Published in final edited form as:

Stat Probab Lett. 2011 July 1; 81(7): 821–828. doi:10.1016/j.spl.2011.02.030.

Higher Order Inference On A Treatment Effect Under Low Regularity Conditions

Lingling Li[a], **Eric Tchetgen Tchetgen**[b], **Aad van der Vaart**[c], and **James M. Robins**[b]

[a] Harvard Pilgrim Health Care Institute and Harvard Medical School

[b] Harvard University

[c] Vrije Universiteit

Abstract

We describe a novel approach to nonparametric point and interval estimation of a treatment effect in the presence of many continuous confounders. We show the problem can be reduced to that of point and interval estimation of the expected conditional covariance between treatment and response given the confounders. Our estimators are higher order U-statistics. The approach applies equally to the regular case where the expected conditional covariance is root-n estimable and to the irregular case where slower non-parametric rates prevail.

Keywords

Minimax; U-statistics; Influence functions; Nonparametric; Semi-parametric; Robust Inference

1. Introduction

We consider perhaps the central problem in biostatistics, epidemiology, and econometrics: the estimation of a treatment effect in the presence of a high dimensional vector X of confounding covariates. To this end, for a binary treatment A and a response Y, let τ be the variance weighted average treatment effect

$$\tau \equiv E[var(A|X)\gamma(X)]/E[var(A|X)]=E[cov(Y,A|X)]/E[var(A|X)],$$
$$\text{with } \gamma(x) \equiv E(Y|A=1,X=x) - E(Y|A=0,X=x),$$

where a simple calculation establishes the equality in the first line, and $\gamma(x)$ is the average treatment effect among subjects with $X = x$ under the assumption of no unmeasured confounding (ignorable treatment assignment given X).

Our motivation for τ as our functional of interest is as follows. The most common model for estimation of a causal effect assumes $\gamma(X) = \beta$ does not depend on X wp 1, which is unlikely to hold exactly. Most semiparametric estimators of β, including those of Robinson (1988)

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

and Donald and Newey (1994), converge in probability to τ even if the assumption $\gamma(X) = \beta$ is false. An alternative motivation is considered by Crump et al. (2006).

We now show that point and interval estimators for τ can be constructed from point and interval estimators for the numerator $E[cov(Y, A|X)]$ of τ. As a consequence, until Section 6, the paper is devoted to constructing point and interval estimators for $E[cov(Y, A|X)]$. In Section 6, we translate these estimators into estimators for τ.

For any fixed $\tau^* \in R$, define $Y(\tau^*) = Y - \tau^* A$ and the corresponding functional

$$\psi(\tau^*)=E[\{Y(\tau^*)-E[Y(\tau^*)|X]\}\{A-E(A|X)\}]=E[cov(Y(\tau^*),A|X)].$$

τ is the unique solution to $\psi(\tau^*) = 0$. Suppose that we can construct point estimators $\hat{\psi}(\tau^*)$ and $(1 - \alpha)$ interval estimators for $\psi(\tau^*)$. Then $\hat{\tau}$ satisfying $\hat{\psi}(\hat{\tau}) = 0$ is an estimator of τ. Further, a $(1 - \alpha)$ confidence set for τ is the set of τ^* for which a $(1 - \alpha)$ interval estimator for $\psi(\tau^*)$ contains zero. Until Section 6, we take $\tau^* = 0$, and consider inference for the expected conditional covariance $\psi \equiv E[cov\{Y, A|X\}]$.

Henceforth, we assume we observe N iid copies of $O = (Y, A, X)$ such that the marginal distribution F_O of X has a Lebesgue density f in R^d that has a compact support. We assume F_O is contained in a nonparametric model $M(\Theta) = \{F(\cdot; \theta); \theta \in \Theta\}$, indexed by the (infinite dimensional) parameter $\theta \in \Theta$. In this notation, our parameter of interest is the unique solution $\tau(\theta)$ to $\psi(\tau^*, \theta) = 0$ with $\psi(\tau^*, \theta) \equiv E_\theta[cov_\theta(Y(\tau^*), A|X)]$ and, until Section 6, we consider inference on

$$\psi(\theta)\equiv\psi(0,\theta)=E_\theta[cov_\theta(Y,A|X)].$$

We let $b: x \mapsto b(x) = E[Y|X = x]$, $p: x \mapsto p(x) = E[A|X = x]$, and $f: x \mapsto f(x)$ denote the components of θ corresponding to the conditional expectations of Y and A given $X = x$ and the density of the marginal distribution F_X of X. Our model $M(\Theta)$ places no restrictions on F_O, other than (i) bounds on the L_p norms of these functions to insure all integrals are bounded and (ii) explicit smoothness bounds that specify that $b(x)$, $p(x)$, and $f(x)$ are in known Hölder classes β_b, β_p, and β_f. Informally, a function $h(x)$ is in a Hölder class β_h if all partial derivatives of $h(\cdot)$ up to order $\lfloor\beta_h\rfloor$ exist and are bounded by a constant C_h and the partial derivatives of order $\lfloor\beta_h\rfloor$ are Hölder with exponent $\beta_h - \lfloor\beta_h\rfloor$ and bound C_h. Recall that a function $q(x)$ is Hölder with exponent a and bound c if $|q(x) - q(x^*)| < c|x - x^*|^a$ with $a < 1$ for all x, x^*. A formal definition of our model and of a Hölder class are given in the web-supplement.

Robins et al. (2009b) proved that in model $M(\Theta)$

$$(\beta_b+\beta_p)/d\geq 1/2 \tag{1}$$

is a necessary condition for the existence of a $\sqrt{N}$–consistent estimator of $\psi(\theta)$.

We introduce a novel class of point and interval estimators for $\psi(\theta)$ that can be applied in both the “regular” case where condition (1) holds and in the “irregular” case where condition (1) does not hold. Our novel estimators are U-statistics. In previous work we derived these estimators using an abstract theory of higher order influence functions (Robins

(2004), Robins et al. (2008) and Robins et al. (2009a)). In this paper we derive these estimators using a much more accessible bias correction procedure.

In section 2 we assume that condition (1) holds. However Robins and Ritov (1997) argue that, in epidemiologic studies in which the dimension d of X is not small, the large sample behavior of estimators derived under asymptotics that assumes condition (1) often fails to provide an accurate guide to their actual finite sample behavior; therefore we study the irregular case in Section 3.

For two sequences of random variables X_N and Y_N, the notation $X_N \lesssim Y_N$ means $X_N \leq CY_N$ for a constant C that is fixed in the context. The notations $X_N \asymp Y_N$ mean $X_N \lesssim Y_N$ and $Y_N \lesssim X_N$. The notations $X_N \sim Y_N$ and $X_N \ll Y_N$ mean that $\frac{X_N}{Y_N} \xrightarrow{P} 1$ and $\frac{X_N}{Y_N} \xrightarrow{P} 0$. For convenience, we will drop the N subscript and write X and Y for X_N and Y_N.

2. Failure of First Order Inference in The Regular Case

By definition, an estimator $\hat{\psi}$ is a regular asymptotically linear (RAL) estimator of $\psi(\theta)$ if and only if

$$N^{-1/2}\left(\widehat{\psi}-\psi(\theta)\right)=N^{-1/2}\sum_{i=1}^{N}U_{1,i}(\theta)+o_P(1), \tag{2}$$

$$U_1(\theta)=\{Y-b(X)\}\{A-p(X)\}-\psi(\theta) \tag{3}$$

Here U_1 (θ) is the so called first order influence function of ψ (θ). By Slutsky's theorem $N^{1/2}(\hat{\psi}-\psi\ (\theta))$ is asymptotically normal with mean zero and variance var $\{U_1\ (\theta)\}$. Thus a RAL estimator converges to ψ (θ) at rate $N^{-\frac{1}{2}}$. Consider the plug-in estimator ψ $(\hat{\theta})$ and the one step estimator $\psi\left(\widehat{\theta}\right)+\frac{1}{N}\sum_{i=1}^{N}U_{1,i}(\widehat{\theta})=\frac{1}{N}\sum_{i=1}^{N}\{Y_i-\widehat{b}(X_i)\}\{A_i-\widehat{p}(X_i)\}$, where $\hat{\theta}$ is a rate-optimal nonparametric estimator of θ (i.e of F_O, Härdle et al. (1998)). If ψ $(\hat{\theta})$ is RAL then so is the one step estimator but not vice-versa, as the onestep estimator may have smaller asymptotic bias with the same asymptotic variance (Bickel et al. (1998)).

In this paper, we require a modified version of the one step estimator in which $\hat{b}$, $\hat{p}$, $\hat{f}$, and thus $\hat{\theta}$ are estimated from a separate, randomly-chosen training sample of size $N - n$, and the modified one-step estimator is $\widehat{\psi}_1 \equiv \psi_1\left(\widehat{\theta}\right)=\frac{1}{n}\sum_{i=1}^{N}\{Y_i-\widehat{b}(X_i)\}\{A_i-\widehat{p}(X_i)\}$, where the sum is over the n subjects in the estimation sample. The original one step estimator and $\hat{\psi}_1$ will generally have the same rate of convergence and order of asymptotic bias if $(N - n) \asymp n$ (which we assume to be true unless stated otherwise). This modification is made because Hölder classes with $\beta < d/2$ are not Donsker (Van der Vaart and Wellner (1996)). Henceforth, all expectations and variances are to be interpreted as conditional on the training sample and thus are random, although for convenience, we sometimes suppress this fact in the notation, especially for variances.

Conditional on the training sample, the estimator $\hat{\psi}_1$ is the sum of n independent random variables. Hence, it is conditionally asymptotically normal with mean $E_\theta[(b(X) - \hat{b}(X))(p(X) - \hat{p}(X))] + \psi(\theta)$ and variance of order $\frac{1}{n}$ (Bickel et al. (1998)). Thus, the interval $\mathcal{C}_1 = \hat{\psi}_1 \pm z_{\alpha/2}s.e.(\hat{\psi}_1)$ is a honest asymptotic confidence interval if and only if the maximal bias

$BI(\widehat{\psi}_1) \equiv \sup_{\theta \in \Theta} |BI(\widehat{\psi}_1, \theta)|$ is $o_P(n^{-1/2})$, where the subscript p reflects the randomness in $BI(\hat{\psi}_1)$ due to the training sample. Thus, $BI(\hat{\psi}_1)$ is of smaller order than $s.e.(\hat{\psi}_1) \asymp n^{-1/2}$. Here $BI(\hat{\psi}_1, \theta) = E_\theta[(b(X) - \hat{b}(X))(p(X) - \hat{p}(X))]$ is the bias under θ. A formal definition of a honest asymptotic confidence interval is given in the web-supplement. In addition, $\hat{\psi}_1$ has a uniform convergence rate of $n^{-\frac{1}{2}}$ (i.e. is $\sqrt{n}$-consistent) if and only if $BI(\widehat{\psi}_1)=O_p\left(n^{-\frac{1}{2}}\right)$.

If $\hat{b}$ and $\hat{p}$ are rate optimal estimators of b and p, they have convergence rates $n^{-\frac{\beta_b}{2\beta_b+d}}$ and $n^{-\frac{\beta_p}{2\beta_p+d}}$. Hence, $BI(\widehat{\psi}_1) \asymp (N-n)^{-\left(\frac{\beta_b}{2\beta_b+d}+\frac{\beta_p}{2\beta_p+d}\right)}$ (i.e., $n^{-\left(\frac{\beta_b}{2\beta_b+d}+\frac{\beta_p}{2\beta_p+d}\right)}$ when $(N - n) \asymp n$). Hence even when condition (1) holds, $BI(\hat{\psi}_1)$ can exceed $O_P(n^{-1/2})$. For example, if $\beta_b = \beta_p$ then for $BI(\hat{\psi}_1)$ to be $O_P(n^{-1/2})$ requires that $\beta_b + \beta_p \geq d$. In fact, if $\beta_p = 0$ holds, then $BI(\hat{\psi}_1) \gg n^{-1/2}$ for any finite β_b. Thus, to construct a uniform $\sqrt{n}$-consistent estimator for $\psi(\theta)$ whenever condition (1) holds, we require an estimator with smaller bias than $\hat{\psi}_1$. To achieve this, we will subtract from $\hat{\psi}_1$ a bias correction term which estimates the bias $BI(\hat{\psi}_1, \theta)$.

3. Second Order U-statistics Estimators

3.1. The Estimator

To motivate our bias correction term, suppose that X were categorical with known probability mass function f. Define the residuals $\hat{\varepsilon}_i \equiv Y_i - \hat{b}(X_i)$, $\hat{\Delta}_j \equiv A_j - \hat{p}(X_j)$, and kernel function $K_f(X_i, X_j)=\frac{I(X_i=X_j)}{f(X_i)}$. Then $\frac{1}{n(n-1)}\sum_{i \neq j}\widehat{\varepsilon}_i K_f(X_i, X_j)\widehat{\Delta}_j$ is an unbiased estimator of $BI(\hat{\psi}_1, \theta)$. Since f is unknown, we use $K_{\hat{f}}(X_i, X_j)$ instead. By analogy, for X continuous, if we could find a “kernel” $K_{f,\infty}(x, X)$ such that

$$\begin{aligned} r(x) &= E_f[K_{f,\infty}(x,X)r(X)] \\ &\equiv \int K_{f,\infty}(x,x^*)r(x^*)f(x^*)\,dx^* \text{ for all } r(\cdot) \in L_2(f), \end{aligned} \tag{4}$$

then the statistic $\frac{1}{n(n-1)}\sum_{i \neq j}\widehat{\varepsilon}_i K_{f,\infty}(X_i, X_j)\widehat{\Delta}_j$ would be unbiased for $BI(\hat{\psi}_1, \theta)$.

A kernel satisfying eq. (4) is referred to as a Dirac delta function wrt to the measure F_X and is known not to exist in $L_2[F_X] \times L_2[F_X]$. However, the above motivates the construction of a class of estimators for $BI(\hat{\psi}_1, \theta)$ using “truncated Dirac kernels”.

Let $\{z_l(\cdot)\} \equiv \{z_l(x); l = 1, 2, \ldots\}$ be dense in $L_2(\mu)$ with μ the Lebesgue measure and let $\bar{z}_k(x)^T = (z_1(x), \ldots, z_k(x))$. Define, for $\hat{f}$ a component of $\hat{\theta}$, $\bar{\varphi}_k(X) = (E_{\hat{f}}[\bar{z}_k(X)\,\bar{z}_k(X)^T])^{-1/2}\,\bar{z}_k(X)$ so $E_{\hat{f}}[\bar{\varphi}_k(X)\,\bar{\varphi}_k(X)^T] = I_{k\times k}$. Here $\hat{f}$ is a rate optimal estimator of f with convergence rate $(N-n)^{-\frac{\beta_f}{2\beta_f+d}}$ in $L_q(\mu)$ for q finite. Let $K_{\hat{f},k}(X_i, X_j) = \bar{\varphi}_k(X_i)^T\,\bar{\varphi}_k(X_j)$. Then, for any $h(x)$, the projection $\Pi_{\hat{f}}[h(x)|\bar{z}k(x)] \equiv \Pi_{\hat{f}}[h(x)|lin\{\bar{z}_k(x)\}]$ under $\hat{f}$ of $h(x)$ onto the subspace $lin\{\bar{z}_k(x)\}$ spanned by the elements of $\bar{z}_k(x)$ is $E_{\hat{f}}[K_{\hat{f},k}(x, X)\,h(X)]$. Thus, by definition, $K_{\hat{f},k}(x, X)$, is the associated projection kernel. Note that $\Pi_{\hat{f}}[h(x)|\bar{z}_k(x)] = \Pi_{\hat{f}}[h(x)|\bar{\varphi}_k(x)]$ since $lin\{\bar{z}_k(x)\}$ and $lin\{\bar{\varphi}_k(x)\}$ are equal.

$K_{\hat{f},k}(x, X)$ is a truncated at k approximation to $K_{\hat{f},\infty}(x, X)$ in the sense that, with $\hat{f}$ substituted for f, it satisfies eq. (4) for $r(x) \in lin\{\bar{z}_k(x)\}$. Our bias corrected estimator is then

$$\widehat{\psi}_{2,k} \equiv \widehat{\psi}_1 - \frac{1}{n(n-1)} \sum_{i \neq j} \widehat{\varepsilon}_i K_{\widehat{f},k}(X_i, X_j) \widehat{\Delta}_j.$$

3.2. Bias and Variance Properties of $\hat{\psi}_{2,k}$

The bias of $\hat{\psi}_{2,k}$ is given in the following theorem proved in the web-supplement. The bias can be decomposed into the sum of two terms-the truncation bias and the estimation bias. The truncation bias is due to the truncated at k approximation $K_{f,k}$ (X_i, X_j) to $K_{f,\infty}$ (X_i, X_j), whereas the estimation bias comes from using $\hat{b}(X_i)$, $\hat{p}(X_i)$, and $\hat{f}(X_i)$ to estimate $b(X_i)$, $p(X_i)$ and $f(X_i)$.

In the following, since $f \in \theta$, we can and sometimes do write the projection operator Π_f as Π_θ. Let $\Pi_\theta^{\perp}[h(X)|\overline{\varphi}_k(X)] = h(X) - \Pi_\theta[h(X)|\overline{\varphi}_k(X)]$ be the projection under θ of $h(X)$ onto the orthocomplement of $lin\{\bar{z}_k\ (X)\} = lin\{\bar{\varphi}_k\ (X)\}$.

Theorem 1—Suppose regularity conditions (A.1)– (A.2) of the web-supplement hold. Then the (conditional) bias $\mathrm{BI}\ (\hat{\psi}_{2,k}, \theta) \equiv \mathrm{E}_\theta[\hat{\psi}_{2,k}] - \psi\ (\theta)$ equals $\mathrm{TB}_k\ (\theta) + \mathrm{EB}_{2,k}\ (\theta)$ where

$$TB_k(\theta) = E\left[\left(\Pi_\theta^{\perp}\left[\left(b(X) - \widehat{b}(X)\right)|\overline{\varphi}_k(X)\right]\right) \times (\Pi_\theta^{\perp}[(p(X) - \widehat{p}(X))|\overline{\varphi}_k(X)])\right] \quad (5)$$

and

$$EB_{2,k}(\theta) = \left\{E_\theta\left[\left(b(X) - \widehat{b}(X)\right)\overline{\varphi}_k(X)^T\right] \times \left[(E_\theta[\overline{\varphi}_k(X)\overline{\varphi}_k^T(X)])^{-} - I_{k\times k}\right] \times E_\theta[\overline{\varphi}_k(X)(p(X) - \widehat{p}(X))]\right\} \quad (6)$$

The next theorem, proved in the web-supplement, derives the orders of TB_k (θ) and $EB_{2,k}$ (θ) for a choice of $\bar{Z}_k \equiv \bar{z}_k$ (X), that provides optimal uniform approximation error of order $k^{-\beta/d}$ for any function $h(x)$ of a d-dimensional x in a Hölder class with exponent β. That is $h(x) - \Pi_\theta[(h(x))|\overline{z}_k(x)] = \Pi_\theta^{\perp}[(h(x))|\overline{z}_k(x)]$ is of order $k^{-\beta/d}$ in sup norm. Polynomial, spline and suitable wavelet bases all satisfy this assumption.

Theorem 2—Suppose that regularity conditions (A.1) – (A.3) of the web-supplement are satisfied. Then with $\mathrm{BI}\ (\hat{\psi}_{2,k}) = \sup_{\theta\in\Theta}|\mathrm{BI}(\hat{\psi}_{2,k}, \theta)|$, $\mathrm{TB}_k = \sup_{\theta\in\Theta}\{\mathrm{TB}_k\ (\theta)\}$, $\mathrm{EB}_2 = \sup_{\theta\in\Theta}|\mathrm{EB}_{2,k}\ (\theta)|$,

$$\begin{gathered} TB_k = O_p(k^{-\frac{\beta_b+\beta_p}{d}}) \\ EB_2 = O_p\left((N-n)^{-\left(\frac{\beta_b}{d+2\beta_b} + \frac{\beta_p}{d+2\beta_p} + \frac{\beta_f}{2\beta_f+d}\right)}\right) \end{gathered} \quad (7)$$

$$\begin{aligned} BI\left(\widehat{\psi}_{2,k}\right) &= \max(TB_k, EB_2) \\ &= O_p\left(\max\left(k^{-\frac{\beta_b+\beta_p}{d}}, (N-n)^{-\left(\frac{\beta_b}{d+2\beta_b} + \frac{\beta_p}{d+2\beta_p} + \frac{\beta_f}{2\beta_f+d}\right)}\right)\right) \end{aligned} \quad (8)$$

Note the order of the maximal bias of $EB_{2,k}\ (\theta)$ does not depend on k. The theorem is proved in the web-supplement. A heuristic argument is as follows. If, as is always possible, our optimal estimates of $\hat{b}(x)$ and $\hat{p}(x)$ are in $lin\{\bar{z}_k\ (x)\} = lin\{\bar{\varphi}_k\ (x)\}$, then $TB_k\ (\theta)$ depends on the product of $\Pi_\theta^\perp[b(X)|\overline{\varphi}_k(X)]$ and $\Pi_\theta^\perp[(p(X))|\overline{\varphi}_k(X)]$, which is $O\left(k^{-\frac{\beta_b+\beta_p}{d}}\right)$. Next, noting

$$\left(E_f\left[\overline{\varphi}_k(X)\overline{\varphi}_k^T(X)\right]\right)^{-1}-I_{k\times k}=\left[\left(I_{k\times k}-E_f\left[\overline{\varphi}_k(X)\overline{\varphi}_k^T(X)\right]\right)\right]\left(E_f\left[\overline{\varphi}_k(X)\overline{\varphi}_k^T(X)\right]\right)^{-}$$

and

$$I_{k\times k}-E_f\left[\overline{\varphi}_k(X)\overline{\varphi}_k^T(X)\right]=E_{\widehat{f}}\left[\left(\frac{\widehat{f}(X)-f(X)}{\widehat{f}(X)}\right)\overline{\varphi}_k(X)\overline{\varphi}_k^T(X)\right],$$

we observe that $EB_{2,k}\ (\theta)$ is a product of terms in $(b(X) - \hat{b}(X))$, $(p(X) - \hat{p}(X))$ and $(f(X) - \hat{f}(X))$.

The following theorem proved in the web-supplement gives the order of the (conditional) variance of $\hat{\psi}_{2,k}$.

Theorem 3—Assume (A.1) – (A.3) are satisfied, then conditional on the training sample,

$$var_\theta\left[\widehat{\psi}_{2,k}\right]\asymp\max\left(\frac{1}{n},\frac{k}{n^2}\right) \tag{9}$$

3.3. Convergence Rate of the Optimal Estimator in the Class $\hat{\psi}_{2,k}$: $k \in \mathcal{N}$}

3.3.1. The regular case - Eq. (1) holds—In this subsection, condition (1) holds so $N^{-1/2}$ is a lower bound on the rate of convergence.

Lemma 4: Given (1) and $(N - n) \asymp n$, (i) $\hat{\psi}_{2,n} \equiv \hat{\psi}_{2,k=n}$ converges at rate $n^{-1/2}$ (and thus is rate minimax) if and only if

$$\frac{\beta_b}{d+2\beta_b}+\frac{\beta_p}{d+2\beta_p}+\frac{\beta_f}{d+2\beta_f}\ \geq\ \frac{1}{2}. \tag{10}$$

and (ii) no estimator $\hat{\psi}_{2,k}$ converges at rate $n^{-1/2}$ if $\hat{\psi}_{2,n}$ does not.

Proof: $var_\theta\left[\widehat{\psi}_{2,k}\right]\asymp\max\left(\frac{1}{n},\frac{k}{n^2}\right)$ has variance of order $O(n^{-1})$ only if $k = O\ (n)$. Among all $\hat{\psi}_{2,k}$ with $k = O\ (n)$, $TB_k=O_p\left(k^{-\frac{\beta_b+\beta_p}{d}}\right)$ is minimized for $k \asymp n$, proving (ii). Further $TB_n = O_p(n^{-1/2})$ by condition (1). Finally, when (10) holds, $EB_2 = O_p(n^{-1/2})$. Hence $\hat{\psi}_{2,n}$ converges at rate $n^{-1/2}$.

Recall $\hat{\psi}_1$ has maximal bias $BI(\hat{\psi}_1) \lesssim n^{-1/2}$ (and thus converges at rate $n^{-1/2}$) if and only if $\frac{\beta_b}{d+2\beta_b}+\frac{\beta_p}{d+2\beta_p}\geq 1/2$. As an example, with $\beta_b=\beta_p=\frac{d}{3}$, the bias of $\hat{\psi}_1$ shrinks to zero at rate

$n^{-\frac{2}{5}} \gg n^{-\frac{1}{2}}$; in contrast, $\hat{\psi}_{2,n}$ converges at rate $n^{-1/2}$ as long as $\beta_f/d > \frac{1}{8}$. Thus the second order U-statistic added to $\hat{\psi}_1$ to form $\hat{\psi}_{2,n}$ has reduced the bias to $O_p(n^{-1/2})$ without any increase in the order of the variance since $k/n^2 \asymp 1/n$ when $k \asymp n$.

In Section 4, we shall construct an estimator that converges at the minimax rate of $n^{-1/2}$ when eq. (1) holds, even though neither $\hat{\psi}_1$ nor $\hat{\psi}_{2,n}$ converges at rate $n^{-1/2}$ because (10) fails to hold and so $EB_2 \gg n^{-1/2}$.

Finally suppose that eqs. (1) and (10) hold with strict inequalities. Then TB_n and EB_2 are $o_p(n^{-1/2})$. Hence, with $(N - n) \asymp n$, $N^{1/2}(\hat{\psi}_{2,n} - \psi(\theta))$ is asymptotically normal with mean zero and finite variance. However $\hat{\psi}_{2,n}$ does not achieve the optimal constant as its asymptotic variance exceeds the semiparametric variance bound $var_\theta[U_1(\theta)]$. This deficiency can be remedied by no longer choosing $(N - n) \asymp n$. Specifically, arguing as on page 379 in Robins et al. (2009b), if we make the ratio $(N - n)/N$ to be of order 1/log (N) rather than of order 1 and take $k_{eff} \asymp n/log(n)$, then $N^{1/2}\left(\widehat{\psi}_{2,k_{eff}} - \psi(\theta)\right) = N^{-1/2}\sum_{i=1}^{N} U_{1,i}(\theta) + o_P(1)$; hence $\hat{\psi}_{2,k_{eff}}$ is asymptotically linear and normal with variance $var_\theta[U_1(\theta)]$ and thus semi-parametric efficient.

3.3.2. The irregular case - Eq. (1) does not hold

Suppose condition (1) does not hold. In that case Robins et al. (2009b) proved that a lower bound for the minimax rate is $n^{-\frac{2(\beta_b+\beta_p)/d}{1+2(\beta_b+\beta_p)/d}} \gg n^{-1/2}$. The following Lemma shows that, if

$$\beta_f \geq d \times \frac{\xi_{\min}(\beta_b,\beta_p,d)}{1-2\xi_{\min}(\beta_b,\beta_p,d)}, \text{ where}$$
$$\xi_{\min}(\beta_b,\beta_p,d) = \frac{2(\beta_b+\beta_p)/d}{1+2(\beta_b+\beta_p)/d} - \frac{\beta_b/d}{1+2\beta_b/d} - \frac{\beta_p/d}{1+2\beta_p/d} \tag{11}$$

holds, $\hat{\psi}_{2,k*}$ with $k_* \asymp n^{\frac{2}{1+2(\beta_b+\beta_p)/d}}$ is rate minimax.

Lemma 5: If (11) holds, i) $\hat{\psi}_{2,k*}$ converges at rate $n^{-\frac{2(\beta_b+\beta_p)/d}{1+2(\beta_b+\beta_p)/d}}$, which is thus minimax and (ii) no estimator $\hat{\psi}_{2,k}$ converges at this rate if $\hat{\psi}_{2,k*}$ does not.

Proof: Consider $\hat{\psi}_{2,k*}$ with $k_* = n^{\frac{2}{1+2(\beta_b+\beta_p)/d}}$. The standard error $\{\max(\frac{1}{n}, \frac{k_*}{n^2})\}^{1/2}$ and the truncation bias TB_{k*} of $\hat{\psi}_{2,k*}$ both are of order $n^{-\frac{2(\beta_b+\beta_p)/d}{1+2(\beta_b+\beta_p)/d}}$, proving (ii). When (11) also holds, $\mathrm{EB}_2 \lesssim n^{-\frac{2(\beta_b+\beta_p)/d}{1+2(\beta_b+\beta_p)/d}}$.

In Section 4, we construct an estimator that, often converges faster (and never slower) than $\hat{\psi}_{2,k*}$, when (11) does not hold, although the rate remains slower than $n^{-\frac{2(\beta_b+\beta_p)/d}{1+2(\beta_b+\beta_p)/d}}$.

4. U-Statistic estimators

We next show that we can construct a new estimator $\widehat{\psi}_{3,k} = \widehat{\psi}_{2,k} - \mathbb{H}_{3,3}^{(k)}$ that subtracts from $\hat{\psi}_{2,k}$ a third order U-statistic, denoted by $\mathbb{H}_{3,3}^{(k)}$, which estimates the estimation bias $EB_{2,k}(\theta)$ of $\hat{\psi}_{2,k}$. In fact we show that we can iterate this process to construct new estimators $\widehat{\psi}_{m,k} = \widehat{\psi}_{m-1,k} - \mathbb{H}_{m,m}^{(k)} = \widehat{\psi}_{2,k} - \sum_{j=3}^{m} \mathbb{H}_{j,j}^{(k)}, m=3, \ldots$ that subtract from $\hat{\psi}_{m-1,k}$ a m^{th} order U-statistic

$\mathbb{H}_{m,m}^{(k)}$, which estimates the estimation bias $EB_{m-1,k}(\theta)$ of $\hat{\psi}_{m-1,k}$. In the web-supplement we prove the following theorem

Theorem 6—Under assumptions (A.1) – (A.3) and with each z_l (x) the tensor product of elements of a univariate compact wavelet basis with optimal approximation properties, for m = 3, …, the estimator $\widehat{\psi}_{m,k}=\widehat{\psi}_{2,k}-\sum_{j=3}^{m}\mathbb{H}_{j,j}^{(k)}$ has (i) truncation bias TB_k (θ) for all m, (ii) estimation bias $EB_{m,k}$ (θ) of smaller order than $EB_{m-1,k}$ (θ), total bias $BI(\hat{\psi}_{m,k}, \theta) \equiv E_\theta[\hat{\psi}_{m,k}] - \psi(\theta) = TB_k(\theta) + EB_{m,k}(\theta)$ and (iii) variance of the same order as $\hat{\psi}_{2,k}$ when k = O (n) but of greater order than that of $\hat{\psi}_{m-1,k}$ when k ≫ n. Here

$$\mathbb{H}_{m,m}^{(k)} \equiv \frac{1}{n(n-1)(n-2)\times\ldots(n-(m-1))}\sum_{i_1\neq i_2\ldots i_3\neq\ldots\neq i_m} H_{m,m,\bar{i}_m}^{(k)} \quad with$$
$$H_{m,m,\bar{i}_m}^{(k)}=(-1)^m\widehat{\varepsilon}_{i_1}\overline{\varphi}_k(X_{i_1})^T\prod_{r=3}^{m}\left\{\left(\overline{\varphi}_k(X_{i_r})\overline{\varphi}_k(X_{i_r})^T-I_{k\times k}\right)\right\}\overline{\varphi}_k(X_{i_2})\widehat{\Delta}_{i_2}. \tag{12}$$

Specifically

$$EB_{m,k}(\theta) \equiv (-1)^m\left\{E_\theta\left[\left(b(X)-\widehat{b}(X)\right)\overline{\varphi}_k(X)^T\right]\times\left\{\left(E_\theta\left[\overline{\varphi}_k(X)\overline{\varphi}_k(X)^T\right]\right)^{-1}-I_{k\times k}\right\}\times\left\{E_\theta\left[\overline{\varphi}_k(X)\overline{\varphi}_k(X)^T\right]-I_{k\times k}\right\}^{m-2}\times E_\theta[\overline{\varphi}_k(X)(p(X)-\widehat{p}(X)\right. \tag{13}$$

$$EB_m=\sup_{\theta\in\Theta}|EB_{m,k}(\theta)| \asymp n^{-\left(\frac{\beta_b}{d+2\beta_b}+\frac{\beta_p}{d+2\beta_p}+\frac{2\beta_f}{2\beta_f+d}\right)} \asymp EB_{m-1}\times n^{-\left(\frac{\beta_f}{2\beta_f+d}\right)} \tag{14}$$

$$var_\theta\left[\widehat{\psi}_{m,k}\right] \asymp \frac{1}{n}\max\left(1,\left(\frac{k}{n}\right)^{m-1}\right) wp\ 1. \tag{15}$$

Remark 1—The assumption that each z_l (x) is the tensor product of compact wavelets is only used in the proof of (iii) for technical reasons. We expect that (iii) holds for many other bases.

Remark 2—In this notation we could write $\widehat{\psi}_{2,k}=\widehat{\psi}_1-\mathbb{H}_{2,2}^{(k)}$ with $\mathbb{H}_{2,2}^{(k)}=\frac{1}{n(n-1)}\sum_{i\neq j}\widehat{\varepsilon}_i K_{\widehat{f},k}(X_i,X_j)\widehat{\Delta}_j$.

4.1. Convergence Rate of the Optimal Estimator in the Class {$\hat{\psi}_{m,k}$: m = 2, …, k ∈ $\mathcal{N}$}

4.1.1. The regular case - Eq. (1) holds—

In this subsection, condition (1) holds so $N^{-1/2}$ is a lower bound on the rate of convergence.

Lemma 7: Given condition (1), $\beta_f > 0$, and $(N - n) \asymp n$, $\hat{\psi}_{m_{opt},n} \equiv \hat{\psi}_{m_{opt},k=n}$ converges at rate $n{-}^{1/2}$ (and thus is rate minimax) where m_{opt} is the smallest integer for which $\rho_m \equiv \frac{\beta_b}{d+2\beta_b}+\frac{\beta_p}{d+2\beta_p}+\frac{(m-1)\beta_f}{2\beta_f+d}>1/2$.

Proof: Since ρ_m increases without bound as $m \to \infty$, m_{opt} always exists when $\beta_f > 0$ and $EB_{m_{opt}}$ is $o_P\left(n^{-\frac{1}{2}}\right)$. Further $var_\theta\left[\widehat{\psi}_{m_{opt},n}\right] \asymp \frac{1}{n}\max\left(1-\left(\frac{n}{n}\right)^{m_{opt}-1}\right) \asymp \frac{1}{n}$ and $TB_n = O\left(n^{-\frac{1}{2}}\right)$ by condition (1).

The key point is the same as in the case discussed in Section 3.3. The U-statistic terms of $\hat{\psi}_{m_{opt},n}$ reduce the order of the estimation bias below $n^{-\frac{1}{2}}$, and yet do not increase the order of the variance or truncation bias. Thus by introducing the U-statistic estimators of arbitrarily large order m, we are able to construct $\sqrt{n}$-consistent estimators for $\psi(\theta)$ for any value of $\beta_f > 0$, as long as condition (1) holds.

Although $\hat{\psi}_{m_{opt},n}$ fails to be semiparametric efficient this defficiency can be remedied as follows.

Lemma 8: Assume condition (1) holds with a strict inequality. Let $(N - n)/N=1/\log(N)$ so $n = N(1 - 1/\log(N))]$. Let m_{opt*} be the smallest integer for which $[\log_N(N/\log(N))]\left\{\frac{\beta_b}{d+2\beta_b}+\frac{\beta_p}{d+2\beta_p}+\frac{(m-1)\beta_f}{2\beta_f+d}\right\}>1/2$, and $k_{eff} \asymp n/\log(n)$. Then (i) $\hat{\psi}_{m_{opt*},k_{eff}}$ has $TBk_{eff} = o_p(N^{-1/2})$, (ii) $EB_{m_{opt*}} = o_p(N^{-1/2})$, and (iii) $N^{1/2}\left(\widehat{\psi}_{m_{opt*},k_{eff}} - \psi(\theta)\right) = N^{-1/2}\sum_{i=1}^{N} U_{1,i}(\theta)+o_P(1)$; hence $\hat{\psi}_{m_{opt*},k_{eff}}$ is semiparametric efficient.

4.1.2. The irregular case - Eq. (1) does not hold—

Suppose condition (1) does not hold so estimation of $\psi(\theta)$ at rate $N^{-1/2}$ is not possible. For any fixed $m \geq 2$, let $k_*(m)=n^{\frac{m}{m-1+2(\beta_b+\beta_p)/d}}$ be the value of k equating the order $\frac{k^{m-1}}{n^m}$ of $var[\hat{\psi}_{m,k}]$ to the order $k^{-2(\beta_b+\beta_p)/d}$ of TB_k^2. (Note k_* of Sec 3.3 is k_* (2)). Thus $var\left[\widehat{\psi}_{m,k_*(m)}\right]=n^{-\frac{2m(\beta_b+\beta_p)/d}{m-1+2(\beta_b+\beta_p)/d}}$. $\hat{\psi}_{m,k*(m)}$ has the optimal rate in the class $\{\hat{\psi}_{m,k}: k \in \mathcal{N}\}$ since $EB_m \asymp n^{-\left(\frac{\beta_b}{d+2\beta_b}+\frac{\beta_p}{d+2\beta_p}+\frac{(m-1)\beta_f}{d+2\beta_f}\right)}$ does not depend on k. This rate is

$$r(m) \equiv \max\left\{n^{-\left(\frac{\beta_b}{d+2\beta_b}+\frac{\beta_p}{d+2\beta_p}+\frac{(m-1)\beta_f}{d+2\beta_f}\right)}, n^{-\frac{m(\beta_b+\beta_p)/d}{m-1/2(\beta_b+\beta_p)/d}}\right\}$$

The optimal estimator in the class $\{\hat{\psi}_{m,k}: m = 2, \ldots; k \in \mathcal{N}\}$ is thus $\hat{\psi}_{m_{eff},,k*(m_{eff})}$ with m_{eff} the minimizer of $r(m)$. As discussed in Section 3.3, if condition (11) holds, then $m_{eff} = 2$, and $\hat{\psi}_{m_{eff},,k*(m_{eff})}$ attains the minimax convergence rate $n^{-\frac{2(\beta_b+\beta_p)}{d+2(\beta_b+\beta_p)}}$. If (11) fails to hold, $\hat{\psi}_{m_{eff},,k*(m_{eff})}$ will not be minimax (Robins et al. (2008)).

5. Confidence Interval Construction

In the regular case where (1) holds with a strict inequality and $\beta_f > \delta$, it follows from Lemma 8 that an honest asymptotic $1 - \alpha$ confidence interval for $\psi(\theta)$ whose width shrinks at rate $n^{-1/2}$ is the Wald interval $C_{m_{opt*},k_{eff}}=\widehat{\psi}_{m_{opt*},k_{eff}} \pm z_\alpha \widehat{se}\left(\widehat{\psi}_{m_{opt*},k_{eff}}\right)$. where

$$\widehat{se}\left(\widehat{\psi}_{m_{opt*},k_{eff}}\right)=n^{-1}\left\{\sum_{i=1}^{n}U_{1i}\left(\widehat{\theta}\right)^{2}\right\}^{1/2}$$

and z_{α} is the upper α–quantile of a N (0, 1) distribution.

Consider now the irregular case. A necessary condition for an $\hat{\psi}$ to center a honest Wald interval $C=\psi \pm z_{\alpha}\widehat{se}(\widehat{\psi})$ is that the order of its bias be less than that of the standard error. The estimator $\hat{\psi}_{m_{eff},,k*(m_{eff})}$ fails to satisfy this condition as its maximal estimation bias $EB_{m_{eff}}$ can dominate its standard error. However the condition is satisfied by the estimator $\hat{\psi}_{m_{eff},\tilde{k}(m_{eff})}$ with m_{eff} as above and $\tilde{k}$ (m_{eff}) equal to the k that equates the variance $\max\left(\frac{1}{n},\frac{k^{(m_{eff}-1)}}{n^{(m_{eff})}}\right)$ to $\{\log n\}\times\left\{\max\left[\{TB_{k}\}^{2},\{EB_{m_{eff}}\}^{2}\right]\right\}=\{\log n\}\max\left(k^{-2(\beta_b+\beta_p)/d},n^{-2\left(\frac{\beta_b}{d+2\beta_b}+\frac{\beta_p}{d+2\beta_p}+\frac{(m_{eff}-1)\beta_f}{d+2\beta_f}\right)}\right)$. The log n factor insures that the order of the standard error exceeds that of the bias. Furthermore $\hat{\psi}_{m_{eff},\tilde{k}(m_{eff})}$ converges at the same rate as the estimator $\hat{\psi}_{m_{eff},,k*(m_{eff})}$ up to a log factor.

In Theorem 10 of the web-supplement we show that, if an estimator $\hat{\psi}_{m,k}$ in our class has bias of lower order than the standard deviation, then, for $k \gg n$, $\left\{\frac{k^{m-1}}{n^{m}}\right\}^{-1/2}\left(\widehat{\psi}_{m,k}-\psi(\theta)\right)$ is conditionally (given the training sample) and unconditionally uniformly asymptotically normal with mean zero and variance that can be consistently estimated. It follows that $C_{m,k}=\widehat{\psi}_{m,k}\pm z_{\alpha}\widehat{se}\left(\widehat{\psi}_{m,k}\right)$ is an honest asymptotic $1-\alpha$ confidence interval for ψ (θ) whose width shrinks at rate $\left\{\frac{k^{m-1}}{n^{m}}\right\}^{1/2}$, where the formula for $\widehat{se}\left(\widehat{\psi}_{m,k}\right)$ is given in Theorem 10 of the web-supplement. Thus, the interval

$$C_{m_{eff},\tilde{k}(m_{eff})}=\widehat{\psi}_{m_{eff},\tilde{k}(m_{eff})}\pm z_{\alpha}\widehat{se}\left(\widehat{\psi}_{m_{eff},\tilde{k}(m_{eff})}\right)$$

shrinks as fast as any interval $C_{m,k}$ in our class.

6. Inference on $\tau(\theta)$

Recall from Section 1 that our ultimate functional of interest, $\tau(\theta) = E_{\theta}[cov_{\theta}(Y, A|X)]/E_{\theta}[var_{\theta}(A|X)]$, is the unique solution to the equation $\psi(\tau, \theta) = 0$ where $\psi(\tau, \theta) = E_{\theta}[\{Y(\tau) - b(X, \tau)\}\{A - p(X)\}]$ with $b(\tau): x \rightarrow b(x, \tau) \equiv E[Y(\tau) \mid X = x]$ and $Y(\tau) = Y - \tau A$. We assume it is $b(\tau)$ for $\tau = \tau(\theta)$ that is known to lie in the Hölder class of smoothness β_b rather than the function b.

Consider first the irregular case where condition 1 fails to hold. As discussed in Section 1, $\{\tau: 0 \in C_{m_{eff},\tilde{k}(m_{eff})}(\tau)\}$ is an honest asymptotic $1-\alpha$ confidence set for $\tau(\theta)$, where $C_{m,k}(\tau)$ and $\hat{\psi}_{m,k}(\tau)$ are $C_{m,k}$ and $\hat{\psi}_{m,k}$ with Y replaced by $Y(\tau)$. Furthermore, it follows from Theorem 6.1 of Robins et al. (2009b) that the width of the confidence set $\{\tau: 0 \in C_{m_{eff},\tilde{k}(m_{eff})}(\tau)\}$ or $\tau(\theta)$ shrinks with increasing n at the same rate $\left\{\frac{1}{n}\left(\frac{\tilde{k}(m_{eff})}{n}\right)^{m_{eff}-1}\right\}^{1/2}$ as does the confidence interval $C_{m_{eff},\tilde{k}(m_{eff})}(\tau)$ for $\psi(\tau, \theta)$. Finally, let $\hat{\tau}_{m_{eff},\tilde{k}(m_{eff})}$ be the solution to

$\hat{\psi}_{m_{eff},\tilde{k}(m_{eff})}(\tau) = 0$. Then, a Taylor expansion around $\tau(\theta)$, shows that $\left\{\frac{1}{n}\left(\frac{\tilde{k}(m_{eff})}{n}\right)^{m_{eff}-1}\right\}^{-1/2}\{\widehat{\tau} - \tau(\theta)\}$ is asymptotically normal with mean zero and a finite variance.

In the regular case where condition 1 holds and $\beta_f > \delta$, we conclude by a similar argument that $\hat{\tau}_{m_{opt*},k_{eff}}$ solving $\hat{\psi}_{m_{opt*},k_{eff}}(\tau) = 0$ is a semiparametric efficient estimator of $\tau(\theta)$ with influence function $\{\partial\psi(\tau,\theta)/\partial\tau\}^{-1}_{\tau=\tau(\theta)}U_1(\theta,\tau(\theta))$, where $U_1(\theta, \tau) = \{Y(\tau) - b(X, \tau)\}\{A - p(X)\} - \psi(\tau, \theta)$ is the efficient influence function of the functional $\psi(\tau, \theta)$.

7. Discussion

Although this paper breaks important new ground, many difficult issues remain. First, we have assumed the maximal possible roughness (as encoded in Hölder exponents and constants) of the nuisance functions *p*, *b*, and *f* to be known apriori. In practice, different subject matter experts will clearly disagree as to the maximal roughness; in addition, the actual smoothnesses of the nuisance functions cannot be empirically estimated. Thus it would be important to have methods that adapt to the unknown smoothness of these functions. However, for honest confidence intervals, the degree of possible adaption to unknown smoothness is small. Therefore an analyst needs to report a mapping from apriori smoothness assumptions encoded in Hölder exponents and constants (or in other measures of smoothness) to the associated $(1 - \alpha)$ honest confidence intervals proposed in this paper. Such a mapping is finally only useful if substantive experts can approximately quantify their informal opinions concerning the smoothness of *p*, *b*, and *f* using a measure of smoothness offered by the analyst. It is an open question which, if any, smoothness measure is suitable for this purpose.

In the irregular case, our results are for rates of convergence. We currently have few results on the constants in front of those rates.

Finally, a general software program to calculate our estimators must first construct a non-parametric d-dimensional density estimator $\hat{f}$ and then compute the k×k matrix $\{E_{\hat{f}}[\bar{z}_k(X)\,\bar{z}_k(X)^T]\}^{-1}$ by numerical integration followed by matrix inversion. As, in practice, k can easily be 500,000, we have yet to solve these computational challenges.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

Bhattacharya RN, Ghosh JK. A class of *U*-statistics and asymptotic normality of the number of *k*-clusters. Journal of Multivariate Analysis. 1992; 43:300–330.

Bickel, P.; Klaassen, C.; Ritov, Y.; Wellner, J. Efficient and adaptive estimation for semiparametric models. Springer Verlag; 1998.

Crump, RK.; Hotz, VJ.; Imbens, GW.; Mitnik, OA. Working Paper. National Bureau of Economic Research; 2006. Moving the Goalposts: Addressing Limited Overlap in the Estimation of Average Treatment Effects by Changing the Estimand; p. 330

Donald S, Newey W. Series estimation of semilinear models. Journal of Multivariate Analysis. 1994; 50:30–40.

Härdle, W.; Kerkyacharian, G.; Picard, D.; Tsybakov, A. Wavelets, approximation, and statistical applications. Springer; New York: 1998.

Robins, J. Optimal structural nested models for optimal sequential decisions. Proceedings of the Second Seattle Symposium in Biostatistics: Analysis of Correlated Data; Springer; 2004. p. 189

Robins, J.; Li, L.; Tchetgen, E.; van der Vaart, A. Working Paper. Department of Biostatistics, Harvard School of Public Health; 2007. Higher order influence functions and minimax estimation of nonlinear functionals.

Robins J, Li L, Tchetgen E, van der Vaart A. Higher order influence functions and minimax estimation of nonlinear functionals. IMS Lecture Notes–Monograph Series Probability and Statistics Models: Essays in Honor of David A. Freedman. 2008; 2:335–421.

Robins J, Li L, Tchetgen E, van der Vaart A. Quadratic semi-parametric Von Mises calculus. Metrika. 2009a; 69:227–247.

Robins J, Tchetgen E, Li L, van der Vaart A. Semiparametric minimax rates. Electronic Journal of Statistics. 2009b; 3:1305–1321.

Robins JM, Ritov Y. Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. Statistics in Medicine. 1997; 16:285–319. [PubMed: 9004398]

Robinson P. Root-N-consistent semiparametric regression. Econometrica: Journal of the Econometric Society. 1988; 56:931–954.

Van der Vaart, A.; Wellner, J. Weak convergence and empirical processes. Springer Verlag; 1996.