

Robust Inference with Higher Order Influence Functions: Part I

Eric Tchetgen*, Lingling Li*, Aad van der Vaart, James Robins*[†],

Departments of Biostatistics* and Epidemiology[†], Harvard University

Department of Mathematics Vrije Universiteit.

KEY WORDS Higher Order Influence functions, Missing at Random, nonparametric inference, U-statistics, Honest Confidence intervals, Weighted average treatment effect

1. Introduction

Suppose we obtain n i.i.d copies of a random vector O with unknown distribution F . Our goal is to construct honest $(1 - \alpha)$ asymptotic confidence intervals (whose width shrinks to zero with increasing n at the fastest possible rate) for a functional $\psi(F)$ in a model that places no restrictions on F , other than, perhaps, bounds on both the L_p norms and the roughness (more generally, the complexity) of certain density and conditional expectation functions. If $\psi(F)$ has a non-zero semiparametric information bound (SIB), it is known that if sufficiently stringent bounds on L_p norms and roughness are assumed, it is possible to construct honest asymptotic confidence intervals whose width shrinks at the usual parametric rate of $n^{-1/2}$. However, with the very high dimensional data collected in many applications, the roughness bounds necessary to obtain $n^{-1/2}$ rates may be considered substantively implausible (Robins and Ritov, 1997) and a new approach is then required. (All references are to be found at the end of our companion paper in this same volume.)

In this paper we implement a novel "inference machine" that takes as input apriori roughness bounds and a functional $\psi(F)$ with a finite semiparametric information bound and outputs honest asymptotic confidence intervals. We have not, as yet, been able to prove that our intervals shrink as fast as possible. However, we suspect that in many settings our shrinkage rates are indeed optimal and the estimator of $\psi(F)$ defined by the midpoint of our interval achieves the minimax rate of convergence, which may often be slower than the parametric rate of $n^{-1/2}$. Our "inference machine" can be extended to parameters $\psi(F)$ with a SIB of zero by approximating $\psi(F)$ by a sequence $\psi_n(F)$ of functionals with non-zero SIB. The particular functionals we consider in this paper are : (i) the weighted average treatment

effect (WATE) of a dichotomous treatment in the presence of high dimensional vector X of confounding factors and (ii) the marginal mean of a response Y , when Y is missing at random (MAR), and data are available on a high dimensional vector of always observed covariates X . We introduce novel confidence intervals for these two substantively important functionals that cover at their nominal level under weaker smoothness assumptions than previous interval estimators, often at the unavoidable price of shrinking to zero at a rate less than $n^{-1/2}$.

Our point and interval estimators are higher order U-statistics. They are derived with a new unified theory of parametric, semi-, and nonparametric statistics due to Robins and van der Vaart (Robins, 2004, Sec. 9) based on higher order scores (i.e., derivatives of the likelihood) and influence functions that applies equally to both the \sqrt{n} and non- \sqrt{n} problems. The theory reproduces the results previously obtained by the modern theory of non-parametric inference, produces many new non- \sqrt{n} results, and most importantly opens up the ability to perform optimal non- \sqrt{n} inference in complex high dimensional models. This theory of higher order influence functions extends the first order semiparametric theory of Bickel et al. (1993) and van der Vaart (1991) by incorporating the theory of higher order scores and Bhattacharyya bases considered by Pfanzagl (1990), McLeish and Small (1994) and Lindsay and Waterman (1996).

2. The Models

2.1 Weighted Average Treatment effect Model

We observe n i.i.d copies of $O_i = (W_i, A_i, X_i)$, $i = 1, \dots, n$, $O_i \sim F(O_i; \theta) \in \mathcal{M}^{WATE}(\Theta) = \{F(O; \theta); \theta \in \Theta\}$, where W_i is the outcome of interest, A_i is a dichotomous exposure, X_i is a d -dimensional vector of continuous covariates, treatment assignment is ignorable and the parameter θ indexes the laws included in model $\mathcal{M}^{WATE}(\Theta)$. The parameter of interest is the functional $\mu(\theta, c)$ given by:

$$\mu(\theta, c) = \frac{E_\theta [c(X)cov_\theta(W, A|X)]}{E_\theta [c(X)var_\theta(A|X)]} \quad (1)$$

where $c(\cdot)$ is a known function. We refer to $\mu(\theta, c)$ as a weighted average treatment effect since (1) can be rewritten :

$$E_\theta \left[d_{c,\theta}(X) \begin{pmatrix} E_\theta(W|A=1, X, T_\theta(X)=1) \\ -E_\theta(W|A=0, X, T_\theta(X)=1) \end{pmatrix} \right]$$

where $d_{c,\theta}(X) = \frac{T_\theta(X)c(X)\text{var}_\theta(A|X)}{E_\theta[c(X)\text{var}_\theta(A|X)]}$ and $T_\theta(X) = I(\text{var}_\theta(A|X) > 0)$. The particular choice $c(X) = \text{var}_\theta(A|X)^{-1}$ recovers the average treatment effect (ATE) functional. Under the semiparametric model

$$\tau(\theta) = E_\theta(W|A=1, X) - E_\theta(W|A=0, X) \quad (2)$$

for all X , $\mu(\theta, c) = \tau(\theta)$ for all $c(X)$. Therefore, when model (2) holds, valid inference on the nonparametric parameter (1) for any choice of $c(X)$ will also yield valid inference for $\tau(\theta)$. In what follows, we shall not assume that (2) holds and will only consider the case where $c \equiv 1$. If for any $\rho \in R$, we define $\psi_\rho(\theta)$ to be

$$E_\theta \{ \{Y(\rho) - E_\theta(Y(\rho)|X)\} \{A - E_\theta(A|X)\} \}$$

with $Y(\rho) = W - \rho A$, it is easy to verify that $\mu(\theta, 1)$ may also be characterized as the $\rho(\theta)$ satisfying $\psi_{\rho(\theta)}(\theta) = 0$. Thus inference on $\mu(\theta, 1)$ is easily obtained from inference on $\psi_\rho(\theta)$. In particular a $(1 - \alpha)$ confidence set for $\mu(\theta, 1)$ is the set of ρ such that a $(1 - \alpha)$ CI for $\psi_\rho(\theta)$ contains 0. Therefore, with no loss of generality, we consider the construction of a $(1 - \alpha)$ CI for $\psi^{WATE}(\theta) = \psi_{\tilde{\rho}}(\theta)$ for a fixed value $\rho = \tilde{\rho}$, and write $Y(\tilde{\rho}) = Y, \cdot$

2.2 Missing At Random Model

We observe *i.i.d* data $(A_i Y_i, A_i, X_i) = O_i \sim F(O_i; \theta) \in \mathcal{M}^{MAR}(\Theta) = \{F(O; \Theta), \theta \in \Theta\}$, where Y_i is an outcome that is not always observed, A_i is the missingness indicator, X_i is a d -dimensional vector of always observed continuous covariates, and $P_\theta(A=1|X) > \delta$ a.e. for some $\delta > 0$. The MAR assumption implies that $P_\theta(A=1|X, Y) = P_\theta(A=1|X)$. Our goal here is to make inference on the functional $\psi^{MAR}(\theta)$ given by:

$$\begin{aligned} \psi^{MAR}(\theta) &= E_\theta \left(\frac{AY}{P_\theta(A=1|X)} \right) \\ &= E_\theta(E_\theta(Y|A=1, X)) \end{aligned} \quad (3)$$

Under MAR, ψ^{MAR} is the marginal mean of Y in the absence of missing data.

2.3 Formalization of The Models

In both models, we assume for all $\theta \in \Theta$, the distribution of \mathbf{X} is supported on the unit cube $[0, 1]^d$ in R^d and has a density with respect to Lebesgue measure. To complete the specification of models \mathcal{M}^{WATE} and \mathcal{M}^{MAR} , we take functions $b = b(\cdot)$, $\omega = \omega(\cdot)$, $g = g(\cdot)$ to be elements of θ , such that in both models $\omega(X)$ denotes $P(A=1|X; \theta)$, $b(X)$ denotes $E(Y|X; \theta)$ and $E(Y|X, A=1; \theta)$ in models \mathcal{M}^{WATE} and \mathcal{M}^{MAR} respectively, and $g(X)$ denotes the marginal density of X and the conditional density of X given $A=1$ in models \mathcal{M}^{WATE} and \mathcal{M}^{MAR} respectively. The model specific definitions of b and g allows a more unified treatment of the two models. We impose the following known bounds on various L_∞ norms in both models : (a.1) $|b(X)| \leq K$ w.p.1 for some constant $K < \infty$, (a.2) $\text{var}(Y|X) \leq K_1$ w.p.1 for some constant $K_1 < \infty$, and (a.3) there exist $\delta^*, M > 0$, such that $\delta^* < g(X) < M$ w.p.1.

Results in Ritov and Bickel (1990) and Robins and Ritov (1997) imply it is not possible to construct honest asymptotic confidence intervals for ψ^{WATE} whose width shrinks to 0 as $n \rightarrow \infty$ without bounds on the roughness of $b(\cdot)$ and $\omega(\cdot)$. Our bounds will be based on the following definition.

Definition 1 A function $h(X)$ with support $\text{supp}(X)$ is said to belong to a Holder ball $H(\beta_h, C_h)$, with Holder exponent $\beta_h > 0$ and radius $C_h > 0$, if and only if all partial derivatives of $h(X)$ up to order $\lfloor \beta_h \rfloor$ exist, and all partial derivatives $\nabla^{\lfloor \beta_h \rfloor}$ of order $\lfloor \beta_h \rfloor$ satisfy

$$\begin{aligned} & \sup_{X \in \text{supp}(X)} \left| \begin{array}{c} \nabla^{\lfloor \beta_h \rfloor} h(X + \delta X) \\ - \nabla^{\lfloor \beta_h \rfloor} h(X) \end{array} \right| \\ & \leq C_h \|\delta X\|^{\beta_h - \lfloor \beta_h \rfloor} \end{aligned}$$

We note that the integrated mean squared error (MSE) and the uniform error minimax rates of convergence for estimation of a marginal density or conditional expectation $h(\cdot) \in H(\beta_h, C_h)$ are $O\left(n^{-\frac{\beta_h}{2\beta_h+d}}\right)$ and $O\left(\left(\frac{n}{\log n}\right)^{-\frac{\beta_h}{2\beta_h+d}}\right)$. We

assume $b(\cdot)$, $\omega(\cdot)$, and $g(\cdot)$ lie in given Holder balls $H(\beta_b, C_b)$, $H(\beta_\omega, C_\omega)$, $H(\beta_g, C_g)$.

Remark 2 In model \mathcal{M}^{MAR} , if one had assumed a priori that $\omega(\cdot)$ and the marginal density of X lay in Holder balls with respective exponents β_ω and β , then β_g would

be $\min(\beta_\omega, \beta)$, since $f(A=1|X)f(X) = f(X|A=1)P(A=1)$.

2.4 First Order Inference

In this section, we consider the previous approaches to the construction of conservative uniform (i.e. honest) asymptotic confidence sets where

Definition 3 \mathcal{C}_n is a conservative uniform asymptotic $(1 - \alpha)$ confidence set for ψ if:

$$\liminf_n \inf_\theta \{\Pr_\theta [\psi \in \mathcal{C}_n] - (1 - \alpha)\} \geq 0 \quad (4)$$

A standard interval estimator is the Wald type interval centered around an asymptotic linear (therefore asymptotically normal) plug-in estimator $\hat{\psi} = \psi(\hat{\theta})$, i.e.:

$$\mathcal{C}_n = \hat{\psi} \pm z_{\alpha/2} \widehat{s.e.}(\hat{\psi}) \quad (5)$$

where $z_{\alpha/2}$ is the $\alpha/2$ quantile of the standard normal, $\hat{\theta}$ is a preliminary estimate of θ , and $\widehat{s.e.}(\hat{\psi})$ is a (say, nonparametric bootstrap) estimator of $\{var(\hat{\psi})\}^{1/2}$. For instance, in model \mathcal{M}^{MAR} , the plug-in estimator

$$\hat{\psi}^{MAR} \equiv \psi^{MAR}(\hat{\theta}) = \mathbb{V}_n[\hat{b}(X)],$$

where $\hat{b}(\cdot)$ is a preliminary nonparametric estimate of $b(\cdot)$, and for any function $V_{p;i_1, \dots, i_p} = v_p(O_{i_1}, O_{i_2}, \dots, O_{i_p})$ of p subjects' data, $\mathbb{V}_n v_p$ is the p th order U-statistic

$$\frac{(n-p)!}{n!} \sum_{i_1 \neq i_2 \dots \neq i_p} V_{p;i_1, \dots, i_p}$$

The bias of $\hat{\psi}^{MAR}$,

$$E_\theta(\hat{\psi}^{MAR}) - \psi^{MAR} = E_\theta(\hat{b}(X_i) - b(X_i)),$$

is "first order" and therefore the interval centered around it will not be a valid asymptotic confidence interval unless $\hat{b}(\cdot)$ is chosen such that $E(\hat{b}(X_i) - b(X_i)) = o(n^{-1/2})$, as bias must be of smaller order than standard deviation $O(n^{-1/2})$ for \mathcal{C}_n to satisfy (4). This prohibits many minimax rate optimal estimators of $b(\cdot)$, whose expected bias does not converge at the usual parametric rate. In such a case,

an "undersmoothed" estimator $\hat{b}(\cdot)$ may have smaller bias. However, a more general solution is to center our interval around the one step update:

$$\begin{aligned} \hat{\psi}_1^{MAR} &= \hat{\psi}^{MAR} + U_{1, \psi^{MAR}}(\hat{\theta}) \\ &= \mathbb{V}_n \left[H_i^{MAR}(\hat{\theta}) \right] \end{aligned}$$

where $U_{1, \psi^{MAR}}(\theta) = \mathbb{V}_n[V_{1, \psi^{MAR}; i}(\theta)]$ is the first order influence function of $\psi^{MAR}(\theta)$, $V_{1, \psi^{MAR}; i}(\theta) = H_i^{MAR}(\theta) - \psi^{MAR}(\theta)$ and

$$H_i^{MAR}(\theta) = \frac{R_i}{\omega(X_i)}(Y - b(X_i)) + b(X_i)$$

See the next sections for a definition of influence functions. The bias of $\hat{\psi}_1^{MAR}$

$$E_\theta \left[\left(\frac{\omega(X_i)}{\widehat{\omega}(X_i)} - 1 \right) (b(X_i) - \hat{b}(X_i)) \right]$$

is of "second order" and thus less than that of the plug-in (Newey et al., 1991).

Consider now the model \mathcal{M}^{WATE} . Define $\epsilon_i = Y_i - b(X_i)$, $\Delta_i = A_i - \omega(X_i)$. Then $U_{1, \psi^{WATE}}(\theta) = \mathbb{V}_n[V_{1, \psi^{WATE}; i}(\theta)] = \mathbb{V}_n[\epsilon_i \Delta_i - \psi^{WATE}(\theta)]$ is the first order influence function of $\psi^{WATE}(\theta)$ and $\hat{\psi}_1^{WATE} = \hat{\psi}^{WATE} + U_{1, \psi^{WATE}}(\hat{\theta}) = \mathbb{V}_n \hat{\epsilon}_i \hat{\Delta}_i$ where $\hat{\epsilon}_i = Y_i - \hat{b}(X_i)$, $\hat{\Delta}_i = A_i - \widehat{\omega}(X_i)$. The bias of $\hat{\psi}_1^{WATE}$ is $E_\theta \left[\left(b(X_i) - \hat{b}(X_i) \right) (\omega(X_i) - \widehat{\omega}(X_i)) \right]$.

The analysis of $\hat{\psi}_1$ in both models is simplified if the functions $\omega(X_i)$ and $b(X_i)$ are estimated from a separate independent training sample (TS). We assume that our actual sample size is N and we randomly divide the N observations into two samples: an analysis sample of size n and a training sample of size $N - n$ where $(N - n)/n = f$, $1 > f > 0$. We obtain our estimates $\hat{b} \equiv \hat{b}(\cdot) = \hat{b}(\cdot; O_{n+1}, \dots, O_N)$, $\widehat{\omega}$ and \hat{g} from the training sample. Sample splitting has no effect on optimal rates of convergence, although in the simplest form described here does affect 'constants'. With care, we believe sample splitting can be avoided, in part by using techniques similar to those of Bickel and Ritov (1988). Now, conditional on the TS, $\hat{\psi}_1$ is a sample average of i.i.d random variables and thus will have variance of $O(n^{-1})$. When $\widehat{\omega}(\cdot)$ and $\hat{b}(\cdot)$ are chosen to be rate optimal, $\hat{\psi}_1$ will have conditional bias $O_p \left(n^{-\left(\frac{\beta_b}{2\beta_b + d} + \frac{\beta_\omega}{2\beta_\omega + d} \right)} \right)$

which is $o(n^{-1/2})$ if, for example, $\frac{\beta_w}{d} = \frac{\beta_b}{d} > 1/2$, but not if $\frac{\beta_w}{d} = \frac{\beta_b}{d} \leq 1/2$. It follows that CIs centered on $\hat{\psi}_1$ will be asymptotically honest when $\frac{\beta_w}{d} = \frac{\beta_b}{d} > 1/2$ but not when $\frac{\beta_w}{d} = \frac{\beta_b}{d} \leq 1/2$. To handle the latter case we introduce higher order influence functions.

3. Higher Order Inference

3.1 Theory of Higher Order Influence Functions

Given n i.i.d observations $\mathbf{O} = \mathbf{O}_n = \{O_i, i = 1, \dots, n\}$ from a model $\mathcal{M}(\Theta) = \{F(O; \theta), \theta \in \Theta\}$, we consider inference on a functional $\psi(\theta)$. In general, $\psi(\theta)$ can be infinite dimensional but here we only consider the one dimensional case. In the following all quantities can depend on the sample size n , including the support of O , the parameter space Θ , and the functional $\psi(\theta)$. We suppress the dependence on n in the notation.

Given function $g(\zeta), \zeta = \{\zeta_1, \dots, \zeta_r\}^T$, define for $m = 0, 1, 2, \dots$, $g_{\setminus i_1, \dots, i_m}(\zeta) = \frac{\partial^m g(\zeta)}{\partial \zeta_{i_1} \dots \partial \zeta_{i_m}}$ with $i_s \in \{1, \dots, r\}$, for $s = 1, 2, \dots, m$ where the \setminus symbol denotes differentiation by the variables occurring to its right. Given a sufficiently smooth r -dimensional parametric submodel $\tilde{\theta}(\zeta)$ mapping $\zeta \in R^r$ injectively into Θ , define $\psi_{\setminus i_1, \dots, i_m}(\theta)$ to be $(\psi \circ \tilde{\theta})_{\setminus i_1, \dots, i_m}(\zeta)|_{\zeta = \tilde{\theta}^{-1}\{\theta\}}$ and $f_{\setminus i_1, \dots, i_m}(\theta)$ to be $(f \circ \tilde{\theta})_{\setminus i_1, \dots, i_m}(\zeta)|_{\zeta = \tilde{\theta}^{-1}\{\theta\}}$ where $f(\mathbf{O}; \theta) \equiv \prod_i f(O_i; \theta)$.

Definition 4 A U -statistic $U_p(\theta) = u_p(\mathbf{O}_n; \theta) = \mathbb{V}_n [V_{p; i_1, \dots, i_p}(\theta)]$ of order p and finite variance is said to be a p th order estimation influence function for $\psi(\theta)$ if (i) $E_\theta [U_p(\theta)] = 0$, $\theta \in \Theta$ and (ii) for $m = 1, 2, \dots, p$ and every suitably smooth r dimensional parametric submodel $\tilde{\theta}(\zeta), r = 1, 2, \dots$,

$$\psi_{\setminus i_1, \dots, i_m}(\theta) = E_\theta [U_p(\theta) S_{i_1, \dots, i_m}(\theta)]$$

where $S_{i_1, \dots, i_m}(\theta) \equiv f_{\setminus i_1, \dots, i_m}(\mathbf{O}, \theta) / f(\mathbf{O}, \theta)$ is a U -statistic of order m . We then say that $\psi(\theta)$ is higher order pathwise differentiable and refer to $S_{i_1, \dots, i_m}(\theta)$ as an m th order score associated with the submodel $\tilde{\theta}(\zeta)$.

Estimation influence functions are useful for deriving point estimators of ψ with small

bias and for deriving interval estimators centered on an estimate of ψ . A key result is the following theorem which is related to results of McLeish and Small (1994)

Theorem 5 :Extended Information Equality Theorem: Given a p th order influence function $U_p(\theta)$, for any smooth submodels $\tilde{\theta}(\zeta)$, $\partial^s E_\theta [U_p(\zeta)] / \partial \zeta_{i_1} \dots \partial \zeta_{i_s} |_{\zeta = \tilde{\theta}^{-1}\{\theta\}} = -\psi_{\setminus i_1, \dots, i_s}(\theta)$. Thus, if the functions $E_\theta [U_p(\theta^*)]$ and $-\psi(\theta^*) - \psi(\theta)$ are Fréchet differentiable w.r.t. θ^* to order $p + 1$ for a norm $\|\cdot\|$,

$$E_\theta [U_p(\theta + \delta\theta)] = -[\psi(\theta + \delta\theta) - \psi(\theta)] + O(\|\delta\theta\|^{p+1})$$

since the functions $E_\theta [U_p(\theta^*)]$ and $-\psi(\theta^*) - \psi(\theta)$ of θ^* have the same Taylor expansion around θ up to order p

Proof. See Robins and van der Vaart (2004). ■

Let $\mathcal{U}_p(\theta)$ be the Hilbert space of all U -statistics of order p with mean zero and finite variance with inner product defined by covariances with respect to the n -fold product measure $F^n(\cdot; \theta)$. Define the p th order tangent space for the model $\mathcal{M}(\Theta)$, $\overline{\mathcal{B}}_p(\theta)$, to be the subspace of $\mathcal{U}_p(\theta)$ formed by taking the closed linear span of all m^{th} order scores, $m = 1, \dots, p$, at θ , as we vary over all regular parametric submodels $\tilde{\theta}(\zeta)$ of our model $\mathcal{M}(\Theta)$. We say a model is (locally) nonparametric for p th order inference if $\overline{\mathcal{B}}_p(\theta) = \mathcal{U}_p(\theta)$.

Our models $\mathcal{M}^{WATE}(\Theta)$ and $\mathcal{M}^{MAR}(\Theta)$ can be shown to be (locally) nonparametric.

Theorem 6 If the model $\mathcal{M}(\Theta)$ is (locally) nonparametric, then (i) there is at most one p th order estimation influence function $U_p(\theta)$ for $\psi(\theta)$. (ii) for $p > 1$, a) If $U_p(\theta)$ exists then $U_{p-1}(\theta)$ exists, $U_p(\theta) = U_{p-1}(\theta) + U_{p,p}(\theta)$ where $U_{p,p}(\theta)$ is a degenerate p th order U -statistic (i.e., $U_{p,p}(\theta) = \mathbb{V}_n [V_{p,p; i_1, \dots, i_p}(\theta)]$ where $E [v_{p,p}(o_{i_1}, \dots, o_{i_p}; \theta)] = 0$ for $l = 1, \dots, p$), $E_\theta [U_{p-1}(\theta) U_{p,p}(\theta)] = 0$, and $\text{var}_\theta [U_{p-1}(\theta)] \leq \text{var}_\theta [U_p(\theta)]$, b) Suppose the $(p-1)$ th order estimation influence function $U_{p-1}(\theta)$ exists. Then the p th order influence function $U_p(\theta)$ exists if and only the functional $v_{p-1, p-1}(o_{i_1}, \dots, o_{i_{p-1}}; \theta)$ has a first order influence function for all $o_{i_1}, \dots, o_{i_{p-1}}$

in a set \mathcal{O}_{p-1} which has probability 1 under $F^{(p-1)}(\cdot, \theta)$. If $U_p(\theta)$ exists then, letting $n^{-1} \sum_{i_p=1}^n [v_{1, v_{p-1, p-1}}(o_{i_1}, \dots, o_{i_{p-1}}, O_{i_p}; \theta)]$ be the first order influence function of the functional $v_{p-1, p-1}(o_{i_1}, \dots, o_{i_{p-1}}; \theta)$, $V_{p, p; i_1, \dots, i_p}(\theta)$ is the degenerate kernel obtained by subtracting from the p th order U -statistic with kernel $p^{-1} v_{1, v_{p-1, p-1}}(O_{i_1}, \dots, O_{i_{p-1}}, O_{i_p}; \theta)$ its projection on all U -statistics of order $p-1$.

If one knows how to calculate first order influence functions, then one can obtain $U_1(\theta), \dots, U_p(\theta)$ recursively using theorem 6b.

We now describe the main heuristic idea behind using higher order influence functions in models in which they exist. Consider the estimator $\hat{\psi}_p = \psi(\hat{\theta}) + U_p(\hat{\theta})$ based on a sample size n , where $\hat{\theta}$ is an initial estimator of θ from a separate training sample that obtains the optimal rate of convergence for θ . Here $U_p(\theta)$ is a p th order estimation influence function. Expanding and evaluating conditionally on the training sample, we find by Theorem 5 that the conditional bias $E_\theta [\hat{\psi}_p - \psi(\theta) | \hat{\theta}] = \psi(\hat{\theta}) - \psi(\theta) + E_\theta [U_p(\hat{\theta}) | \hat{\theta}]$ is $O_p(\|\hat{\theta} - \theta\|^{p+1})$ which decreases with p if, as we assume, $\|\hat{\theta} - \theta\| = O_p(n^{-\delta})$ for some $\delta > 0$.

Under weak conditions we would expect that $var_\theta \{U_p(\hat{\theta}) - U_p(\theta) | \hat{\theta}\} / var [U_p(\theta)] = O_p(\|\hat{\theta} - \theta\|)$ so $var_\theta [U_p(\hat{\theta}) | \hat{\theta}] = var_\theta [U_p(\theta)] (1 + O_p(\|\hat{\theta} - \theta\|))$. Now, by Theorem 6a, $var_\theta [U_p(\theta)]$ increases with p . Thus, the asymptotic mean squared error among the candidate estimators $\hat{\psi}_p$ of $\psi(\theta)$ is minimized at $p = p_{opt}$ for which the squared bias of $O_p(\|\hat{\theta} - \theta\|^{2(p+1)})$ and the variance $var_\theta [U_p(\theta)]$ are of the same order. By choosing p^* slightly greater than p_{opt} we could guarantee that, asymptotically, variance dominates bias and construct honest asymptotic confidence intervals centered at $\hat{\psi}_{p^*}$.

Unfortunately in models $\mathcal{M}^{WATE}(\Theta)$ and $\mathcal{M}^{MAR}(\Theta)$ respectively, it follows from Theorem 6b and the dependence of the first order influence functions $U_{1, \psi^{MAR}}(\theta)$ and $U_{1, \psi^{WATE}}(\theta)$ on the infinite dimensional parameters $b(\cdot)$ and $\omega(\cdot)$, that higher order influence functions for $\psi^{WATE}(\theta)$ and $\psi^{MAR}(\theta)$ do not exist.

To overcome this difficulty we proceed as follows. In both models, we introduce a new parameter $\bar{\psi}_m(\theta)$ that (i) depends on the sample size through a positive integer index $m = m(n)$ (which we refer to as the truncation index), (ii) has influence functions $U_{p, \bar{\psi}_m}(\theta)$ of all orders p , (iii) equals $\psi(\theta)$ on a large subset $\Theta_{sub, m}$ of Θ with $\hat{\theta} \in \Theta_{sub, m}$, which guarantees that the plug-ins $\psi(\hat{\theta})$ and $\bar{\psi}_m(\hat{\theta})$ are equal.

We then define the estimator $\hat{\psi}_{p, m} \equiv \psi(\hat{\theta}) + U_{p, \bar{\psi}_m}(\hat{\theta})$. The conditional bias $E[\hat{\psi}_{p, m} | \hat{\theta}] - \psi(\theta)$ of $\hat{\psi}_{p, m}$ is $TB_m(\theta) + EB_p(\theta)$ where the $TB_m(\theta) = \bar{\psi}_m(\theta) - \psi(\theta)$ is zero on $\Theta_{sub, m}$ and does not depend on p and the estimation bias $EB_p(\theta) = E[\hat{\psi}_{p, m} | \hat{\theta}] - \bar{\psi}_m(\theta)$ is $O_p(\|\hat{\theta} - \theta\|^{p+1})$ by the above argument and in many models including $\mathcal{M}^{WATE}(\Theta)$ and $\mathcal{M}^{MAR}(\Theta)$ the order can be shown not to depend further on m . Under weak conditions, the conditional variance of $\hat{\psi}_{p, m}$ is of the order of $var_\theta [U_{p, \bar{\psi}_m}(\theta)]$

The rate of convergence of $\hat{\psi}_{p, m}$ can depend on the choice of $\bar{\psi}_m(\theta)$. Nevertheless, many different choices of $\bar{\psi}_m(\theta)$ result in $\hat{\psi}_{p, m}$ that achieve what we conjecture to be the optimal rate. The choice among such "optimal" $\bar{\psi}_m(\theta)$ can then be based on computational considerations. For a class of models that includes both $\mathcal{M}^{WATE}(\Theta)$ and $\mathcal{M}^{MAR}(\Theta)$, we have developed a general method for deriving "optimal" $\bar{\psi}_m(\theta)$ that minimize the computational complexity of $\hat{\psi}_{p, m}$. To do so requires choosing $\bar{\psi}_m(\theta)$ such that the first order influence functions of $\bar{\psi}_m(\theta)$ and $\psi(\theta)$ are equal at $\hat{\theta}$, i.e., $U_{1, \bar{\psi}_m}(\hat{\theta}) = U_{1, \psi}(\hat{\theta})$ w.p.1. We applied this methodology to derive the functionals $\bar{\psi}_m^{WATE}(\theta)$ and $\bar{\psi}_m^{MAR}(\theta)$ given below. Due to space limitations, we do not describe the methodology here. We define $\bar{\psi}_m^{WATE}(\theta) = \psi^{WATE}(\theta) + TB_m^{WATE}(\theta)$ and $\bar{\psi}_m^{MAR}(\theta) = \psi^{MAR}(\theta) + TB_m^{MAR}(\theta)$ where $TB_m^{WATE}(\theta) = E_\theta \left[\left\{ b(X) - \bar{b}_m^{WATE}(X) \right\} \left\{ \omega(X) - \bar{\omega}_m^{WATE}(X) \right\} \right]$ and $TB_m^{MAR}(\theta) = E_\theta \left[\left\{ b(X) - \bar{b}_m^{MAR}(X) \right\} \left\{ 1 - \frac{\omega(X)}{\bar{\omega}_m^{MAR}(X)} \right\} \right]$

with

$$\begin{aligned}
& \bar{\omega}_m^{WATE}(X) = \hat{\omega}(X) \\
& + E_\theta [K_m^g(X, X^*) \{\omega(X^*) - \hat{\omega}(X^*)\} | X] \\
& \{\bar{\omega}_m^{MAR}(X)\}^{-1} = \{\hat{\omega}(X)\}^{-1} \\
& + E_\theta \left[K_m^g(X, X^*) \left\{ \begin{array}{l} \{\omega(X^*)\}^{-1} \\ -\{\hat{\omega}(X^*)\}^{-1} \end{array} \right\} | X \right] \\
& \bar{b}_m(X) = \hat{b}(X) \\
& + E_\theta \left[K_m^g(X, X^*) \left\{ b(X^*) - \hat{b}(X^*) \right\} | X \right]
\end{aligned}$$

where we use the model specific definitions of g given in Sec. 2.3, so that $\bar{b}_m^{MAR}(X) \neq \bar{b}_m^{WATE}(X)$. Here $K_m^g(\cdot, \cdot) = K_{m(n)}^g(\cdot, \cdot)$ is a kernel which satisfies the following conditions : (a.4) there is an m -dimensional subspace $L_m \subset L_2(g)$ such that the operator $\mathbf{K}_m^g(\cdot):L_2(g) \rightarrow L_m$, $(\mathbf{K}_m^g h(\cdot))(x) \equiv \int K_m^g(y, x) h(y)g(y)dy$ equals $h(x)$ for all $h \in L_m$, $(\mathbf{K}_m^g h(\cdot))(x) = 0$ for all $h^\perp \in L_m^\perp$, where L_m^\perp is the orthogonal complement of L_m in $L_2(g)$, and for any (x, y) , $g_1(\cdot)$ and $g_2(\cdot)$, $(\mathbf{K}_m^{g_1} K_m^{g_2}(\cdot, x))(y) = K_m^{g_2}(y, x)$, (a.5) For any $h(\cdot) \in H(\beta_h, C_h)$ with $\beta_h \leq \max(\beta_\omega, \beta_b)$; $\int (h(x) - \mathbf{K}_m^g h(x))^2 g(x)dx = O(m^{-\frac{2\beta_h}{d}})$, (a.6) $\| \prod_{s=1}^J K_m^{\hat{g}}(X_{i_s}, X_{i_{s+1}}) \|_g^2 = O(m^J)$ where $\| \cdot \|_g$ is the $L_2(g)$ - norm and $J \geq 1$, (a.7) for all $x_1, x_2 \in [0, 1]^d$, $K_m^g(x_1, x_2) = K_m^{\hat{g}}(x_1, x_2) (1 + O(\|g - \hat{g}\|_\infty))$ where $\| \cdot \|_\infty$ is the L_∞ - norm.

The set $\Theta_{sub,m}^{WATE}$ on which $\bar{\psi}_m^{WATE}(\theta) = \psi^{WATE}(\theta)$ is precisely the set of θ for which either $b(x) - \hat{b}(x)$ or $\omega(x) - \hat{\omega}(x)$ is in the m -dimensional space L_m . The set $\Theta_{sub,m}^{MAR}$ on which $\bar{\psi}_m^{MAR}(\theta) = \psi^{MAR}(\theta)$ is the set of θ for which either $b(x) - \hat{b}(x)$ or $\{\omega(x)\}^{-1} - \{\hat{\omega}(x)\}^{-1}$ is in the m -dimensional space L_m . Condition (a.4) makes $K_m^g(\cdot, \cdot)$ an orthogonal projection kernel onto L_m . In the Appendix of our companion paper (Li. et al., 2005) we show the following. In both models, Condition (a.5) ensures that the projection has small approximation error and that

$$TB_m \equiv \sup_{\theta \in \Theta} TB_m(\theta) = O\left(m^{-\frac{\beta_b + \beta_\omega}{d}}\right).$$

Condition (a.6) insures that for all $\theta \in \Theta$

$$var_\theta \left[U_{p, \bar{\psi}_m}(\theta) \right] = O\left(\frac{1}{n} \max\left(1, \left(\frac{m}{n}\right)^{p-1}\right)\right)$$

Finally condition (a.7) ensures that the kernel can be well estimated pointwise. We show how one can construct kernels that satisfy (a.4) – (a.7) below.

In the Appendix of our companion paper we show that, in both models, for all $\theta \in \Theta$

$$EB_p(\theta) = O_P\left(\left(\frac{\log n}{n}\right)^{\frac{(p-1)\beta_g}{d+2\beta_g}} n^{-\left(\frac{\beta_b}{d+2\beta_b} + \frac{\beta_\omega}{d+2\beta_\omega}\right)}\right)$$

A more detailed analysis shows that the $\log n$ factor can be removed. For a given p , the estimator $\psi_{p, m_{opt}(p)}(\hat{\theta})$ that has minimum asymptotic MSE among candidates $\psi_{p, m}(\hat{\theta})$ uses the value $m_{opt}(p) = m_{opt}(p, n)$ of m that equates the order $O\left(\frac{1}{n} \max\left(1, \left(\frac{m}{n}\right)^{p-1}\right)\right)$ of $var_\theta \left[U_{p, \bar{\psi}_{m(n)}}(\theta) \right]$ to the order

$$\begin{aligned}
& \max \left[\{TB_{m(n)}\}^2, \{EB_p(\theta)\}^2 \right] = \\
& \max \left[\left(\frac{\log n}{n}\right)^{\frac{2(p-1)\beta_g}{d+2\beta_g}} n^{-\left(\frac{2\beta_b}{d+2\beta_b} + \frac{2\beta_\omega}{d+2\beta_\omega}\right)}, \right. \\
& \left. m^{-\frac{2(\beta_b + \beta_\omega)}{d}} \right]
\end{aligned}$$

of the maximal squared bias. The estimator $\psi_{p_{opt}, m_{opt}(p_{opt})}(\hat{\theta})$ that has minimum asymptotic MSE among all candidates $\psi_{p, m}(\hat{\theta})$ is the estimator $\psi_{p, m_{opt}(p)}(\hat{\theta})$ which minimizes $O\left(\frac{1}{n} \max\left(1, \left(\frac{m_{opt}(p, n)}{n}\right)^{p-1}\right)\right)$.

To illustrate these results, consider the WATE model. Suppose that $\frac{\beta_\omega}{d} = \frac{\beta_b}{d} = \frac{\beta_g}{d} = 3/10$. Then $EB_1(\theta) = O(n^{-3/8}) > O(n^{-1/2})$ so $\hat{\psi}_1$ fails to be $n^{1/2}$ - consistent. However, $\hat{\psi}_{p, m}$ is $n^{1/2}$ - consistent for all $p > 2$ if $m = n$. If, $\frac{\beta_\omega}{d} = \frac{\beta_b}{d} = 1/4$, and $\frac{\beta_g}{d} = 1/5$; then, again, $\hat{\psi}_1$ fails to be $n^{1/2}$ - consistent. Now, however, $\hat{\psi}_{2, m_{opt}(2, n)} = \hat{\psi}_{2, n^{21/20}}$ (and thus $\hat{\psi}_{2, m}$ for any m) also fail to be $n^{1/2}$ - consistent. However $\hat{\psi}_{p, m}$ is $n^{1/2}$ - consistent if $p \geq 3$ and $m = n$. In general, whenever $\frac{\beta_\omega}{d} = \frac{\beta_b}{d} = 1/4$ and $\frac{\beta_g}{d} > 0$, $\hat{\psi}_{p, m}$ will fail to be $n^{1/2}$ - consistent for $p < 1 + 1/6 \left(1 + 2\frac{\beta_g}{d}\right) / \frac{\beta_g}{d}$ whatever be m , while $\hat{\psi}_{p, m}$ will be $n^{1/2}$ - consistent for $p \geq 1 + 1/6 \left(1 + 2\frac{\beta_g}{d}\right) / \frac{\beta_g}{d}$ and $m = n$. When $\frac{\beta_\omega}{d} = \frac{\beta_b}{d} < 1/4$, $m_{opt}(p, n)$ will increase faster

than n for all $p \geq 2$; therefore $\widehat{\psi}_{p_{opt}, m_{opt}(p_{opt}, n)}$ (and thus $\widehat{\psi}_{p, m}$ for any p and m) will fail to be $n^{1/2}$ -consistent and thus usual parametric rates are not achievable. For example suppose $\frac{\beta_\omega}{d} = \frac{\beta_b}{d} = 1/8$ and $\frac{\beta_g}{d} = 1/5$. Here $p_{opt} = 3$ with $m_{opt}(3, n) = n^{1+2/5}$ so $\widehat{\psi}_{3, m=(n^{1+2/5})}$ is the optimal-rate estimator in our class. Results analogous to the above have been previously reported in the literature for the "integral of the square and of the cube of a density" functionals (Bickel and Ritov, 1988; Kerkycharian and Picard, 1996). However these problems are simpler in the sense that estimation bias is not an issue, so one only has to balance variance with truncation bias.

Our companion paper presents simulation results to determine whether our asymptotic results are relevant for finite samples.

3.2 Higher Order Estimated Influence Functions in the WATE and MAR models and Construction of Confidence Intervals

In this section, we provide explicit formula for $\widehat{\psi}_{p, m} \equiv \psi(\widehat{\theta}) + U_{p, \overline{\psi}_m}(\widehat{\theta})$. By Theorem 6, $U_{p, \overline{\psi}_m}(\widehat{\theta}) = U_{1, \overline{\psi}_m}(\widehat{\theta}) + \sum_{j=2}^p U_{j, j, \overline{\psi}_m}(\widehat{\theta})$, with $U_{j, j, \overline{\psi}_m}(\theta) = \mathbb{V}_n \left[V_{j, j, \overline{\psi}_m; i_1, \dots, i_j}(\theta) \right]$.

Now $\psi(\widehat{\theta})$ and $U_{1, \overline{\psi}_m}(\widehat{\theta}) = U_{1, \psi}(\widehat{\theta})$ are defined in Sec 2.4 for both models. Next for $p \geq 2$,

$$\begin{aligned} & V_{p, p, \overline{\psi}_m^{WATE}; i_1, \dots, i_p}(\widehat{\theta}) \\ &= \widehat{\Delta}_{i_1} \times \\ & \sum_{j=0}^{p-2} \left\{ \begin{array}{c} c(p, j) \times \\ \left(\prod_{s=1}^{j+1} K_m^{\widehat{g}}(X_{i_s}, X_{i_{s+1}}) \right) \\ \times \widehat{\epsilon}_{i_{j+2}} \end{array} \right\} \end{aligned}$$

and

$$\begin{aligned} & V_{p, p, \overline{\psi}_m^{MAR}; i_1, \dots, i_p}(\widehat{\theta}) \\ &= \frac{\widehat{\Delta}_{i_1}}{\widehat{\omega}_{i_1}} \times \\ & \sum_{j=0}^{p-2} \left\{ \begin{array}{c} c(p, j) \\ \left(\prod_{s=1}^{j+1} K_m^{\widehat{g}}(X_{i_s}, X_{i_{s+1}}) \right) \\ A_{i_{s+1}}/\widehat{\pi} \\ \times \widehat{\epsilon}_{i_{j+2}} \end{array} \right\} \end{aligned}$$

where $c(p, j) = \binom{p-2}{j} (-1)^{(j+1)}$, $\widehat{\pi} = P_{\widehat{\theta}}(A = 1)$ is the proportion of the training sample with $A = 1$, and $\widehat{\epsilon}_i = Y_i - \widehat{b}(X_i)$, $\widehat{\Delta}_i = A_i - \widehat{\omega}(X_i)$.

It remains to define a projection kernel $K_m^g(x, y)$ which satisfies conditions (a.4) – (a.7). First consider the case where X is univariate. Assume that $m = 2^M$ for a positive integer M . We first define an m -dimensional subspace of $L_2[0, 1]$ spanned by an orthonormal basis $\{\phi_{M, k}(x), k = 0, \dots, m\}$, where $\phi_{M, k}(x)$ is to be a dilated (by m) and translated (by k) appropriately chosen compactly supported 'father' wavelet (Mallat 1998)–e.g. periodic wavelets, folded wavelets or Daubechies' boundary wavelets with enough vanishing moments are examples. Next, assuming that $\widehat{g}(\cdot)$ is everywhere positive on $[0, 1]$, we define the m -dimensional subspace L_m as:

$$L_m = \left\{ \begin{array}{c} a_m^T \Phi_m(x) : \\ \Phi_m(x) \\ = \left(\frac{\phi_{M, 1}(x)}{\sqrt{\widehat{g}(x)}}, \frac{\phi_{M, 2}(x)}{\sqrt{\widehat{g}(x)}}, \dots, \frac{\phi_{M, m}(x)}{\sqrt{\widehat{g}(x)}} \right)^T, \\ a_m \in R^m \end{array} \right\}.$$

Then $K_m^g(x, y)$ is given by:

$$K_m^g(x, y) = \Phi_m^T(x) (E_g(\Phi_m(X)\Phi_m^T(X)))^{-1} \Phi_m(y)$$

where $E_g(\Phi_m(X)\Phi_m^T(X)) = \int \Phi_m(X)\Phi_m^T(X)g(X)dX$. It follows immediately that $K_m^{\widehat{g}}(x, y) = \Phi_m^T(x)\Phi_m(y)$ which greatly simplifies the computation of $\widehat{\psi}_{p, m}$ by removing the need to invert a $m \times m$ matrix. The same construction can be used in the multivariate case by replacing the univariate basis with a multivariate orthonormal basis formed by the tensor product of the bases $\{\phi_{M, k}(x_1), k = 1, \dots, m\}, \dots, \{\phi_{M, k}(x_d), k = 1, \dots, m\}$ where $x = (x_1, \dots, x_d) \in [0, 1]^d$ is the d -dimensional vector of covariates.

Proposition 7 *Under assumption (a.8) that there is a κ such that $\widehat{g}(\cdot) > \kappa > 0$ on the unit cube in R^d , and $\widehat{g}(x) \in H(\beta^*, C_{\beta^*})$ with $\beta^* = \max(\beta_b, \beta_\omega)$, the kernel $K_m^g(x, y)$ satisfies conditions (a.4)-(a.7).*

A proof is provided in the Appendix in our companion paper.

We now turn to the construction of confidence intervals. For a kernel $K_m^g(\cdot, \cdot)$ based on compact wavelet bases, we prove elsewhere

that $\widehat{\psi}_{p,m}$ is asymptotically normal conditional on the training sample, provided $m(n) \rightarrow \infty$. Therefore Wald type CIs $\mathcal{C}_n^{p,m}$ centered around $\widehat{\psi}_{p,m}$ can be considered. Specifically we define

$$\mathcal{C}_n^{p,m} = \widehat{\psi}_{p,m} \pm 1.96 \sqrt{\widehat{var}(\widehat{\psi}_{p,m}|\widehat{\theta})}$$

where $\widehat{var}(\widehat{\psi}_{p,m}|\widehat{\theta}) = \widehat{var}(\widehat{\psi}_{1,\psi}|\widehat{\theta}) + \sum_{j=2}^p \widehat{var}(U_{j,j,\bar{\psi}_m}(\widehat{\theta})|\widehat{\theta})$,

$$\widehat{var}(\widehat{\psi}_{1,\psi}|\widehat{\theta}) = \mathbb{V}_n \left(V_{1,\psi;i_1}(\widehat{\theta})^2 \right)$$

and, for $p > 1$, $\frac{(n)!}{(n-p)!} \widehat{var}(U_{p,p,\bar{\psi}_m^{WATE}}(\widehat{\theta})|\widehat{\theta}) =$

$$\mathbb{V}_n \left\{ \left[\sum_{j=0}^{p-2} \left(\begin{array}{c} \widehat{\Delta}_{i_1}^2 \times \widehat{\epsilon}_{i_p}^2 \times \\ c(p,j) \times \\ \prod_{s=1}^{j+1} K_m^{\widehat{g}}(X_{i_s}, X_{i_{s+1}}) \end{array} \right) \right]^2 \right\} + \mathbb{V}_n \left\{ \left[\sum_{j=0}^{p-2} \left(\begin{array}{c} (\widehat{\Delta}_{i_1} \widehat{\epsilon}_{i_1}) \times (\widehat{\Delta}_{i_p} \widehat{\epsilon}_{i_p}) \times \\ c(p,j) \\ \times \prod_{s=1}^{j+1} K_m^{\widehat{g}}(X_{i_s}, X_{i_{s+1}}) \end{array} \right) \right]^2 \right\} \\ \frac{(n)!}{(n-p)!} \widehat{var}(U_{p,p,\bar{\psi}_m^{MAR}}(\widehat{\theta})|\widehat{\theta}) = \mathbb{V}_n \left\{ \left[\sum_{j=0}^{p-2} \left(\begin{array}{c} \left(\frac{\widehat{\Delta}_{i_1}}{\widehat{\omega}_{i_1}} \right)^2 \times \widehat{\epsilon}_{i_p}^2 \times \\ c(p,j) \times \\ \prod_{s=1}^{j+1} K_m^{\widehat{g}}(X_{i_s}, X_{i_{s+1}}) \\ \times A_{i_{s+1}}/\widehat{\pi} \end{array} \right) \right]^2 \right\}$$

Note that $\widehat{var}(\widehat{\psi}_{p,m}|\widehat{\theta})$ is an unbiased estimator of $var_{\widehat{\theta}}(\widehat{\psi}_{p,m}|\widehat{\theta})$.

We now analyze the properties of our interval estimator for $p > 1$. We add to our model assumptions bounds on higher order conditional moments of the statistics $V_{j,j,\bar{\psi}_m;i_1,\dots,i_j}(\widehat{\theta})$ that comprise $\widehat{\psi}_{p,m}$. Given such bounds we can show that $var_{\widehat{\theta}}(\widehat{\psi}_{p,m}|\widehat{\theta})$ equals $var_{\theta}(\widehat{\psi}_{p,m}|\widehat{\theta})(1 + o_p(1))$ and that, for $m = m(n) = n^\eta, \eta > 0$, $\left\{ \frac{\widehat{\psi}_{p,m} - E[\widehat{\psi}_{p,m}|\widehat{\theta}]}{\widehat{var}(\widehat{\psi}_{p,m}|\widehat{\theta})^{1/2}} \right\}$

converges in law to a $N(0,1)$ distribution uniformly over the model. Suppose that $m_{opt}(p) = n^{\eta_p}, \eta_p > 0$. Then if $\eta > \eta_p$,

$\left[E[\widehat{\psi}_{p,m}|\widehat{\theta}] / \widehat{var}(\widehat{\psi}_{p,m}|\widehat{\theta})^{1/2} \right] = o_p(1)$ uniformly over the model. Thus $\left\{ \frac{\widehat{\psi}_{p,m}}{\widehat{var}(\widehat{\psi}_{p,m}|\widehat{\theta})^{1/2}} \right\}$

converges uniformly to a $N(0,1)$ and the interval $\mathcal{C}_n^{p,m}$ is a conservative uniform asymptotic $(1 - \alpha)$ CI for $\psi(\theta)$ in both of our models.

4. Discussion

Different subject matter experts will disagree as to the maximum roughness of the functions b, ω, g . In contrast to point estimates, for honest confidence intervals, the degree of adaption to unknown smoothness is minimal. Therefore an analyst should report a mapping from apriori roughness assumptions encoded in the exponents and radii of Holder balls (or by other measures of roughness) to the associated optimal $(1 - \alpha)$ honest confidence intervals proposed in this paper. Elsewhere we describe how our methods can be used to construct adaptive point estimates.

Our methods also apply directly to the analysis of models whose dimension increases with sample size. Indeed, such a model is the submodel $\mathcal{M}(\Theta_{sub,n^\eta})$, η known, of the WATE model on which $\psi(\theta) = \bar{\psi}_{n^\eta}(\theta)$. The dimensions of $b(x) = \alpha_{b,n^\eta}^T \Phi_{n^\eta}(x)$ and $\omega(x) = \alpha_{\omega,n^\eta}^T \Phi_{n^\eta}(x)$ increase as n^η . Valid point and interval estimation for $\psi(\theta)$ can be based on the estimators $\widehat{\psi}_{p,m}$. In order to find the optimal choice of p and m in this model, we note (i) there is truncation bias only for $m < n^\eta$, (ii) the variance remains $O\left(\frac{1}{n} \max\left(1, \left(\frac{m}{n}\right)^{p-1}\right)\right)$, and (iii) the estimation bias order will be determined by β_g and additional model restrictions (if any) placed on the fraction of non-zero components or on the rate of decay of the components of the vectors α_{ω,n^η} and α_{b,n^η} as n increases.