the derivations in Appendix A.6 for $\psi(\theta) = -\mathsf{E}_\theta[Y(a)]$ to the case under the proximal causal inference framework considered in Miao, Shi, and Tchetgen Tchetgen (2018), Tchetgen Tchetgen et al. (2020), and Cui et al. (2020). The purpose of this section is to show that our framework can be extended to the endogeneity settings which economists are extremely interested in; e.g. **newey1991uniform**; Newey and Powell (2003). We only consider $\psi(\theta) = -\mathsf{E}_\theta[Y(a)]$, since it is yet unclear what is the "right" generalization under endogeneity for the general DR functionals considered in Rotnitzky, Smucler, and Robins (2021). We leave such generalization to future work.

We immediately begin with the formal setup, followed with a theorem on the forms of HOIFs of $\psi(\theta)$. Instead of observing i.i.d. $O_i = (X_i, A_i, Y_i)_{i=1}^n$ under strong ignorability, there might exist latent confounding variable U, so $\psi(\theta)$ is not point identified in general. But suppose we observe $O_i = (X_i, Z_i, A_i, W_i, Y_i)_{i=1}^n$, where W and Z are "proxy variables" of Y and A respectively, satisfying the following proximal causal identification conditions given in Tchetgen Tchetgen et al. (2020):

Condition C.1.

(1) *Consistency: $Y = Y_A$ almost surely;*

(2) *No direct effect from the treatment A and treatment proxy Z to the outcome proxy W: $W_{a,z} = W$ for all a and z almost surely;*

(3) *No direct effect from the treatment proxy Z to the outcome Y: $Y_{a,z} = Y_a$ for all a and z almost surely;*

(4) $(Z, A) \perp\!\!\!\perp (Y_a, W)|(U, X)$ *for* $a = 0, 1$;

(5) $0 < \Pr(A = a|U, X) < 1$ *almost surely for* $a = 0, 1$;

(6) *For any* a, x, *if* $\mathsf{E}_\theta[g(U)|W, A = a, X = x] = 0$ *almost surely, then* $g(U) = 0$ *almost surely;*

(7) *For any* a, x, *if* $\mathsf{E}_\theta[g(W)|Z, A = a, X = x] = 0$ *almost surely, then* $g(W) = 0$ *almost surely;*

For more details of proximal causal inference and the intuition behind the above regularity conditions, we refer interested readers to many recent papers on this topic (Miao, Shi, and Tchetgen Tchetgen, 2018; Tchetgen Tchetgen et al., 2020; Cui et al., 2020; Kallus, Mao, and Uehara, 2021; Shpitser, Wood-Doughty, and Tchetgen Tchetgen, 2021).

Then we have

Theorem C.1. *Under the above Condition C.1 (Assumptions 1, and 4-9 of Cui et al. (2020)), together with the following conditions on the so-called confounding bridge functions r and q (Assumption 10 of Cui et al. (2020)): Denote $\Psi(z, a, x) \equiv$ $\mathsf{E}_\theta[(Y - r(W, A = a, X))^2|Z = z, A = a, X = x]$ and $\Gamma(w, a, x) \equiv \mathsf{E}_\theta\left[\left(q(Z, A, X) - \frac{1}{f(A|W,X)}\right)^2 |W = w, A = a, X = x\right]$:*

(1) $0 < \inf\limits_{z,a,x} \Psi(z, a, x) \leqslant \sup\limits_{z,a,x} \Psi(z, a, x) < \infty$;

(2) $0 < \inf\limits_{w,a,x} \Gamma(w, a, x) \leqslant \sup\limits_{w,a,x} \Gamma(w, a, x) < \infty$;

(3) *Let* $T : L_2(W, A, X) \to L_2(Z, A, X)$ *be the conditional expectation operator given by* $T(g) \equiv \mathsf{E}_\theta[g(W, A, X)|Z, A, X]$

and the adjoint $T' : L_2(Z, A, X) \to L_2(W, A, X)$ *be* $T'(g) \equiv \mathsf{E}_\theta[g(Z, A, X)|W, A, X]$.

Suppose that the inverse mapping $\{T'\Psi^{-1}T\}^{-1}$ *exists and is uniquely defined and that*

$T'\Psi^{-1}\mathsf{E}_\theta[(Y - r(W, A, X))(r(W, A = a, X) - \psi(\theta))|Z, A, X] \in Domain(\{T'\Psi^{-1}T\}^{-1})$

$\mathbb{1}\{A = a\}/f(A|W, X) \in Domain(\{T'\Psi^{-1}T\}^{-1})$.

Then the confounding bridge functions r and q that solve integral equations

(C.1)
$$\mathsf{E}_\theta[Y|X, A = a, Z] = \int r(w, A = a, X)\mathrm{d}F_\theta(w|X, Z),$$

$$\frac{1}{\Pr_\theta(A = a|X, W)} = \int q(z, A = a, X)\mathrm{d}F_\theta(z|X, W, A = a).$$

are uniquely identified. The nonparametric first order influence function of $\psi(\theta) = -\mathsf{E}_\theta[Y(a)]$ is given by

(C.2) $\mathsf{IF}_{1,\psi}(\theta) = \mathbb{1}\{A = a\}q(Z, A = a, X)r(W, A = a, X) - \mathbb{1}\{A = a\}Yq(Z, A = a, X) - r(W, A = a, X) - \psi(\theta)$.

In the main text, all the notations related to b and p are extended to h and q. For the truncated nuisance functions $\widetilde{r}_{k,\theta}$ and $\widetilde{q}_{k,\theta}$, we use k-dimensional basis $\bar{\mathfrak{z}}_k(W, A, X)$ and $\bar{\mathsf{z}}_k(Z, A, X)$, respectively.

Furthermore, the HOIFs of order m of $\widetilde{\psi}_k(\theta) = \mathsf{E}_\theta[\mathcal{H}(\widetilde{r}_{k,\theta}, \widetilde{q}_{k,\theta})]$ are $\mathbb{IF}_{m,k}(\Sigma_k^{-1}) = \mathbb{IF}_1 - \mathbb{IF}_{22 \to mm,k}(\Sigma_k^{-1})$, where $\mathbb{IF}_{\ell\ell \to mm,k}(\Sigma_k^{-1}) := \sum_{j=\ell}^m \mathbb{IF}_{jj,k}(\Sigma_k^{-1})$, $\Sigma_k := \mathsf{E}_\theta[\mathbb{1}\{A = a\}\bar{\mathfrak{z}}_k(W, A = a, X)\bar{\mathsf{z}}_k(Z, A = a, X)^\top]$ which is assumed to be

invertible,

(C.3)
$$\mathbb{IF}_{22,k}(\Sigma_k^{-1}) := \frac{1}{n(n-1)} \sum_{1 \leqslant i_1 \neq i_2 \leqslant n} \mathsf{IF}_{22,k,\bar{i}_2}(\Sigma_k^{-1})$$

and for $j > 2$,

(C.4)
$$\mathbb{IF}_{jj,k}(\Sigma_k^{-1}) := \frac{(n-j)!}{n!} \sum_{1 \leqslant i_1 \neq \dots \neq i_j \leqslant n} \mathsf{IF}_{jj,k,\bar{i}_j}(\Sigma_k^{-1}).$$

Here

(C.5)
$$\mathsf{IF}_{22,k,\bar{i}_2}(\Sigma_k^{-1}) \equiv \mathsf{IF}_{22,\widetilde{\psi}_k,\bar{i}_2}(\theta) = [\mathbb{1}\{A = a\}(Y - \widetilde{r}_{k,\theta}(W, A = a, X))\bar{\mathsf{z}}_k(Z, A = a, X)]_{i_1}^\top \Sigma_k^{-1} [\bar{\mathfrak{z}}_k(W, A = a, X)(\mathbb{1}\{A = a\}\widetilde{q}_{k,\theta}(Z, A = a, X) - 1)]_{i_2},$$

and

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

$$= (-1)^j \left[\mathbb{1}\{A = a\}(Y - \widetilde{r}_{k,\theta}(W, A = a, X))\bar{\mathsf{z}}_k(Z, A = a, X) \right]_{i_1}^\top \left\{ \prod_{s=3}^j \Sigma_k^{-1} \left(\left[\mathbb{1}\{A = a\}\bar{\mathfrak{z}}_k(W, A = a, X)\bar{\mathsf{z}}_k(Z, A = a, X)^\top \right]_{i_s} - \Sigma_k \right) \right\}$$

$$\times \Sigma_k^{-1} [\bar{\mathfrak{z}}_k(W, A = a, X)(\mathbb{1}\{A = a\}\widetilde{q}_{k,\theta}(Z, A = a, X) - 1)]_{i_2}.$$

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 $\bar{i}_j := \{i_1, \dots, i_j\}$.

Note that the first order influence function in the above theorem has appeared in Cui et al. (2020). We state it in the above theorem for the sake of completeness. The forms of HOIFs are new, but are straightforward applications of the derivation given in Appendix A.6 for DR functionals.
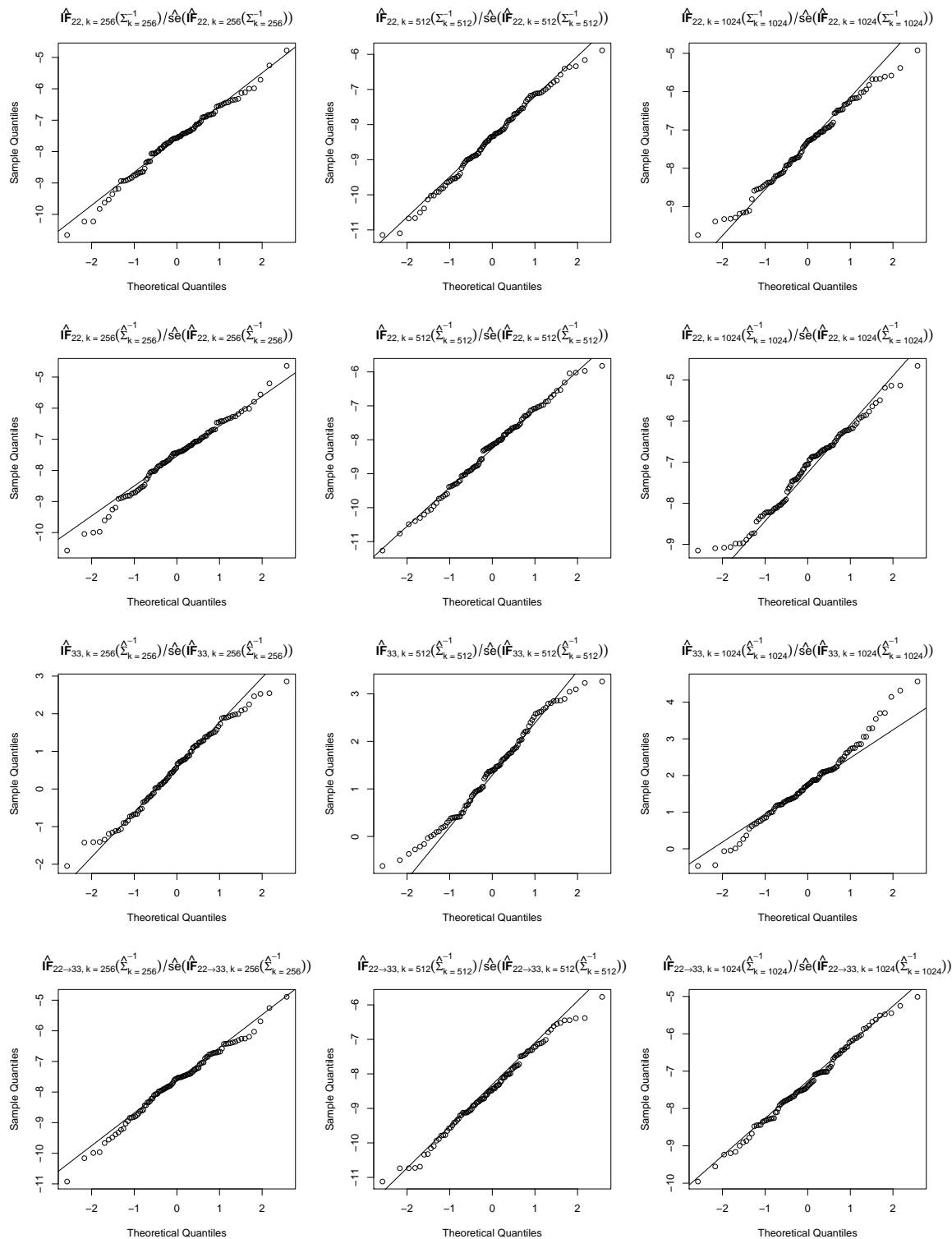
The figures mentioned in Section 5 are collected here.

FIGURE 1. For data generating mechanism in simulation setup I: normal qq-plots for $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})/\widehat{\text{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]$ (the first row), $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})/\widehat{\text{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})]$ (the second row), $\widehat{\mathbb{IF}}_{33,k}(\widehat{\Sigma}_k^{-1})/\widehat{\text{s.e.}}[\widehat{\mathbb{IF}}_{33,k}(\widehat{\Sigma}_k^{-1})]$ (the third row) and $\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})/\widehat{\text{s.e.}}[\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})]$ (the fourth row) for $k = 256$ (left panels), $k = 512$ (middle panels) and $k = 1024$ (right panels).

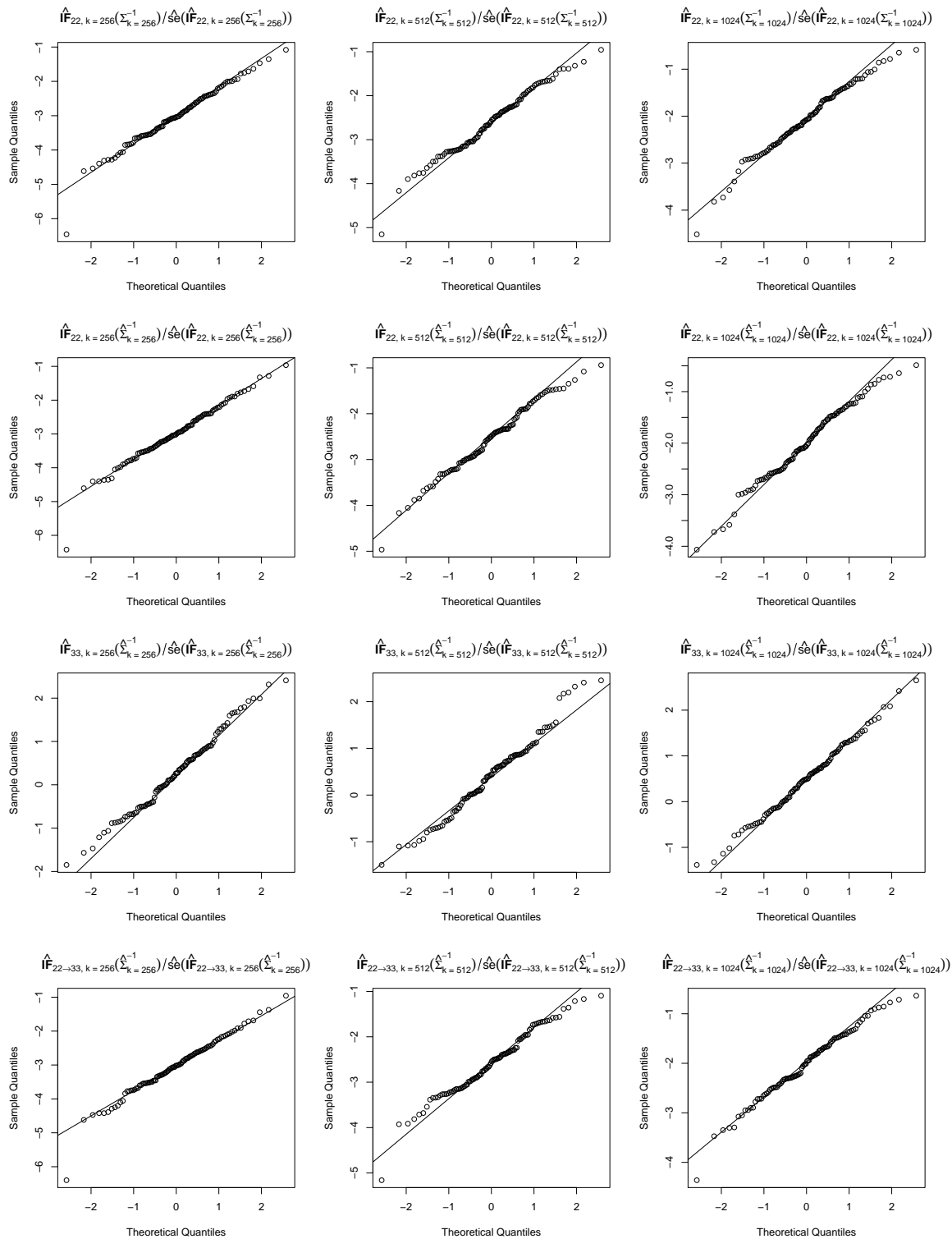FIGURE 2. For data generating mechanism in simulation setup II: normal qq-plots for $\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})/\widehat{\text{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\Sigma_k^{-1})]$ (the first row), $\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})/\widehat{\text{s.e.}}[\widehat{\mathbb{IF}}_{22,k}(\widehat{\Sigma}_k^{-1})]$ (the second row), $\widehat{\mathbb{IF}}_{33,k}(\widehat{\Sigma}_k^{-1})/\widehat{\text{s.e.}}[\widehat{\mathbb{IF}}_{33,k}(\widehat{\Sigma}_k^{-1})]$ (the third row) and $\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})/\widehat{\text{s.e.}}[\widehat{\mathbb{IF}}_{22\to33,k}(\widehat{\Sigma}_k^{-1})]$ (the fourth row) for $k = 256$ (left panels), $k = 512$ (middle panels) and $k = 1024$ (right panels).

Arcones, Miguel A and Evarist Giné (1992). "On the Bootstrap of U and V Statistics". In: *The Annals of Statistics* 20.2, pp. 655–674.

Baker, Rose (2008). "An order-statistics-based method for constructing multivariate distributions with fixed marginals". In: *Journal of Multivariate Analysis* 99.10, pp. 2312–2327.

Belloni, Alexandre et al. (2015). "Some new asymptotic theory for least squares series: Pointwise and uniform results". In: *Journal of Econometrics* 186.2, pp. 345–366.

Bhattacharya, Rabi N and Jayanta K Ghosh (1992). "A class of U-statistics and asymptotic normality of the number of k-clusters". In: *Journal of Multivariate Analysis* 43.2, pp. 300–330.

Chernozhukov, Victor et al. (2020). "Adversarial Estimation of Riesz Representers". In: *arXiv preprint arXiv:2101.00009*.

Cui, Yifan et al. (2020). "Semiparametric proximal causal inference". In: *arXiv preprint arXiv:2011.08411*.

Ghassami, AmirEmad et al. (2022). "Minimax kernel machine learning for a class of doubly robust functionals with application to proximal causal inference". In: *International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 7210–7239.

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

Huskova, Marie and Paul Janssen (1993). "Consistency of the Generalized Bootstrap for Degenerate U-Statistics". In: *The Annals of Statistics* 21.4, pp. 1811–1823.

Kallus, Nathan, Xiaojie Mao, and Masatoshi Uehara (2021). "Causal Inference Under Unmeasured Confounding With Negative Controls: A Minimax Learning Approach". In: *arXiv preprint arXiv:2103.14029*.

Liu, Lin et al. (2017). "Semiparametric efficient empirical higher order influence function estimators". In: *arXiv preprint arXiv:1705.07577*.

Miao, Wang, Xu Shi, and Eric Tchetgen Tchetgen (2018). "A Confounding Bridge Approach for Double Negative Control Inference on Causal Effects". In: *arXiv preprint arXiv:1808.04945*.

Mosimann, James E (1962). "On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions". In: *Biometrika* 49.1-2, pp. 65–82.

Newcomer, Justin T, Nagaraj K Neerchal, and Jorge G Morel (2008). *Computation of higher order moments from two multinomial overdispersion likelihood models*. Tech. rep. Department of Mathematics and Statistics, University of Maryland.

Newey, Whitney K and James L Powell (2003). "Instrumental variable estimation of nonparametric models". In: *Econometrica* 71.5, pp. 1565–1578.

Robins, James et al. (2008). "Higher order influence functions and minimax estimation of nonlinear functionals". In: *Probability and Statistics: Essays in Honor of David A. Freedman*. Institute of Mathematical Statistics, pp. 335–421.

Rotnitzky, Andrea, Ezequiel Smucler, and James M Robins (2021). "Characterization of parameters with a mixed bias property". In: *Biometrika* 108.1, pp. 231–238.

Shpitser, Ilya, Zach Wood-Doughty, and Eric J Tchetgen Tchetgen (2021). "The Proximal ID Algorithm". In: *arXiv preprint arXiv:2108.06818*.

Tchetgen Tchetgen, Eric J et al. (2020). "An Introduction to Proximal Causal Learning". In: *arXiv preprint arXiv:2009.10982*.