cemmap
centre for microdata methods and practice

Semiparametric efficient empirical higher order influence function estimators

Rajarshi Mukherjee
Whitney K. Newey
James M. Robins

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP30/17

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

SEMIPARAMETRIC EFFICIENT EMPIRICAL HIGHER ORDER INFLUENCE FUNCTION ESTIMATORS

By Rajarshi Mukherjee*, Whitney K. Newey†, and James M. Robins‡

Robins et al. (2008, 2016b) applied the theory of higher order influence functions (HOIFs) to derive an estimator of the mean of an outcome Y in a missing data model with Y missing at random conditional on a vector X of continuous covariates; their estimator, in contrast to previous estimators, is semiparametric efficient under minimal conditions. However the Robins et al. (2008, 2016b) estimator depends on a non-parametric estimate of the density of X. In this paper, we introduce a new HOIF estimator that has the same asymptotic properties as their estimator but does not require non-parametric estimation of a multivariate density, which is important because accurate estimation of a high dimensional density is not feasible at the moderate sample sizes often encountered in applications. We also show that our estimator can be generalized to the entire class of functionals considered by Robins et al. (2008) which include the average effect of a treatment on a response Y when a vector X suffices to control confounding and the expected conditional variance of a response Y given a vector X.

1. Introduction. (Robins et al., 2008, 2016b) introduced novel U-statistic based estimators of nonlinear functionals in semi- and non-parametric models. Construction of these estimators was based on the theory of Higher Order Influence Functions (henceforth referred to as HOIFs). HOIFs are U-statistics that represent higher order derivatives of a functional. The authors' used the HOIFs to construct rate minimax estimators of an important class of functionals in models with $n^{-1/2}$ minimax rates and in higher complexity models with slower minimax rates, where the model complexity was defined in terms of Hölder smoothness classes. This class of functionals is of central importance in biostatistics, epidemiology, economics, and other social sciences and is formally defined in Section 3 below. As specific examples, the class includes the mean of a response Y when Y is missing at random, the average effect of a treatment on a response Y when treatment assignment is ignorable given a vector X of baseline covariates, and the expected conditional covariance of two variables given a vector X. Robins et al. (2008) describe other important functionals in the class. Following Robins et al. (2008), we shall refer to functionals as $\sqrt{n}-estimable$ if the minimax rate of estimation is $n^{-1/2}$ and to be $non\text{-}\sqrt{n}-estimable$ if slower.

One may wonder why higher order influence functions are of interest in the $\sqrt{n}$ case. Surprisingly in this case, HOIFs estimators offer a free lunch, at least asymptotically : one may obtain semiparametric efficiency with HOIFs estimators whose variance is dominated by the linear term associated with the usual first order influence function but whose bias is corrected using higher order influence functions. Moreover, for many functionals, no estimator, other than a HOIF estimator, has been constructed that is $\sqrt{n}-consistent$, much less efficient, under the minimal conditions needed for semiparametric efficiency.

*Stein Fellow, Department of Statistics, Stanford University
†Professor, Department of Economics, Massachusetts Institute of Technology
‡Professor, Department of Biostatistics, Harvard University
AMS 2000 subject classifications: Primary 62G05, 62G20, 62F25
Keywords and phrases: Higher Order Influence Functions, Doubly Robust Functionals, Semiparametric Efficiency, Higher Order U-Statistics

The contribution of this paper is a new HOIF estimator for $\sqrt{n}-estimable$ parameters that, unlike previous HOIF estimators, does not require non-parametric estimation of a high dimensional density g. This is important because accurate high dimensional non-parametric density estimation is generally infeasible at the sample sizes often encountered.

The idea behind our new estimator is exceedingly simple. All HOIFs estimators considered heretofore have required an estimate of the inverse of a large covariance matrix whose entries are expectations under an estimate of a density g. For $\sqrt{n}-estimable$ parameters we shall need to consider matrices with up to $n/(\log n)^3$ rows where n is the sample size . In the non-$\sqrt{n}$ case the number of rows is strictly greater than n and less than n^2. Our new HOIFs estimator simply uses an empirical inverse covariance matrix, thereby avoiding estimation of g. Of course this is only possible in the $\sqrt{n}$ case, as the empirical inverse covariance matrix does not exist if the number of rows exceeds n. We refer to the new estimators as empirical HOIF estimators. Our main technical contribution is a proof that our new estimator is minimax and in fact efficient in the semiparametric sense in the $\sqrt{n}$ case. For the sake of concreteness will we first consider the specific example of missing data with the variable of interest missing at random. We then provide general results that apply to all functionals in our class.

The rest of the paper is organized as follows. In Section 2.1 we introduce the missing at random model and the functional we wish to estimate. In Section 2.2 we introduce our new empirical HOIF estimator. In Section 2.3 we analyze the large sample properties of our estimator and compare its behavior to the HOIF estimators of Robins et al. (2008, 2016b). In Section 2.4 we show the empirical HOIF estimator is semiparametric efficient under *minimal conditions* when complexity of the model is defined in terms of Holder smoothness classes. In Section 3 we extend the results of Section 2 to the more general class of doubly robust functionals studied by Robins et al. (2008). Section 4 discusses implications of the results. Finally we collect our proofs and required technical lemmas in Section 5 and 6 respectively.

2. A New Higher order Influence Function Estimator in a Missing Data Model.

2.1. *Observation Scheme.* We observe N i.i.d copies of observed data $W=(AY,A,X)$. Here $A\in\{0,1\}$ is the indicator of the event that a binary response $Y\in\{0,1\}$ is observed and X is a d-dimensional vector of covariates with density $f(x)$ with respect to the Lebesgue measure on a compact set in $\mathbb{R}^d$, which we assume to be $[0,1]^d$ from now on. Define

$$\begin{aligned} B &:= b(X)=\mathbb{E}(Y|A=1,X) \\ \Pi &:= \pi(X)=\mathbb{P}(A=1|X) \end{aligned}$$

where $x\mapsto b(x)$ is the outcome regression function and $x\mapsto\pi(x)$ is the propensity score. We are interested in estimating $\psi=\mathbb{E}\left[\frac{AY}{\pi(X)}\right]=\mathbb{E}[b(X)]=\int b(x)f(x)dx$. Interest in ψ lies in the fact that it is the marginal mean of Y under the missing at random (MAR) assumption that $\mathbb{P}(A=1|X,Y)=\pi(X)$. It will useful to parametrize the model by $\theta=(b,p,g)$ for functions $x\mapsto b(x), x\mapsto p(x), x\mapsto g(x)$ where $x\mapsto p(x)=1/\pi(x), x\mapsto g(x)=\mathbb{E}[A|X=x]f(x)=\pi(x)f(x)=f(x|A=1)\mathbb{P}(A=1)$. Further, it is easy to see that the parameters b,p,g are variation independent. As discussed in Robins et al. (2008, 2016b), the parametrization (b,p,g) is much more natural than (b,p,f), as will be evident from the formulas provided below. We also assume that g is absolutely continuous with respect to the Lebesgue measure μ. In view of this parametrization we write the corresponding probability measure, expectation, and variance operators as $\mathbb{P}_\theta,\mathbb{E}_\theta$, and var_θ respectively. Finally, in terms of this parametrization, we can write the functional $\theta\mapsto\psi(\theta)$ of interest as

$$\chi(\mathbb{P}_\theta) = \psi(\theta) = \int b(x)\, p(x)\, g(x)\, dx.$$

We assume that the law of W belongs to a model

$$\mathcal{M}(\Theta) = \{\mathbb{P}_\theta, \theta \in \Theta\}.$$

where for some $\sigma > 0,\ M > 0$,

$$\Theta \subseteq \{\theta : \inf_x \pi(x) \geq \sigma,\ \inf_x g(x) \geq \sigma,\ \sup_x g(x) \leq M\}. \tag{2.1}$$

We will assume the model $\mathcal{M}(\Theta)$ is locally non-parametric (in the sense that the tangent space at each $\theta \in \Theta$ equals $L_2(\mathbb{P}_\theta)$). Then it is well-known (Robins and Ritov, 1997; Tsiatis, 2007) that the unique first order influence function for ψ at θ is

$$IF_1(\theta) = Ap(X)(Y - b(X)) + b(X) - \psi(\theta),$$

which we can also write as $AP(Y - B) + B - \psi(\theta)$ in our notation.

In Section 2.4, we study a particular Θ defined by membership of the functions b, p, g in certain Hölder smoothness balls and show that the proposed estimator is adaptive and semiparametric efficient in the corresponding model $M(\Theta)$. However, for now, we work with any Θ satisfying (2.1).

We are now ready to define both the estimators of Robins et al. (2008, 2016b) and then the new estimator of this paper, followed by their analyses.

2.2. *The Estimators.* Our estimators will depend on a random variable H_1 that will vary depending on the functional in the doubly robust class of Robins et al. (2008) under investigation in Section 3. $H_1 = h_1(W)$ will either be nonnegative w.p.1 or non-positive w.p.1. In our MAR example, we have

$$H_1 = -A.$$

which is non-positive w.p.1. We shall consider estimators $\widehat{\psi}_{m,k}$ constructed as follows where the indices m and k are defined below.

(i) The sample is randomly split into 2 parts: an estimation sample of size n and a training sample of size $n_{tr} = N - n$ with $n/N \to c^*$ and $n \to \infty$ with $0 < c^* < 1$.

(ii) Estimators $\widehat{g}, \widehat{b}, \widehat{p}$ are constructed from the training sample data.We do not restrict the form of these estimators. Let $\widehat{\theta} = \left(\widehat{b}, \widehat{p}, \widehat{g}\right)$.

(iii) Given a complete sequence of basis functions $z_1(x), z_2(x), \ldots$, for $L_2[0,1]^d$, let $\overline{z}_k(x) = (z_1(x), z_2(x), \ldots, z_k(x))^T$, $Z_k = z_k(X)$, $\overline{Z}_k = (Z_1, Z_2 ..., Z_k)^T$, and define the following covariance matrices

$$\begin{aligned}
\Omega &= \mathbb{E}_\theta\left[|H_1|\overline{Z}_k\overline{Z}_k^T\right] = \int \overline{z}_k(x)\,\overline{z}_k^T(x)\, g(x)\, dx, \\
\widehat{\Omega}_k^{ac} &= \mathbb{E}_{\widehat{\theta}}\left[|H_1|\overline{Z}_k\overline{Z}_k^T\right] = \int \overline{z}_k(x)\,\overline{z}_k^T(x)\, \widehat{g}(x)\, dx, \\
\widehat{\Omega}^{emp} &= n_{tr}^{-1} \sum_{i \in training} \left[|H_1|\overline{Z}_k\overline{Z}_k^T\right]_i.
\end{aligned}$$

(iv) Set

$$\widehat{\psi}_1 = \widehat{\psi} + n^{-1}\sum_i^n \widehat{IF}_{1,i}$$

where $\widehat{\psi}$ and $\widehat{IF}_1$ are $\psi(\theta)$ and $IF_1(\theta)$ with $\widehat{\theta}$ replacing θ. The estimator $\widehat{\psi}_1$ is the usual one-step estimator that adds the estimated first order influence function to the plug-in estimator.

(v) Let $\varepsilon_b = H_1(Y-B)$, $\varepsilon_p = H_1 P - 1$. For $m = 2, ...,$ and any invertible $\widehat{\Omega}$ define

$$\widehat{\psi}_{m,k}(\widehat{\Omega}) \equiv \widehat{\psi} + \sum_{j=2}^m \widehat{\mathbb{IF}}_{j,j,k}(\widehat{\Omega})$$

where $\widehat{\mathbb{IF}}_{j,j,k}$ is the jth order U-statistic

$$\widehat{\mathbb{IF}}_{j,j,k}(\widehat{\Omega}) = \frac{j!}{n!}\sum_{i_1\neq i_2\neq\cdots\neq i_j} \widehat{IF}_{j,j,k,\bar{i}_j}(\widehat{\Omega}),$$

and where all the sums are only over subjects in the estimation sample with distinct coordinate multi-indices $\bar{i}_j := \{i_1, i_2, \ldots, i_j\}$, and for $j \geq 2$

$$\widehat{IF}_{2,2,k,\bar{i}_2}(\widehat{\Omega}) = -(-1)^{I\left(h_1(W_{i_1})\leq 0\right)}\left[\varepsilon_{\widehat{p}}\overline{Z}_k^T\right]_{i_1}\widehat{\Omega}^{-1}\left[\overline{Z}_k\varepsilon_{\widehat{b}}\right]_{i_2}$$

$$\widehat{IF}_{j,j,k,\bar{i}_j}(\widehat{\Omega}) = (-1)^{j-1}(-1)^{I\left(h_1(W_{i_1})\leq 0\right)}\left\{\begin{array}{c}\left[\varepsilon_{\widehat{p}}\overline{Z}_k^T\right]_{i_1}\widehat{\Omega}^{-1}\times \\ \prod_{s=3}^{j}\left[\left\{\left(|H_1|\overline{Z}_k\overline{Z}_k^T\right)_{i_s} - \widehat{\Omega}\right\}\widehat{\Omega}^{-1}\right] \\ \times\left[\overline{Z}_k\varepsilon_{\widehat{b}}\right]_{i_2}\end{array}\right\}, j > 2.$$

Finally we define

$$\hat{\psi}^{ac}_{m,k} := \hat{\psi}_{m,k}(\widehat{\Omega}^{ac}), \quad \hat{\psi}^{emp}_{m,k} := \hat{\psi}_{m,k}(\widehat{\Omega}^{emp}),$$

where, by convention, we define an estimator to be zero if the associated covariance estimator $\widehat{\Omega}^{ac}$ or $\widehat{\Omega}^{emp}$ fails to be invertible. Note that $\widehat{\psi}_1$ is the sample average of $A\widehat{P}(Y-\widehat{B})+\widehat{B}$ and thus does not depend on $\widehat{g}$.

REMARK 1. *In our MAR model, regression estimators $\widehat{b}, \widehat{\pi} = 1/\widehat{p}$ and density estimator $\widehat{f}(x|A=1) = \widehat{g}(x)\left\{(n_{tr})^{-1}\sum_{i=1}^{n_{tr}} A_i\right\}^{-1}$ could, for example, be constructed from training sample data by using multiple machine learning algorithms to construct candidate estimators and then using cross validation to choose the best candidate.*

2.3. *Analysis of the Estimators.* Robins et al. (2008, 2016b) analyzed the estimator $\widehat{\psi}^{ac}_{m,k}$. In this paper, we shall analyze the estimator $\widehat{\psi}^{emp}_{m,k}$, which has the advantage of not requiring an estimate $\widehat{g}$ of g. The following theorem of Robins et al. (2008, 2016b) gives the conditional bias for any $\widehat{\Omega}$ estimated from the training sample.

THEOREM 1. *For any invertible* $\widehat{\Omega}$ *one has conditional on the training sample,* $\mathbb{E}_\theta\left[\widehat{\psi}_{m,k}-\psi\left(\theta\right)\right]=EB_{m,k}\left(\theta\right)+TB_k\left(\theta\right)$

$$\begin{aligned}
EB_{m,k}\left(\theta\right) &= (-1)^{(m-1)+I(h_1(W)\leq 0)}\left\{\begin{array}{c}\mathbb{E}_\theta\left[H_1\left(P-\widehat{P}\right)\overline{Z}_k^T\right]\Omega^{-1}\left[\left\{\Omega-\widehat{\Omega}\right\}\widehat{\Omega}^{-1}\right]^{m-1}\\ \times\mathbb{E}_\theta\left[\overline{Z}_k H_1\left(B-\widehat{B}\right)\right]\end{array}\right\}\\
&\quad TB_k\left(\theta\right)\\
&= (-1)^{I(h_1(W)\leq 0)}\left\{\begin{array}{c}\int dx g\left(x\right)\left(b-\widehat{b}\right)\left(x\right)\left(p-\widehat{p}\right)\left(x\right)\\ -\int\int g\left(x_1\right)g\left(x_2\right)\left(b-\widehat{b}\right)\left(x_1\right)K_{g,k}\left(x_1,x_2\right)\left(p-\widehat{p}\right)\left(x_2\right)dx_2dx_1\end{array}\right\}\\
&= (-1)^{I(h_1(W)\leq 0)}\left[\int dx g\left(x\right)\left(I-\Pi_{g,k}\right)\left[\left(b-\widehat{b}\right)\right]\left(x\right)\left(I-\Pi_{g,k}\right)\left[p-\widehat{p}\right]\left(x\right)\right],
\end{aligned}$$

with

$$K_{g,k}\left(x',x\right)=\overline{z}_k^T\left(x'\right)\Omega^{-1}\overline{z}_k\left(x\right)$$

the orthogonal projection kernel onto $\overline{z}_k\left(x\right)$ *in* $L_2\left(g\right)$, *and*

$$\Pi_{g,k}\left[h\right]\left(x\right)=\int dx' g\left(x'\right)h\left(x'\right)K_{g,k}\left(x,x'\right)$$

the corresponding orthogonal projection of any function $x\mapsto h\left(x\right)$, *and* $I\left[h\right]\left(x\right)=h\left(x\right)$.

To proceed further we require the following definition.

DEFINITION 1. *We say that a choice of basis functions* $\{z_l,\ l\geq 1\}$, *and tuple of functions* $\widetilde{\theta}=(\widetilde{b},\widetilde{p},\widetilde{g})$ *in* $\mathbb{R}^{[0,1]^d}$ *satisfies Condition(B) if the following hold for some* $1<B<\infty$ *and every* $n,k\geq 1$

A.1 $\sup_x \overline{z}_k^T\left(x\right)\overline{z}_k\left(x\right)\leq B\cdot k$.
A.2 $\frac{1}{B}\leq\lambda_{\min}(\Omega)\leq\lambda_{\max}(\Omega)\leq B$.
A.3 $\left\|\frac{d\mathbb{P}_\theta}{d\mathbb{P}_{\widetilde{\theta}}}\right\|_\infty\leq B$.

Now we are ready to further analyze $\hat{\psi}_{m,k}^{ac}$ and $\hat{\psi}_{m,k}^{emp}$. In particular, the following theorem concerning $\widehat{\psi}_{m,k}^{ac}$ can be easily derived from Remark 3.18 following (Robins et al., 2008, Theorem 3.17) and (Robins et al., 2016b, Theorem 8.1).

THEOREM 2. *Assume that* $\{z_l,\ l\geq 1\}$ *and* $\hat{\theta}=(\hat{b},\hat{p},\hat{g})$ *satisfy Condition(B). Then there exists* $c>1$ *such that the following hold conditional on the training sample restricted to the event that* $\widehat{\Omega}^{ac}$ *is invertible.*

1. $TB_k\left(\theta\right)=O\left(\left\|\left(I-\Pi_{g,k}\right)\left[\left(b-\widehat{b}\right)\right]\right\|_2\left\|\left(I-\Pi_{g,k}\right)\left[\left(p-\widehat{p}\right)\right]\right\|_2\right)$,
2. $EB_{m,k}^{ac}\left(\theta\right)=O\left(\left\|b-\widehat{b}\right\|_{m+1}\left\|p-\widehat{p}\right\|_{m+1}\left\|g-\widehat{g}\right\|_{m+1}^{m-1}\right)$,
3. $var_\theta[\widehat{\psi}_{m,k}^{ac}]\leq\sum_{j=1}^m\frac{c^jk^{j-1}}{\binom{n}{j}}$.

An analogous theorem for $\widehat{\psi}_{m,k}^{emp}$ is stated below, which is the main result of this paper.

THEOREM 3. *Assume that* $\{z_l,\ l \geq 1\}$ *and* $\widetilde{\theta} = (\hat{b}, \hat{p}, g)$ *satisfy Condition(B) and that* $\|\widehat{\Omega} - \Omega\|_{op} \leq 1/2B$. *Then there exists* $c > 1$ *such that the following hold conditional on the training sample restricted to the event that* $\widehat{\Omega}^{emp}$ *is invertible.*

1. $TB_k(\theta) = O\left(\left\|(I - \Pi_{g,k})\left[\left(b - \widehat{b}\right)\right]\right\|_2 \|(I - \Pi_{g,k})[(p - \widehat{p})]\|_2\right)$.
2. $EB_{m,k}^{emp}(\theta) = O\left(\left\|b - \widehat{b}\right\|_2 \|p - \widehat{p}\|_2 \|\widehat{\Omega} - \Omega\|_{op}^{m-1}\right)$.
3. $var_\theta[\widehat{\psi}_{m,k}^{emp}] \leq \sum_{l=0}^{m-2} \left(\sum_{j=(l+2)\vee 3}^{m} \binom{j-2}{l}\right)^2 \frac{c^{l+2}k^{l+1}}{\binom{n}{l+2}}$.

A few remarks are in order about the statement and implications of Theorem 3. First, we make a clarification about the Condition(B) holding with $\widetilde{\theta} = (\hat{b}, \hat{p}, g)$. In particular, note that we do not assume that g is known; rather only that $\mathbb{P}_{\widetilde{\theta}}$ with $\widetilde{\theta} = (\hat{b}, \hat{p}, g)$ satisfies $\left\|\frac{d\mathbb{P}_\theta}{d\mathbb{P}_{\widetilde{\theta}}}\right\|_\infty \leq B$. Next we note that the upper bound on the variance of $\hat{\psi}_{m,k}^{emp}$ is typically larger than that of $\hat{\psi}_{m,k}^{ac}$ in Theorem 2. We do not believe this to be a simple artifact of the proof but rather arises from the fact that the empirical measure $\frac{1}{n_{tr}} \sum_{i \in training} \delta_{X_i}$ is not absolutely continuous with respect to the Lebesgue measure. Obtaining the variance bound in Theorem 3 was the main technical challenge of the paper.

We now show that by allowing k and m to grow with n we may be able to obtain semiparametric efficient estimators of ψ. In the context of $\hat{\psi}^{ac}$, the following theorem is closely related to and is proved exactly like (Robins et al., 2016b, Theorem 8.2) and is the main step needed to show semiparametric efficiency.

THEOREM 4. *Assume the following.*

(i) $k(n) = n/(\ln n)^2$ *and* $m(n) = \ln n$ *and define* $\widehat{\psi}_n^{ac} = \widehat{\psi}_{m(n),k(n)}^{ac}$.

(ii) The conditions of Theorem 2 hold, $\left\|\widehat{b} - b\right\|_\infty$ *and* $\|\widehat{p} - p\|_\infty$ *are* $O_{\mathbb{P}_\theta}(1)$, *and there exists some* $\delta > 0$ *such* $\|\hat{g} - g\|_\infty = O_{\mathbb{P}_\theta}(n^{-\delta})$.

(iii) $TB_{k(n)}(\theta) = o_{\mathbb{P}_\theta}\left(n^{-1/2}\right)$.

Then

$$n^{1/2}\left(\widehat{\psi}_n^{ac} - \psi(\theta)\right) = n^{1/2}\sum_{i=1}^{n} IF_{1,i}(\theta) + o_{\mathbb{P}_\theta}(1).$$

An immediate corollary is that under the conditions of Theorem 4 is that $\left(\widehat{\psi}_n^{ac} + \widehat{\psi}_{n_t}^{\overline{ac}}\right)/2$ is semiparametric efficient at θ where $\widehat{\psi}_{n_t}^{\overline{ac}}$ is $\widehat{\psi}_n^{ac}$ but with the roles of the training and estimation sample reversed. This follows from the fact that any asymptotically linear estimator with the efficient influence function as its influence function is regular and semiparametric efficient (Van der Vaart, 2000). Below we provide the analogous theorem for $\widehat{\psi}_n^{emp}$.

THEOREM 5. *Assume the following.*

(i) $k(n) = n/(\ln n)^3$ *and* $m(n) = \sqrt{\ln n}$ *and define* $\widehat{\psi}_n^{emp} = \widehat{\psi}_{m(n),k(n)}^{emp}$.

(ii) The conditions of Theorem 3 hold, $\left\|\widehat{b} - b\right\|_\infty$ *and* $\|\widehat{p} - p\|_\infty$ *are* $O_{\mathbb{P}_\theta}(1)$.

(iii) $TB_{k(n)}(\theta) = o_{\mathbb{P}_\theta}\left(n^{-1/2}\right)$.

Then

$$n^{1/2}\left(\widehat{\psi}_n^{emp} - \psi(\theta)\right) = n^{1/2}\sum_{i=1}^{n} IF_{1,i}(\theta) + o_{\mathbb{P}_\theta}(1).$$

Hence, once again, $\left(\widehat{\psi}_n^{emp} + \widehat{\psi}_{n_t}^{\overline{emp}}\right)/2$ is semiparametric efficient at θ under the conditions of Theorem 5. Finally, it is immediate from a comparison of Theorem 4 and Theorem 5, that an advantage of the latter is that it requires effectively no assumptions on the function g.

2.4. *Adaptive Efficient Estimation.* In this section we show that we can use our empirical HOIF estimators to obtain adaptive semiparametric efficient estimators when Θ assumes the functions b, p live in Hölder balls. Following Robins et al. (2008, 2016b) we define the complexity of the model $\mathcal{M}(\Theta)$ in terms of Hölder smoothness classes defined as follows.

DEFINITION 2. *A function $x \mapsto h(x)$ with domain a compact subset of D of R^d is said to belong to a Hölder ball $H(\beta, C)$, with Hölder exponent $\beta > 0$ and radius $C > 0$, if and only if h is uniformly bounded by C, all partial derivatives of h up to order $\lfloor\beta\rfloor$ exist and are bounded, and all partial derivatives $\nabla^{\lfloor\beta\rfloor}$ of order $\lfloor\beta\rfloor$ satisfy*

$$\sup_{x, x+\delta x \in D} \left|\nabla^{\lfloor\beta\rfloor} h(x+\delta x) - \nabla^{\lfloor\beta\rfloor} h(x)\right| \leq C||\delta x||^{\beta - \lfloor\beta\rfloor}.$$

To construct adaptive semiparametric efficient estimators over Hölder balls we use specific bases that satisfy (A.1) and (A.2) of Condition (B) that additionally give optimal rates of approximation for Hölder classes. In particular, we shall assume our basis $\{z_l(x), l = 1, ..,\}$ has optimal approximation properties in $L_2(\mu)$ for Holder balls $H(\beta, C)$ i.e.,

$$sup_{h \in H(\beta,C)} inf_{\varsigma_l} \int_{[0,1]^d} \left(h(x) - \sum_{l=1}^{k} \varsigma_l z_l(x)\right)^2 dx = O\left(k^{-2\beta/d}\right). \tag{2.2}$$

where given any $\{z_l,\ l \geq 1\}$ satisfying (2.2) the O-notation only depends on the Hölder radius C. The basis of $d - fold$ tensor products of B-splines of order s satisfies (2.2) for all $0 < \beta < s + 1$ (Belloni et al., 2015; Newey, 1997). The basis consisting of $d - fold$ tensor products of a univariate Daubechies compact wavelet basis with mother wavelet $\varphi_w(u)$ satisfying

$$\int_{R^1} u^m \varphi_w(u)\, du = 0, m = 0, 1, ..., M$$

also satisfies (2.2) for $\beta < M + 1$ (Härdle et al., 1998). In addition both of these bases satisfy (A.1) and (A.2) of Condition (B) for some large but fixed $1 < B < \infty$ (Belloni et al., 2015; Härdle et al., 1998; Newey, 1997; Robins et al., 2016a).

THEOREM 6. *Assume the following:*

(i) The conditions of Theorem 5 hold and $\{z_l, l \geq 1\}$ satisfy (2.2).
(ii) b, p $\widehat{b}$,$\widehat{p}$ lie in $H(\beta_b, C_b)$ and $H(\beta_p, C_p)$ with $C_p > \frac{1}{\sigma}$.
(iii) $\beta = (\beta_b + \beta_p)/2$ satisfies $\frac{d}{4} < \beta < \beta_{\max}$ for some known $\beta_{\max}$.

Then the estimator $\widehat{\psi}_{m(n),k(n)}^{emp}$ defined in Theorem 5 satisfies

$$TB_k(\theta) = O_{\mathbb{P}_\theta}\left(k^{-\frac{2\beta}{d}}\right) = o_{\mathbb{P}_\theta}\left(n^{-1/2}\right).$$

As an immediate consequence of Theorem 6 we have that $\left(\widehat{\psi}_n^{emp} + \widehat{\psi}_{n_t}^{\overline{emp}}\right)/2$ is semiparametric efficient at any $\mathbb{P}_\theta$ that satisfies conditions of the lemma. Moreover, this result is adaptive over

any $\beta \in (\frac{d}{4}, \beta_{\max})$. Interestingly, the knowledge of an upper bound $\beta_{\max}$ only becomes crucial in constructing a sequence of basis functions $\{z_l,\ l \geq 1\}$ satisfying (2.2) and is not required anywhere else in the analysis. An analogous Theorem for $\hat{\psi}^{ac}_{m,k}$ was proved in (Robins et al., 2016b, Theorem 8.2) with additional conditions on g and $\hat{g}$.

REMARK 2. *When b, p satisfy (ii) in Theorem 6, the following estimators $\hat{b}, \hat{p}$ will do so as well (van der Vaart, Dudoit and Laan, 2006) when the basis $\{z_l, l \geq 1\}$ are compactly supported Daubechies wavelets of sufficient regularity (at least $2\beta_{\max}$): $\widehat{b}(x) = \sum_{l=1}^{k_b} \widehat{\eta}_l z_l(x)$ and $\widehat{p}(x) = 1/\widehat{\pi}(x)$ with $\widehat{\pi}(x) = \sum_{l=1}^{k_\pi} \widehat{\alpha}_l z_l(x)$ with parameters estimated by least squares and k_b and k_π chosen by cross validation, all in the training sample. Note, however, the choices $\widehat{b}(x) \equiv 0$ and $1/\widehat{p}(x) \equiv c$ for $1/c \in (0,1)$, satisfy the conditions of Theorem 6 with $\left\|\widehat{b} - b\right\|_\infty$ and $\|\widehat{p} - p\|_\infty$ being $O_{\mathbb{P}_\theta}(1)$. Thus we obtain the surprising conclusion that our estimators $\hat{b}, \hat{p}$ do not need to be consistent for b and p to obtain a semiparametric efficient estimator $\left(\widehat{\psi}^{emp}_n + \widehat{\psi}^{\overline{emp}}_{n_t}\right)/2$ of $\psi(\theta)$ if $\beta/d > 1/4$! In fact we can even ignore the range of $\widehat{p}$ and choose $\widehat{b} = \widehat{p} \equiv 0$ and still preserve semiparametric efficiency. The explanation of this fact is that when we choose $\widehat{b} = \widehat{p} \equiv 0$, then, although $\widehat{\psi}, \widehat{IF}_1$, and $\widehat{\psi}_1$ are all identically zero, nonetheless $\sum_{j=2}^{m(n)} \widehat{\mathbb{IF}}_{j,j,k(n)}(\widehat{\Omega}^{emp})$ is an estimate of $\left[\int dx g(x) \Pi_{g,k}[b](x) \Pi_{g,k}[p](x)\right]$ with bias $\left\|\widehat{\Omega} - \Omega\right\|_{op}^{m(n)-1} = O_{\mathbb{P}_\theta}\left[\left\{\{k(n)\log(k(n))/n\}^{1/2}\right\}^{m(n)-1}\right] = o_{\mathbb{P}_\theta}\left(n^{-1/2}\right)$ for $\widehat{\Omega}$ constructed using Daubechies compactly supported wavelet bases of suitable regularity or B-Spline bases of suitable order.*

REMARK 3. *Suppose model $\mathcal{M}(\Theta)$ restricts b and p to lie in pre-specified Hölder balls $H(\beta_b, C_b)$, $H(\beta_p, C_p)$. Robins et al (2010) show that the minimax rate for estimating ψ when g is a known function is $n^{-1/2} + n^{-\frac{4\beta/d}{4\beta/d+1}}$. Hence the minimax rate is slower than $n^{-1/2}$ whether g is known or unknown in the model $\mathcal{M}(\Theta)$ when $\beta/d < 1/4$. However, even in such a model there exist parameters, $\theta^* = (b^*, p^*, g^*) \in \Theta$ in which b^* and p^* happen to lie in smaller Holder balls $H(\beta^*_b, C^*_b)$, $H(\beta^*_p, C^*_p)$ with $\left(\beta^*_b + \beta^*_p\right)/2d > 1/4$. Thus $\left(\widehat{\psi}^{ac}_n + \widehat{\psi}^{\overline{ac}}_{n_t}\right)/2$ and $\left(\widehat{\psi}^{emp}_n + \widehat{\psi}^{\overline{emp}}_{n_t}\right)/2$ will be semiparametric efficient at θ^* under the assumptions in Theorem 4 and 5, even though both will converge to $\psi(\theta)$ at a rate slower than $n^{-1/2}$ at nearly all $\theta \in \Theta$.*

REMARK 4. *Note even when b and p lie in Holder balls $H(\beta_b, C_b)$ and $H(\beta_p, C_p)$ with $\beta = \left(\beta_b + \beta_p\right)/2 > d/4$, we still need for $\widehat{b}$,$\widehat{p}$ to lie in these Holder balls with probability approaching one to insure, by Lemma 1, that $TB_k(\theta) = o_p\left(n^{-1/2}\right)$. This may place restrictions on the machine learning algorithms we can use to estimate b and p. As an example suppose (i) we use multiple machine learning algorithms to construct candidate estimators and then use cross validation to choose the best candidate and (ii) the aforementioned series estimators $\widehat{b}(x) = \sum_{l=1}^{k_b} \widehat{\eta}_l z_l(x)$ and $\widehat{p}(x) = 1/\widehat{\pi}(x)$ with $\widehat{\pi}(x) = \sum_{l=1}^{k_\pi} \widehat{\alpha}_l z_l(x)$ are included among the candidates. If the only candidates were these series estimators, we know that $TB_k(\theta) = o_p\left(n^{-1/2}\right)$ for $k = n/(\log n)^3$ and we would be efficient. Nonetheless it may be the case at the particular law $\theta^* = (b^*, p^*, g^*)$ that generated the data, another pair of candidates $\widehat{b}$ and $\widehat{p}$ are chosen with high probability over these series estimators because for these laws, $\widehat{b}$ and $\widehat{p}$ converge to b and p at faster rates than the series estimators. However, faster rates of convergence does not imply that the associated truncation bias $TB_k(\theta) = \int dx g(x)(I - \Pi_{g,k})\left[\left(b - \widehat{b}\right)\right](x)(I - \Pi_{g,k})[p - \widehat{p}](x)$ is less than the truncation bias of the series estimator and thus no guarantee it is $o_p\left(n^{-1/2}\right)$. It is an interesting open question to identify the subset of machine learning algorithm that would give such a guarantee.*

3. A Class of Doubly Robust Functionals. In this section we extend our results to incorporate a general class of double robust functionals studied in Robins et al. (2008). We consider N i.i.d observations $W=(X,V)$ from a law $\mathbb{P}_\theta$ with $\theta\in\Theta$ and wish to make inference on a functional $\chi(\mathbb{P}_\theta)=\psi(\theta)$. We make the following 4 assumptions :

Ai) For all $\theta\in\Theta$, the distribution of X is supported on a compact set in $\mathbb{R}^d$ which we take to be $[0,1]^d$ and has a density $f(x)$ with respect to the Lebesgue measure.

Aii) The parameter θ contains components $b=b(\cdot)$ and $p=p(\cdot)$, $b:[0,1]^d\to\mathbb{R}$ and $p:[0,1]^d\to\mathbb{R}$ such that the functional ψ of interest has a first order influence function $\mathbb{IF}_{1,\psi}(\theta)=N^{-1}\sum_i IF_{1,\psi}(\theta)$, where

$$IF_{1,\psi}(\theta)=H(b,p)-\psi(\theta), \tag{3.1}$$

$$\text{with } H(b,p)\equiv h(W,b(X),p(X))$$

$$\equiv b(X)p(X)h_1(W)+b(X)h_2(W)+p(X)h_3(W)+h_4(W) \tag{3.2}$$

$$\equiv BPH_1+BH_2+PH_3+H_4,$$

and the known functions $h_1(\cdot),h_2(\cdot),h_3(\cdot),h_4(\cdot)$ do not depend on θ. Furthermore $h_1(\cdot)$ is either nowhere negative or nowhere positive on the support of X.

Aiii) $\theta=(b,p,g,c)$ has the product parameter space. $\Theta=\Theta_b\times\Theta_p\times\Theta_g\times\Theta_c$ with $g(x)\equiv E[\{|H_1|\}|X=x]f(x)$ bounded away from zero and infinity and absolutely continuous wrt to Lesbegue measure on the support of X

Aiv) The model $\mathcal{M}(\Theta)$ for $\mathbb{P}_\theta$ satisifes 2.1 and is locally nonparametric in the sense that the tangent space at each $\mathbb{P}_\theta\in\mathcal{M}(\Theta)$ is all of $\mathbf{L}_2(\mathbb{P}_\theta)$.

Our missing data example is the special case with $H_1=-A,H_2=1,H_3=AY,H_4=0$, $p(X)=1/pr(A=1|X),b(X)=E[Y|A=1,X],g(X)=E[A|X]f(X)$

Robins et al. (2008) prove the $H(b,p)$ is doubly robust for $\psi(\theta)$in the sense that

$$E_\theta[H(b^*,p)]=E_\theta[H(b,p^*)]=E_\theta[H(b,p)]=\psi(\theta)$$

for any $\theta\in\Theta$ and functions $b^*(x)$ and $p^*(x)$. Specifically they prove the following.

Theorem : Double-Robustness: Assume Ai)-Aiv) hold. Then

$$\begin{aligned}
\psi(\theta) &\equiv = E_\theta[H_4]-E[BPH_1]\\
&= E_\theta[H_4]-(-1)^{[I(h_1(W)\le 0)]}\int b(x)p(x)g(x)\,dx\\
E_\theta[\{H_1B+H_3\}|X] &= E_\theta[\{H_1P+H_2\}|X]=0\ w.p.1\\
E_\theta[H(b^*,p^*)]-E_\theta[H(b,p)] &= (-1)^{[I(h_1(W)\le 0)]}\left\{\int[b-b^*](x)[p-p^*](x)g(x)\,dx\right\}
\end{aligned}$$

The development in (Robins et al., 2008, Theorem 3.2 and Lemma 3.3) show that results we have obtained only require that Ai)-Aiv) are true. Thus we have the following.

THEOREM 7. *Assume Ai)-Aiv) and redefine* $\varepsilon_b=\{BH_1+H_3\}$, $\varepsilon_p=\{H_1P+H_2\}$, $g(x)=E[|H_1||X=x]f(x)$. *Then the conclusions of Theorem 1-Theorem 6 continue to hold under same conditions on the redefined* $\theta=(b,p,g)$.

4. Discussions. We have shown that for $\sqrt{n}$-estimable parameters the asymptotic properties of our new empirical HOIF estimators are identical to those of the HOIF estimators of Robins et. al (2008,2016), yet eliminate the need to construct multivariate density estimates. In particular the new estimators are semiparametric efficient under minimal conditions.

5. Proofs.

PROOF OF THEOREM 3. We divide our proof into bias and variance computations respectively. Throughout the proof $\widehat{\Omega}$ stands for $\widehat{\Omega}^{emp}$. Throughout we assume $I(h_1(W) \leq 0) = 1$ almost surely. The case $I(h_1(W) \geq 0$ requires obvious sign changes in various place.

Bias Bound : By analysis similar to Robins et al. (2008),

$$EB^{emp}_{m,k}(\theta) = (-1)^m \mathbb{E}_\theta[H_1(P-\hat{P})\bar{Z}_k^T]\Omega^{-1}\left[\left\{\Omega-\hat{\Omega}\right\}\hat{\Omega}^{-1}\right]^{m-1}\mathbb{E}_\theta[\bar{Z}_k H_1(B-\hat{B})].$$

We next show that under the assumptions of Theorem 3

$$\left|EB^{emp}_{m,k}(\theta)\right| = O\left(\|\widehat{\Omega}-\Omega\|_{op}^{m-1}\left\{\mathbb{E}_\theta[(B-\hat{B})^2]\mathbb{E}_\theta[(P-\hat{P})^2]\right\}^{1/2}\right).$$

For any $m > 1$ we see that

$$\|A^m\|_{op} = \left(\max_j |\Lambda_{jj}|\right)^m = \left(\|A\|_{op}\right)^m.$$

Let $\hat{1}$ denote the indicator function for the event that $\lambda_{\max}(\hat{\Omega}^{-1}) \leq C^{-1}$.
By Cauchy-Schwartz Inequality,

$$\left|EB^{emp}_{m,k}\right| \leq \left\|\mathbb{E}_\theta[H_1(P-\hat{P})\bar{Z}_k^T]\Omega^{-1/2}\right\|\left\|\Omega^{-1/2}\left[\left\{\Omega-\hat{\Omega}\right\}\hat{\Omega}^{-1}\right]^{m-1}\mathbb{E}_\theta[\bar{Z}_k H_1(B-\hat{B})]\right\|.$$

Note that $\left\|\mathbb{E}_\theta[H_1(P-\hat{P})\bar{Z}_k^T]\Omega^{-1/2}\right\|^2$ is the second moment of the linear projection of $-(P-\hat{P})$ on $\bar{Z}_k$ under g, so that

$$\left\|\mathbb{E}_\theta[H_1(P-\hat{P})\bar{Z}_k^T]\Omega^{-1/2}\right\| \leq \left\{\mathbb{E}_\theta[(P-\hat{P})^2]\right\}^{1/2}.$$

Also, note that for $H = \Omega^{-1/2}\left[\left\{\Omega-\hat{\Omega}\right\}\hat{\Omega}^{-1}\right]^{m-1}$, in the positive semi-definite sense

$$\begin{aligned}
\hat{1}H^TH &= \hat{1}\left[\hat{\Omega}^{-1}\left\{\Omega-\hat{\Omega}\right\}\right]^{m-1}\Omega^{-1}\left[\left\{\Omega-\hat{\Omega}\right\}\hat{\Omega}^{-1}\right]^{m-1} \\
&\leq \hat{1}C^{-1}\left[\hat{\Omega}^{-1}\left\{\Omega-\hat{\Omega}\right\}\right]^{m-1}\left[\left\{\Omega-\hat{\Omega}\right\}\hat{\Omega}^{-1}\right]^{m-1} \\
&= \hat{1}C^{-1}\left[\hat{\Omega}^{-1}\left\{\Omega-\hat{\Omega}\right\}\right]^{m-2}\hat{\Omega}^{-1}\left\{\Omega-\hat{\Omega}\right\}^2\hat{\Omega}^{-1}\left[\left\{\Omega-\hat{\Omega}\right\}\hat{\Omega}^{-1}\right]^{m-2} \\
&\leq \left\|\Omega-\hat{\Omega}\right\|_{op}^2\hat{1}C^{-1}\left[\hat{\Omega}^{-1}\left\{\Omega-\hat{\Omega}\right\}\right]^{m-2}\hat{\Omega}^{-2}\left[\left\{\Omega-\hat{\Omega}\right\}\hat{\Omega}^{-1}\right]^{m-2} \\
&\leq \left\|\Omega-\hat{\Omega}\right\|_{op}^2\hat{1}C^{-3}\left[\hat{\Omega}^{-1}\left\{\Omega-\hat{\Omega}\right\}\right]^{m-2}\left[\left\{\Omega-\hat{\Omega}\right\}\hat{\Omega}^{-1}\right]^{m-2}.
\end{aligned}$$

Repeating this argument (i.e. by induction) we have

$$\hat{1}H^TH \leq \hat{1}\left\|\Omega-\hat{\Omega}\right\|_{op}^{2(m-1)}C^{-2(m-1)-1}I.$$

Next, since $I \leq \Omega^{-1}C$ in the p.s.d. sense we have

$$\hat{1}H^TH \leq \hat{1}\left\|\Omega-\hat{\Omega}\right\|_{op}^{2(m-1)}C^{-2(m-1)}\Omega^{-1}.$$

It then follows that

$$\begin{aligned}
&\hat{1}\left\|\Omega^{-1/2}\left[\left\{\Omega-\hat{\Omega}\right\}\hat{\Omega}^{-1}\right]^{m-1}\mathbb{E}_\theta[\bar{Z}_k H_1(B-\hat{B})]\right\|^2 \\
&=\hat{1}\left\|H_1\mathbb{E}_\theta[\bar{Z}_k H_1(B-\hat{B})]\right\|^2=\hat{1}\mathbb{E}_\theta[H_1(B-\hat{B})\bar{Z}_k^T]H^T H\mathbb{E}_\theta[\bar{Z}_k H_1(B-\hat{B})] \\
&\leq \hat{1}\left\|\Omega-\hat{\Omega}\right\|_{op}^{2(m-1)} C^{-2(m-1)}\mathbb{E}_\theta[H_1(B-\hat{B})\bar{Z}_k^T]\Omega^{-1}\mathbb{E}_\theta[\bar{Z}_k H_1(B-\hat{B})]. \\
&\leq \hat{1}\left\|\Omega-\hat{\Omega}\right\|_{op}^{2(m-1)} C^{-2(m-1)}\mathbb{E}_\theta[(B-\hat{B})^2],
\end{aligned}$$

where the last inequality follows by $\mathbb{E}_\theta[H_1(B-\hat{B})\bar{Z}_k^T]\Omega^{-1}\mathbb{E}_\theta[\bar{Z}_k H_1(B-\hat{B})]$ being the expected square of the projection of $B-\hat{B}$ on $\bar{Z}_k$ under g. Therefore we have

$$\hat{1}\left|EB_{m,k}^{emp}\right|\leq \hat{1}\left\|\Omega-\hat{\Omega}\right\|_{op}^{m-1} C^{-(m-1)}\left\{\mathbb{E}_\theta[(B-\hat{B})^2]\mathbb{E}_\theta[(P-\hat{P})^2]\right\}^{1/2}$$

This completes the bound for the bias.

Variance Bound : In this section we put $\varepsilon_{\hat{b}}=H_1(Y-\hat{b}(X))$ and $\varepsilon_{\hat{p}}=|H_1|\hat{p}(X)-1$.

To control the variance of $\hat{\psi}_{m,k}$ we begin by analyzing the variance of $\mathbb{U}_n(\widehat{IF}_{22,\bar{i}_2})$. Letting $\hat{\theta}=(\hat{b},\hat{p},\hat{g})$ for any $\hat{g}$ that makes $\mathbb{P}_{\hat{\theta}}$ absolutely continuous with respect to $\mathbb{P}_\theta$ we have the following inequality by Lemma 9

$$\mathbb{E}_\theta\left(\mathbb{U}_n(\widehat{IF}_{2,2,k,\bar{i}_2})\right)^2\leq 4\left(1+\left\|\frac{d\mathbb{P}_\theta}{d\mathbb{P}_{\hat{\theta}}}\right\|_\infty\right)^4\mathbb{E}_{\hat{\theta}}\left(\mathbb{U}_n(\widehat{IF}_{22,\bar{i}_2})\right)^2. \tag{5.1}$$

Now note that for any choice of $\hat{g}$, $\mathbb{U}_n(\widehat{IF}_{22,\bar{i}_2})$ is a second order degenerate U-statistics under $\mathbb{P}_{\hat{\theta}}$. Therefore by Lemma 8 for any $1\leq i_1\neq i_2$ and an universal constant $C>0$

$$\mathbb{E}_{\hat{\theta}}\left(\mathbb{U}_n(\widehat{IF}_{2,2,k,\bar{i}_2})\right)^2\leq\frac{C}{n^2}\mathbb{E}_{\hat{\theta}}\left(\left(\varepsilon_{\hat{p}}\overline{Z}_k^T\right)_{i_1}\widehat{\Omega}^{-1}\left(\overline{Z}_k\varepsilon_{\hat{b}}\right)_{i_1}\right)^2. \tag{5.2}$$

Now

$$\begin{aligned}
&\mathbb{E}_{\hat{\theta}}\left(\left(\varepsilon_{\hat{p}}\overline{Z}_k^T\right)_{i_1}\widehat{\Omega}^{-1}\left(\overline{Z}_k\varepsilon_{\hat{b}}\right)_{i_1}\right)^2 \\
&\leq\|\varepsilon_{\hat{p}}\varepsilon_{\hat{b}}\|_\infty^2\mathbb{E}_{\hat{\theta}}\left(\overline{Z}_k^T(X_{i_1})\widehat{\Omega}^{-1}\overline{Z}_k(X_{i_2})\overline{Z}_k^T(X_{i_2})\widehat{\Omega}^{-1}\overline{Z}_k(X_{i_1})\right) \\
&\leq C\|\varepsilon_{\hat{p}}\varepsilon_{\hat{b}}\|_\infty^2 k\frac{\mathbb{E}_\theta(\lambda_{\max}(\overline{Z}_k\overline{Z}_k^T))}{\lambda_{\min}(\widehat{\Omega})}.
\end{aligned}\tag{5.3}$$

Above the last inequality follows by Lemma 11.

Now note that for any $\{z_l,\ l\geq 1\}$ satisfying Condition(B) and $\|\widehat{\Omega}-\Omega\|_{op}\leq 1/2B$, $\lambda_{\min}(\widehat{\Omega})=\frac{1}{\lambda_{\max}(\widehat{\Omega}^{-1})}\geq\frac{1}{\lambda_{\max}(\widehat{\Omega})}$. Using this fact along with (5.1), (5.2), and (5.3), one has that there exists a constant c depending on the choice of basis functions and $\|\frac{d\mathbb{P}_\theta}{d\mathbb{P}_{\hat{\theta}}}\|_\infty$ such that

$$\mathbb{E}_\theta\left(\mathbb{U}_n(\widehat{IF}_{2,2,k,\bar{i}_2})\right)^2\leq c^2\frac{k}{n^2}. \tag{5.4}$$

For a general $j \geq 3$ note that

$$\begin{aligned}
&\widehat{IF}_{j,j,k} \\
&= \mathbb{U}_n \left((-1)^j \left\{ \begin{array}{c} \left[\varepsilon_{\widehat{p}} \overline{Z}_k^T\right]_{i_1} \widehat{\Omega}^{-1} \times \\ \prod_{s=3}^{j} \left[\left\{ \left(|H_1| \overline{Z}_k \overline{Z}_k^T \right)_{i_s} - \widehat{\Omega} \right\} \widehat{\Omega}^{-1} \right] \\ \times \left[\overline{Z}_k \varepsilon_{\widehat{b}}\right]_{i_2} \end{array} \right\} \right) \\
&= \sum_{l=0}^{j-2} \sum_{\{t_1,\ldots,t_l\} \subseteq \{3,\ldots,j\}} \mathbb{U}_n \left((-1)^j \left\{ \begin{array}{c} \left[\varepsilon_{\widehat{p}} \overline{Z}_k^T\right]_{i_1} \widehat{\Omega}^{-1} \times \\ \prod_{s=3}^{j} \left[\kappa_s^{(t_1,\ldots,t_l)} (H_1 \overline{Z}_k \overline{Z}_k^T)_{i_s} \widehat{\Omega}^{-1} \right] \\ \times \left[\overline{Z}_k \varepsilon_{\widehat{b}}\right]_{i_2} \end{array} \right\} \right),
\end{aligned} \tag{5.5}$$

where

$$\kappa_s^{(t_1,\ldots,t_l)} (H_1 \overline{Z}_k \overline{Z}_k^T) = \begin{cases} |H_1| \overline{Z}_k \overline{Z}_k^T - \Omega & \text{if } s \in \{t_1, \ldots, t_l\} \\ \Omega - \widehat{\Omega} & \text{o.w.} \end{cases}$$

Fix $\{t_1, \ldots, t_l\} \subseteq \{3, \ldots, j\}$. Then letting $\hat{\theta} = (\hat{b}, \hat{p}, \hat{g})$ for any $\hat{g}$ that makes $\mathbb{P}_{\hat{\theta}}$ absolutely continuous with respect to $\mathbb{P}_\theta$ we have the following inequality by Lemma 9

$$\begin{aligned}
&\mathbb{E}_\theta \left(\mathbb{U}_n \left((-1)^j \left\{ \begin{array}{c} \left[\varepsilon_{\widehat{p}} \overline{Z}_k^T\right]_{i_1} \widehat{\Omega}^{-1} \times \\ \prod_{s=3}^{j} \left[\kappa_s^{(t_1,\ldots,t_l)} (H_1 \overline{Z}_k \overline{Z}_k^T)_{i_s} \widehat{\Omega}^{-1} \right] \\ \times \left[\overline{Z}_k \varepsilon_{\widehat{b}}\right]_{i_2} \end{array} \right\} \right) \right)^2 \\
&\leq 2(l+2) \left(1 + \left\| \frac{d\mathbb{P}_\theta}{d\mathbb{P}_{\hat{\theta}}} \right\|_\infty \right)^{2(l+2)} \\
&\times \mathbb{E}_{\hat{\theta}} \left(\mathbb{U}_n \left((-1)^j \left\{ \begin{array}{c} \left[\varepsilon_{\widehat{p}} \overline{Z}_k^T\right]_{i_1} \widehat{\Omega}^{-1} \times \\ \prod_{s=3}^{j} \left[\kappa_s^{(t_1,\ldots,t_l)} (H_1 \overline{Z}_k \overline{Z}_k^T)_{i_s} \widehat{\Omega}^{-1} \right] \\ \times \left[\overline{Z}_k \varepsilon_{\widehat{b}}\right]_{i_2} \end{array} \right\} \right) \right)^2.
\end{aligned} \tag{5.6}$$

Reducing the computation to a degenerate U-statistics under $\hat{\theta}$ can now be achieved by taking $\hat{g} = g$. This in turn allows us to invoke Lemma 8 to conclude that

$$\mathbb{E}_{\hat{\theta}} \left(\mathbb{U}_n \left((-1)^j \left\{ \begin{array}{c} \left[\varepsilon_{\widehat{p}} \overline{Z}_k^T\right]_{i_1} \widehat{\Omega}^{-1} \times \\ \prod_{s=3}^{j} \left[\kappa_s^{(t_1,\ldots,t_l)} (H_1 \overline{Z}_k \overline{Z}_k^T)_{i_s} \widehat{\Omega}^{-1} \right] \\ \times \left[\overline{Z}_k \varepsilon_{\widehat{b}}\right]_{i_2} \end{array} \right\} \right) \right)^2$$

$$
\begin{aligned}
&\leq \frac{1}{\binom{n}{l+2}}\mathbb{E}_{\hat{\theta}}\left(\left\{\begin{array}{c}\left[\varepsilon_{\widehat{p}}\overline{Z}_k^T\right]_{i_1}\widehat{\Omega}^{-1}\times\\ \prod_{s=3}^{j}\left[\kappa_s^{(t_1,\ldots,t_l)}(H_1\overline{Z}_k\overline{Z}_k^T)_{i_s}\widehat{\Omega}^{-1}\right]\\ \times\left[\overline{Z}_k\varepsilon_{\widehat{b}}\right]_{i_2}\end{array}\right\}\right)^2\\
&= \frac{1}{\binom{n}{l+2}}\mathbb{E}_{\hat{\theta}}\left(\left\{\begin{array}{c}\left[\varepsilon_{\widehat{p}}\overline{Z}_k^T\right]_{1}\widehat{\Omega}^{-1}\times\\ \prod_{s=3}^{j}\left[\kappa_s^{(t_1,\ldots,t_l)}(H_1\overline{Z}_k\overline{Z}_k^T)_{s}\widehat{\Omega}^{-1}\right]\\ \times\left[\overline{Z}_k\varepsilon_{\widehat{b}}\right]_{2}\end{array}\right\}\right)^2\\
&= \frac{1}{\binom{n}{l+2}}\mathbb{E}_{\hat{\theta}}\left(\left\{\begin{array}{c}\left[\varepsilon_{\widehat{p}}\overline{Z}_k^T\right]_{1}\widehat{\Omega}^{-1}\times\\ \prod_{r=0}^{l-1}\left[\{(\Omega-\widehat{\Omega})\widehat{\Omega}^{-1}\}^{t_{r+1}-t_r}(|H_1|\overline{Z}_k\overline{Z}_k^T-\Omega)_{t_{r+1}}\right]\\ \times\{(\Omega-\widehat{\Omega})\widehat{\Omega}^{-1}\}^{j-t_l}\left[\overline{Z}_k\varepsilon_{\widehat{b}}\right]_{2}\end{array}\right\}\right)^2,
\end{aligned}
$$

where $t_0 = 3$. Since the projections contract norm it is enough to control

$$
\begin{aligned}
&\frac{1}{\binom{n}{l+2}}\mathbb{E}_{\hat{\theta}}\left(\left\{\begin{array}{c}\left[\varepsilon_{\widehat{p}}\overline{Z}_k^T\right]_{1}\widehat{\Omega}^{-1}\times\\ \prod_{r=0}^{l-1}\left[\{(\Omega-\widehat{\Omega})\widehat{\Omega}^{-1}\}^{t_{r+1}-t_r}(|H_1|\overline{Z}_k\overline{Z}_k^T)_{t_{r+1}}\right]\\ \times\{(\Omega-\widehat{\Omega})\widehat{\Omega}^{-1}\}^{j-t_l}\left[\overline{Z}_k\varepsilon_{\widehat{b}}\right]_{2}\end{array}\right\}\right)^2\\
&\leq \frac{k^{l+1}}{\binom{n}{l+2}}\left(\frac{\|\Omega-\widehat{\Omega}\|_{op}}{\lambda_{\min}(\widehat{\Omega})}\right)^{2\sum_{r=0}^{l}(t_{r+1}-t_r)}\times\frac{\left(\|H_1\|_\infty\mathbb{E}_\theta\left(\lambda_{\max}(\overline{Z}_k\overline{Z}_k^T)\right)\right)^{l+2}}{\lambda_{\min}^2(\widehat{\Omega})},
\end{aligned}
\tag{5.7}
$$

where $t_{l+1} = j$ and the last inequality follows by Lemma 11. The occurrence of $\mathbb{E}_\theta$ in the right hand side of the inequality is due to fact that we have used $\hat{g} = g$ in our $\hat{\theta}$ and this will allow to use the generating distribution of X in the expectation calculation with respect to $\hat{\theta}$.

Therefore combining (5.5), and (5.7) we have

$$
\begin{aligned}
&\mathbb{E}_\theta\left(\mathbb{U}_n\left((-1)^j\left\{\begin{array}{c}\left[\varepsilon_{\widehat{p}}\overline{Z}_k^T\right]_{i_1}\widehat{\Omega}^{-1}\times\\ \prod_{s=3}^{j}\left[\kappa_s^{(t_1,\ldots,t_l)}(|H_1|\overline{Z}_k\overline{Z}_k^T)_{i_s}\widehat{\Omega}^{-1}\right]\\ \times\left[\overline{Z}_k\varepsilon_{\widehat{b}}\right]_{i_2}\end{array}\right\}\right)\right)^2\\
&\leq 2l\left(1+\left\|\frac{d\mathbb{P}_\theta}{d\mathbb{P}_{\hat{\theta}}}\right\|_\infty\right)^{2l}\\
&\times\mathbb{E}_{\hat{\theta}}\left(\mathbb{U}_n\left((-1)^j\left\{\begin{array}{c}\left[\varepsilon_{\widehat{p}}\overline{Z}_k^T\right]_{i_1}\widehat{\Omega}^{-1}\times\\ \prod_{s=3}^{j}\left[\kappa_s^{(t_1,\ldots,t_l)}(H_1\overline{Z}_k\overline{Z}_k^T)_{i_s}\widehat{\Omega}^{-1}\right]\\ \times\left[\overline{Z}_k\varepsilon_{\widehat{b}}\right]_{i_2}\end{array}\right\}\right)\right)^2
\end{aligned}
$$

$$\leq \frac{2(l+2)\left(\left(1+\left\|\frac{d\mathbb{P}_\theta}{d\mathbb{P}_{\hat{\theta}}}\right\|_\infty\right)^{2(l+2)}\|H_1\|_\infty \mathbb{E}_\theta\left(\lambda_{\max}(\overline{Z}_k\overline{Z}_k^T)\right)\right)^{l+2}}{\binom{n}{l+2}}$$
$$\times \prod_{r=0}^{l}\left\{\left(\frac{\|\Omega-\widehat{\Omega}\|_{op}}{\lambda_{\min}(\widehat{\Omega})}\right)^{2(t_{r+1}-t_r)}\right\} \times \lambda_{\max}^2(\widehat{\Omega}^{-1}) \times k^{l+1}.$$

Therefore under the assumptions of Theorem 5 one has by using the fact that $x^j > 2x$ for any $x > 2$, we have that there exists a $c > 2$ depending on M, B such that

$$\mathbb{E}_\theta\left(\sum_{j=3}^{m}\widehat{IF}_{j,j,k}\right)^2$$
$$= \mathbb{E}_\theta\left(\sum_{j=3}^{m}\sum_{l=0}^{j-2}\sum_{\substack{\{t_1,\ldots,t_l\}\\ \subseteq\{3,\ldots,j\}}}\mathbb{U}_n\left((-1)^j\left\{\begin{array}{c}\left[\varepsilon_{\widehat{p}}\overline{Z}_k^T\right]_{i_1}\widehat{\Omega}^{-1}\times\\ \prod_{s=3}^{j}\left[\kappa_s^{(t_1,\ldots,t_l)}(H_1\overline{Z}_k\overline{Z}_k^T)_{i_s}\widehat{\Omega}^{-1}\right]\\ \times\left[\overline{Z}_k\varepsilon_{\widehat{b}}\right]_{i_2}\end{array}\right\}\right)\right)^2$$
$$= \mathbb{E}_\theta\left(\sum_{l=0}^{m-2}\sum_{j=l+2\vee 3}^{m}\sum_{\substack{\{t_1,\ldots,t_l\}\\ \subseteq\{3,\ldots,j\}}}\mathbb{U}_n\left((-1)^j\left\{\begin{array}{c}\left[\varepsilon_{\widehat{p}}\overline{Z}_k^T\right]_{i_1}\widehat{\Omega}^{-1}\times\\ \prod_{s=3}^{j}\left[\kappa_s^{(t_1,\ldots,t_l)}(H_1\overline{Z}_k\overline{Z}_k^T)_{i_s}\widehat{\Omega}^{-1}\right]\\ \times\left[\overline{Z}_k\varepsilon_{\widehat{b}}\right]_{i_2}\end{array}\right\}\right)\right)^2$$
$$\leq \sum_{l=0}^{m-2}2^{-l}\mathbb{E}_\theta\sum_{l=0}^{m-2}2^l\left(\sum_{j=(l+2)\vee 3}^{m}\sum_{\substack{\{t_1,\ldots,t_l\}\\ \subseteq\{3,\ldots,j\}}}\mathbb{U}_n\left((-1)^j\left\{\begin{array}{c}\left[\varepsilon_{\widehat{p}}\overline{Z}_k^T\right]_{i_1}\widehat{\Omega}^{-1}\times\\ \prod_{s=3}^{j}\left[\kappa_s^{(t_1,\ldots,t_l)}(H_1\overline{Z}_k\overline{Z}_k^T)_{i_s}\widehat{\Omega}^{-1}\right]\\ \times\left[\overline{Z}_k\varepsilon_{\widehat{b}}\right]_{i_2}\end{array}\right\}\right)\right)^2$$
$$\leq \mathbb{E}_\theta\sum_{l=0}^{m-2}\left\{\begin{array}{c}2^l\left(\sum_{j=(l+2)\vee 3}^{m}\binom{j-2}{l}\right)\\ \times\sum_{j=(l+2)\vee 3}^{m}\sum_{\substack{\{t_1,\ldots,t_l\}\\ \subseteq\{3,\ldots,j\}}}\left(\mathbb{U}_n\left((-1)^j\left\{\begin{array}{c}\left[\varepsilon_{\widehat{p}}\overline{Z}_k^T\right]_{i_1}\widehat{\Omega}^{-1}\times\\ \prod_{s=3}^{j}\left[\kappa_s^{(t_1,\ldots,t_l)}(H_1\overline{Z}_k\overline{Z}_k^T)_{i_s}\widehat{\Omega}^{-1}\right]\\ \times\left[\overline{Z}_k\varepsilon_{\widehat{b}}\right]_{i_2}\end{array}\right\}\right)\right)^2\end{array}\right\}$$
$$\leq \sum_{l=0}^{m-2}\left(\sum_{j=(l+2)\vee 3}^{m}\binom{j-2}{l}\right)^2\frac{c^{l+2}k^{l+1}}{\binom{n}{l+2}}.$$

∎

Proof of Theorem 5. The bias control is trivial since $k\log k \ll n$. By simple change of variable $l \to l+2$

$$\sum_{l=0}^{m-2}\left(\sum_{j=(l+2)\vee 3}^{m}\binom{j-2}{l}\right)^2\frac{c^{l+2}k^{l+1}}{\binom{n}{l+2}} = \sum_{l=2}^{m}\left(\sum_{j=l\vee 3}^{m}\binom{j-2}{l-2}\right)^2\frac{c^l k^{l-1}}{\binom{n}{l}}$$

$$
\begin{aligned}
&\leq \sum_{l=2}^{m}\left(\sum_{j=l}^{m}\binom{j-2}{l-2}\right)^2 \frac{c^l k^{l-1}}{\binom{n}{l}} \\
&\overset{(1)}{=} \sum_{l=2}^{m}\left(\binom{m-1}{l-1}\right)^2 \frac{c^l k^{l-1}}{\binom{n}{l}} \\
&\overset{(2)}{\leq} \sum_{l=2}^{m}\left(\frac{m-1}{l-1}\right)^{2(l-1)} \frac{(e^2c)^l k^{l-1}}{\binom{n}{l}} \\
&\leq \sum_{l=2}^{m}\left(\frac{m}{l}\right)^{2l} \frac{c^{*l} k^{l-1}}{\binom{n}{l}} \quad (c^* = e^2c/4) \\
&\overset{(3)}{\leq} \frac{1}{n}\sum_{l=2}^{m}\left(\frac{m}{l}\right)^{2l}\left(\frac{2c^*kl}{n}\right)^{l-1} \frac{2c^* l\sqrt{l}}{e^l} \\
&= \frac{1}{n}\sum_{l=2}^{m}\left(\frac{2c^*km^2}{nl}\right)^{l-1} \frac{2c^*m^2}{e^l\sqrt{l}}.
\end{aligned}
$$

Above (1) and (2) follow by 12 part (ii) and part (i) respectively, and (3) follows by Stirling's approximation bound and Lemma 12 part (i). Letting $k = n/\log n^{\theta_1}$ and $m = \frac{\log n^{\theta_2/2}}{\sqrt{2c^*}}$ we have

$$
\sum_{l=0}^{m-2}\left(\sum_{j=(l+2)\vee 3}^{m}\binom{j-2}{l}\right)^2 \frac{c^{l+2}k^{l+1}}{\binom{n}{l+2}} \leq \frac{2c^*}{n}\sum_{l=2}^{m}\left(\frac{\log n^{\theta_2-\theta_1}}{l}\right)^{l-1} \frac{2c^*\log n^{\theta_2}}{e^l\sqrt{l}}.
$$

Therefore taking $\theta_1 = 3$ and $\theta_2 = 1$ we have desired result by the dominated convergence theorem. ∎

6. Technical Lemmas.

LEMMA 8. *Suppose $O_1, \ldots, O_n \sim \mathbb{P}$ be i.i.d random vectors taking values in a measurable space χ and let $f : \chi^m \to \mathbb{R}$ be a symmetric function of $m \in \mathbb{N}$ arguments. Let f_l denote the l^{th} order degenerate component $(l = 0, \ldots, m)$ of the Hoeffding decomposition of f into m orthogonal components under $\mathbb{P}$. Then*

$$
\mathrm{Var}_{\mathbb{P}}(\mathbb{U}_n(f)) = \sum_{l=1}^{m}\binom{m}{l}\frac{1}{\binom{n}{l}}\mathbb{E}_{\mathbb{P}}f_l^2.
$$

PROOF. Proof is simple by Hoeffding's decomposition. ∎

LEMMA 9 (Lemma 13.1 of Robins et al. (2016b)). *For a measurable space χ with any two probability measures $\mathbb{P} \ll \mathbb{Q}$ and $f : \chi^m \to \mathbb{R}$ any measurable function of $m \in \mathbb{N}$ arguments*

$$
\mathrm{E}_{\mathbb{P}}(\mathbb{U}_n(f))^2 \leq 2m\left(1 + \left\|\frac{d\mathbb{P}}{d\mathbb{Q}}\right\|_\infty\right)^{2m}\mathrm{E}_{\mathbb{Q}}(\mathbb{U}_n(f))^2.
$$

LEMMA 10 (Rudelson (1999)). *Let $Q_1, ..., Q_n$ be a sequence of independent symmetric non-negative $k \times k$-matrix valued random variables with $k \geq 2$ such that $Q = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(Q_i)$ and*

$\sup_{i=1,\ldots,n} \|Q_i\|_{op} \leq M$ *a.s.. Then for* $\hat{Q} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(Q_i)$ *and an absolute constant* $C > 0$

$$\|\hat{Q} - Q\|_{op} \leq C\left(\frac{M \log k}{n} + \sqrt{\frac{M\|Q\|_{op} \log k}{n}}\right).$$

LEMMA 11. *For any given sequences of* $k \times k$ *matrices* $M_0, M_1, \ldots, M_l$ *one has for a constant* C *depending on the choice of basis functions*

$$\mathbb{E}_\theta \left(\left\{ \begin{array}{c} \left[\overline{Z}_k^T\right]_1 M_0 \times \\ \prod_{r=1}^{l-1} \left[M_r (H_1 \overline{Z}_k \overline{Z}_k^T)_{r+2}\right] \\ \times M_l \left[\overline{Z}_k\right]_2 \end{array} \right\} \right)^2$$
$$\leq \left(\|H_1\|_\infty^l \mathbb{E}_\theta^l (\lambda_{\max}(\overline{Z}_k \overline{Z}_k^T)) \prod_{r=0}^{l} (\lambda_{\max}(M_r))^2 \right) k^{l+1},$$

where the expectation is taken over the distribution of $X_1, \ldots, X_l$ *with* $M_0, \ldots, M_l$ *treated as fixed.*

PROOF. The proof follows by writing out the expectation as a multiple integral and then arguing as Lemma 12.4 of Robins et al. (2016b) in conjunction with repeated use of the variational formula of operator norm. ∎

LEMMA 12. *For any three positive integers* $M \geq N \geq K$ *the following hold*

(i) $(N/2)^K/K! \leq \binom{N}{K} \leq \left(\frac{eN}{K}\right)^K$.

(ii) $\sum_{T=N}^{M} \binom{T}{N} = \binom{M+1}{N+1}$.

References.

BELLONI, A., CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics* **186** 345–366.

HÄRDLE, W., KERKYACHARIAN, G., TSYBAKOV, A. and PICARD, D. (1998). *Wavelets, approximation, and statistical applications*. Springer.

NEWEY, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of econometrics* **79** 147–168.

ROBINS, J. M. and RITOV, Y. (1997). Toward a Curse of Dimensionality Appropriate (CODA) Asymptotic Theory for Semi-parametric Models. *Statistics in medicine* **16** 285–319.

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

ROBINS, J. M., LI, L., TCHETGEN, E. T. and VAN DER VAART, A. (2016a). Asymptotic normality of quadratic estimators. *Stochastic Processes and their Applications* **126** 3733–3759.

ROBINS, J., LI, L., MUKHERJEE, R., TCHETGEN, E. T. and VAN DER VAART, A. (2016b). Minimax Estimation of a Functional on a Structured High- Dimensional Model. *The Annals of Statistics (To Appear)*.

RUDELSON, M. (1999). Random vectors in the isotropic position. *Journal of Functional Analysis* **164** 60–72.

TSIATIS, A. (2007). *Semiparametric theory and missing data*. Springer Science & Business Media.

VAN DER VAART, A. W. (2000). *Asymptotic statistics* **3**. Cambridge university press.

VAN DER VAART, A. W., DUDOIT, S. and LAAN, M. J. (2006). Oracle inequalities for multi-fold cross validation. *Statistics & Decisions* **24** 351–371.