

Robust and Honest Confidence Intervals for Causal Effects: Application of a Unified Theory of Parametric, Semi and Nonparametric Statistics Based On Higher Dimensional Influence Functions.

James Robins, Harvard School of Public Health

(the talk is based on joint work with Aad van der Vaart, Eric Tchetgen, and Lingling Li)

Suppose for n (say, 5000) subjects, data are available on a dichotomous treatment R , a continuous response Y , and a high (say 10) dimensional vector of pre-treatment covariates X with compact support.

We wish to estimate one wishes to estimate the average treatment effect (ACE) of R on Y .

$$ATE = \tau = E[Y(1)] - E[Y(0)]$$

That is we wish to estimate τ in the simplest MSM

$$MSM : Y_a = \varrho + \tau a$$

We assume no unmeasured confounding (ignorability given X).

$$(Y(1), Y(0)) \text{ independent of } A|X$$

and

$$0 < \text{pr}(A = 1|X) = p(X) = P < 1 \text{ for all } X$$

Problem: How do we control confounding by the 10 dimensional vector X of measured covariates?

Note a weighted average treatment effect $w - ATE$ is $w(X)$

$$w - ATE = \int w(X) ATE(X)$$

where

$$ATE(X) = E[Y(1)|X] - E[Y(0)|X]$$

$w(X)$ a user supplied density for X

ATE has $w = f_X$

More efficient estimators of parameters with $w(X)$ proportional to $var(A|X) = p(X)(1 - p(X))$ - Imbens et al

Simple stratification or matching will not work because no two subjects are terribly close in 10 dimensional space

and matching does not utilize the smoothness of $p(X)$ or $E[Y|A, X]$ with potentially large increases in bias and variance

Common analytic approaches include

(i) fitting by OLS a ‘working’ outcome regression (OR) model for the regression of Y on R and Z

$$\begin{aligned} E[Y|A, X] \\ = \sum_{k=1}^{10} \eta_0 + \tau A + \sum_{k=1}^{10} \eta_{1k} X_k + \sum_{l=1}^{10} \sum_{m=1}^{10} \eta_{2,lm} X_1 X_2 \end{aligned}$$

possibly including interactions with A . (WITHOUT INTERACTION τ IS THE ATE)

(ii) subclassification by, matching on, or inverse probability of treatment weighting by, an estimate of the propensity score based on a working logistic model for the probability of treatment given Z

$$\begin{aligned} \log \text{it pr}(A = 1|X) \\ = \alpha_0 + \alpha_{1k}X_k + \sum_{l=1}^{10} \sum_{m=1}^{10} \alpha_{2,lm}X_1X_2 \end{aligned}$$

Both approaches estimate the treatment effect at the usual parametric rate of square root n if their respective working model is correct.

However, the OR approach is biased if the working outcome regression model is misspecified, while the propensity approaches are biased if the working propensity model is misspecified.

A much improved approach is to use a doubly robust estimator that is guaranteed to estimate the treatment effect at the usual parametric rate if either (but not necessarily both) of the two working models are correct.

This can be obtained by fitting the model OR model with the additon of one critical covariate Q where

$$Q = \frac{1}{\widehat{P}} \text{ if } A = 1$$

$$Q = \frac{1}{1 - \widehat{P}} \text{ if } A = 0$$

.

$$E[Y|A, X]$$

$$= \sum_{k=1}^{10} \eta_0 + \tau A + \eta_{1k} X_k + \sum_{l=1}^{10} \sum_{m=1}^{10} \eta_{2,lm} X_1 X_2 + \varphi Q$$

This solves the long standing question of whether better to control for confounding using propensity scores or outcome regression. That is, use both and compute a DR estimator.

Often surprisingly little loss of efficiency compared to ML even if outcome model correct.

However even a doubly robust estimator has the following problem .

We get square root n rates and valid CI of radius $1/\text{square root } n$ if either working model is correct but an inconsistent estimate and invalid confidence intervals that under-cover if both models are sufficiently wrong.

But models are sure to be wrong.

Further with high dimensional Z , due to lack of power, we cannot use the data to determine whether even fairly large working models are sufficiently close to being correct that confounding is controlled.

Thus it would be a more honest to use confidence intervals that

(i) shrink to zero (with increasing sample size) at much slower rates than the usual parametric rate of $1/\sqrt{n}$ but

(ii) are much more robust to misspecification of the working models, in the sense that even if the models are not quite correct, the now larger confidence intervals still include the true treatment effect at their nominal coverage rate.

Such confidence intervals did not exist. But now they do.

These interval estimators are centered on a higher dimensional U-statistic estimator of the ATE .

These U-statistic estimators are derived using a new unified theory of parametric, semi , and nonparametric statistics based on higher order scores (i.e., derivatives of the likelihood), and higher order influence functions that

applies equally to both the square-root-n and non-square-root n problems,

reproduces the results previously obtained by the modern theory of non-parametric inference,

produces many new non-root- n results,

and most importantly opens up the ability to perform optimal non-root n inference in complex high dimensional models.

Specifically, all longitudinal causal and CAR censored models in book with van der laan on complex longitudinal data

I'll will show the estimators later as hard even to explain the notation in a short time.

How well one can estimate either the propensity score or the outcome regression depends on the smoothness (complexity) of the true propensity function and outcome regression function as a function of X .

On measure of smoothness is the number of partial derivatives if they cannot be too big. More on this later.

Not too big formalized by the diameter of a Holder or Sobolev ball.

Specifically how well one can estimate either the propensity score or the outcome regression depends on the ratio of the highest degree of derivatives to the dimension of X . Doubly robust estimators work only if both functions have a ratio of greater than $1/2$ (i.e. in our example more than 5th derivatives that are not too big).

More technically it is not the number of derivatives but the number of derivatives that are not big. This last concept measures the wiggleness of a function.

Often one would not believe 5 small derivatives based on substantive knowledge, determined by showing an expert wiggly functions and asking which are plausible.

A whole open area: substantively useful measures of complexity. Methodology can take as input any such measure

We need to know the number of derivatives to decide which higher order U-stat estimator gives a CI that both covers at the nominal rate and is as narrow as possible.

Thus an analyst should report a sensitivity analysis mapping assumed smoothness to optimal CI.

Less smoothness, bigger intervals

.

Do estimation in $A = 1$ and then in $A = 0$. So ignore data on $Y(0)$ and it is a missing outcome problem.

$$O = (AY, X, A) = (AY(1), X, A)$$

$$L = (Y(1), X, A)$$

$X = \text{high dim vector of always observed cov}$

$A = \text{binary treatment}$

$$\psi = E[Y(1)]$$

$$pr[A = 1|Y(1), X] = pr[A = 1|X]$$

MAR=ignorable=no unmeasured conf

$$\psi = E [b (X)] = E \left[\frac{A}{p (X)} Y \right]$$

$$\begin{aligned} b (X) &= E [Y | X, A = 1] \\ &= E [Y (1) | X, A = 1] \\ &= E [Y (1) | X] \end{aligned}$$

$$p (X) = pr [A = 1 | X] = E [A | X]$$

$$\psi = E [b (X)] = E \left[\frac{A}{p (X)} Y \right]$$

All longitudinal causal and CAR censored models in book
with van der laan on complex longitudinal data

The statistics:

First estimate the marginal density $g(X)$ at optimal rate using kernels, wavelets, or log tensor splines $n^{-\frac{\beta_g}{2\beta_g+d}}$, $d = 10$.

If splines use $n^{\frac{1}{2\beta_g/d+1}}$ terms. For example if $\beta_g = 1$, we use $n^{\frac{1}{1.2}} = 5000^{(1/2)} = 1209$ tensor spline basis functions

and fit

$$g(X) = c \exp \left\{ \sum_{l=1}^{1209} \omega_l s_l(X) \right\}$$

Next define

$$\hat{z}_l(X) = \phi_l(X) / \hat{g}(X)^{1/2}$$

where $\{\phi_l(x), 1, 2, \dots\}$ is d -fold tensor product of ON basis for $L_2(\mu)$ in \mathbb{R}^1 so is ON basis for $L_2(\mu)$ in \mathbb{R}^d .

ON basis for $L_2(\mu)$ in \mathbb{R}^1 could be Fourier, orthogonal polynomial, spline or compact wavelet basis with appropriate no. of vanishing moments

$$\{\widehat{Z}_l = \widehat{z}_l(x), l = 1, 2, \dots\}$$

is ON basis for $L_2(\widehat{g})$ in \mathbb{R}^d

Fit by OLS model

$$\begin{aligned} E[Y|A = 1, X] &= b(X) = \\ &= \eta_l^T \hat{Z}_M \end{aligned}$$

to obtain $\hat{b}(X) = \hat{\eta}_l^T \hat{Z}_M$

and by logistic regression

$$\begin{aligned} \log \textit{it pr}(A = 1|X) \\ &= \log \textit{it } p(X) \\ &= \alpha_l^T \hat{Z}_M \end{aligned}$$

Assuming, for simplicity $b(X)$ and $p(X)$ both have β – *derivatives*,

$$M = n^{\frac{1}{2\beta/d+1}}$$

Compute $\hat{\psi}_{DR,opt} = n^{-1} \sum_{i=1}^n \tilde{b}(X_i)$ where $\tilde{b}(x)$ is the predicted value from the fit

$$\begin{aligned} E[Y|A=1, X] &= b(X) = \\ &= \eta_l^T \hat{Z}_M + \varphi \frac{A}{\hat{p}(X)} \end{aligned}$$

$$\widehat{\psi}_2 = \widehat{\psi}_{DR,opt} + \widehat{IF}_{2,2}$$

$$\widehat{IF}_{2,2} = [n\left(n-1\right)]^{-1}\sum_{i\neq j}\widehat{h}_2\left(O_i,O_j\right)$$

$$\widehat{h}_2\left(O_i,O_j\right)\\ =\frac{A_i}{\widehat{p}\left(X_i\right)}\left(Y_i-\widehat{b}\left(X_i\right)\right)\widehat{\bar{Z}}_{ki}^T\widehat{\bar{Z}}_{kj}\times$$

$$\left(\frac{A_j-\widehat{p}\left(X_j\right)}{\widehat{p}\left(X_j\right)}\right)$$

$$=\frac{A_i}{\widehat{p}\left(X_i\right)}\left(Y_i-\widehat{b}\left(X_i\right)\right)\sum_{l=1}^k\widehat{Z}_{li}\widehat{Z}_{lj}$$

$$\left(\frac{A_j-\widehat{p}\left(X_j\right)}{\widehat{p}\left(X_j\right)}\right)$$

$$k=k\left(n\right)$$

$$=\max\left(n^{\frac{2}{1+4\beta/d}},n^{2-\frac{4\beta/d}{2\beta/d+1}-\frac{2\beta_g/d}{2\beta_g/d+1}}\right)$$

If $\beta = 1, \beta_g = 1,$

$$\begin{aligned}
 & k \\
 &= \left(5000^{\left(\frac{2}{1+.4}\right)}, 5000^{\left(2-\frac{.4}{1.2}-\frac{.2}{1.2}\right)} \right) \\
 &= \max(192, 420, 353, 550) \\
 &= 353, 550
 \end{aligned}$$

Suprisingly $\hat{\psi}_2$ is AN with variance that can easily be consistently estimated so standard wald intervals based on

$$\hat{\psi}_2 \pm 1.96 \left[\widehat{var} \left(\hat{\psi}_2 \right) \right]^{1/2}$$

Results depend on smoothness β_g of g because $k > n = 5000$, so beyond the empirical.

Ill-posed problem. Depends on reality that is not fully empirically verifiable.

Ritove -Bickel and Laurie-Davies: Ignore ill posed problems

Response: Causal inference because of confounding is an illposed problem

$\sum_{l=1}^k \widehat{Z}_{li} \widehat{Z}_{lj}$ is a Dirac kernel

$$\begin{aligned}
 & E \left[\sum_{l=1}^k \widehat{z}_l (X_i) \widehat{z}_l (X_j) h (X_j) | X_i \right] \\
 &= \int \sum_{l=1}^k \widehat{z}_l (X_i) \widehat{z}_l (X_j) h (X_j) g (X_j) dX_j \\
 &\rightarrow h (X_i) \text{ as } k \rightarrow \infty
 \end{aligned}$$

$$\int h (X_j) K (X_i, X_j) dX_j = h (X_i)$$

$$K (X_i, X_j) = \delta (X_i - X_j)$$

$\delta (X_i - X_j)$ not in L_2 — generalized function.

$$\widehat{\psi}_3 = \widehat{\psi}_2 + \widehat{IF}_{3,3}$$

$$\widehat{IF}_{3,3} = [n(n-1)(n-2)]^{-1} \sum_{i \neq j \neq s} \widehat{h}_3(O_i, O_j, O_s)$$

$$\begin{aligned} &\widehat{h}_3(O_i, O_j, O_s) \\ &= \frac{A_i}{\widehat{p}(X_i)} \left(Y_i - \widehat{b}(X_i) \right) \overline{Z}_{ki}^T \times \\ &\quad \left\{ \frac{A_s}{\widehat{p}(X_s)} \overline{Z}_{ks} \overline{Z}_{ks}^T - I \right\} \times \\ &\quad \overline{Z}_{kj} \left(\frac{A_j - \widehat{p}(X_j)}{\widehat{p}(X_j)} \right) \end{aligned}$$

$$\widehat{\psi}_3 \text{ is AN}$$

$$\widehat{\psi}_m = \widehat{\psi}_{m-1} + \widehat{IF}_{m,m}$$

$$\begin{aligned}\widehat{IF}_{m,m} &= (-1)^m \frac{1}{n \times \dots \times (n - m + 1)} \\ &\times \sum_{r_1 \neq \dots \neq r_m} \widehat{h}_m(O_{r_1}, \dots, O_{r_m}) \\ \widehat{h}_3(O_{r_1}, \dots, O_{r_m}) &= \\ &= \frac{A_{r_1}}{\widehat{p}(X_{r_1})} \left(Y_{r_1} - \widehat{b}(X_{r_1}) \right) \times \overline{Z}_{kr_1}^T \\ &\prod_{s=2}^{m-1} \left\{ \frac{A_{r_s}}{\widehat{p}(X_{r_s})} \overline{Z}_{kr_s} \overline{Z}_{kr_s}^T - I \right\} \times \\ &\left(\frac{A_{r_m} - \widehat{p}(X_{r_m})}{\widehat{p}(X_{r_m})} \right) \overline{Z}_{kr_m}\end{aligned}$$

$$\widehat{\psi}_m \text{ is AN}$$

Randomized Trial with Non compliance: 5000 subjects

Data

$$O = (R, A, X, Y) =$$

$X = d = 10$ dim *vector* of cont cov

$R =$ *randomization* estimator

$A =$ *binary treatment*

$Y =$ *cont response*

$$pr [R = 1 | X] \text{ known}$$

assumptions

1.If $R = 0$ then $A = 0$ so no defiers or always takers.

2.Exclusion Restriction: $Y (R = r, A = a) = Y (A = a)$
for $r = 0, 1$ and $a = 0, 1$.

Treatment Effect in the Treated (Compliers)

$$\begin{aligned}\psi(X) &= E[Y(1) - Y(0) | X, A = 1, R = 1] \\ &= E[Y(1) - Y(0) | X, A = 1]\end{aligned}$$

Goal Estimate and Confidence Interval for

$$d_{opt}(X) = I(\psi(X) > 0)$$

$d_{opt}(x)$ says treat if and only if mean $Y(1)$ *exceeds* mean $Y(0)$ in subjects with $X = x$.

Comments:

$d_{opt}(X)$ may not be of interest after the trial as treated will change if heterogeneity in treatment effect

Terribly ambitious. Looking for qualitative interaction at each level x of X . Crazy without confidence intervals.

Richard Peto admonitions.

Equally of interest in RCT with no noncompliance

If only interest qualitative interaction could look at ITT parameter

$$\psi_{ITT}(X) = E[Y|R=1, X] - E[Y|R=0, X]$$

Not possible if we redefine

$$d_{opt}(X) = I(\psi(X) > 5)$$

Don't give treatment unless large clinical effect.

Task 1: Construct Adaptive Estimate $\hat{\psi}_{adap}(X)$ of $\psi(X)$.
 $\hat{d}_{opt, adap}(X) = I(\hat{\psi}_{adap}(X) > 0)$.

Step 1: Consider 20,000 models $j = 1, \dots, 20,000$

$\psi(L) = \psi_j(X; \eta_j)$ with the dimension of η_j from 1 to 4000 sa

as well as many different functional forms

Let $\hat{\eta}_j$ solve

$$\sum_{i=1}^{4000} \left\{ Y_i - A_i \psi_j(X_i; \eta_j) \right\} (A_i - P_i) \left[\partial \psi_j(X_i; \eta_j) / \partial \eta_j \right]$$

based on 4000 random observations.

$$\hat{d}_j(X) = I[\psi_j(X_i; \hat{\eta}_j) > 0]$$

With remaining 1000 observations choose among the $\hat{\eta}_j$ by cross validation.

\hat{j}_{opt} is the j satisfying

$$\arg \max_j \left\{ \sum_{i=1}^{1000} Y I \left(A = \hat{d}_j (X) \right) / f (A|X) \right\}$$

$$f (A|L) = p (X)^A (1 - p (X))^{1-A}$$

that has the greatest expected outcome.

Then

$$\hat{\psi}_{adap} (X) = \psi_{\hat{j}_{opt}} \left(X_i; \hat{\eta}_{\hat{j}_{opt}} \right)$$

$$\hat{d}_{opt, adap} (X) = I \left(\hat{\psi}_{adap} (X) > 0 \right)$$

Task 2: Construct Adaptive 95% Confidence Interval for $\psi(X)$ centered on $\hat{\psi}_{adap}(X)$.

Need to assume $\psi(X)$ has β derivatives and $g(X)$ has β_g derivatives

so again will require map from smoothness assumptions to CI.

M in the following will depend on β

k in the following will depend on β, β_g

Again define

$$\hat{z}_l(X) = \phi_l(X) / \hat{g}(X)^{1/2}$$

where $\{\phi_l(x), 1, 2, \dots\}$ is d -fold tensor product of ON basis for $L_2(\mu)$ in \mathbb{R}^1 so is ON basis for $L_2(\mu)$ in \mathbb{R}^d .

Let

$$\begin{aligned}\hat{\psi}_{adap,k}(X) &= \hat{\bar{\theta}}_{adap,k}^T \hat{\bar{z}}_k(X) \\ \hat{\bar{\theta}}_{adap,k} &= \hat{E} \left[\hat{\psi}_{adap}(X) \hat{\bar{z}}_k(X) \right] \\ &= \int \hat{\psi}_{adap}(X) \hat{\bar{z}}_k(X) \hat{g}(X) dX\end{aligned}$$

$$\begin{aligned}\psi_k(X) &= \bar{\theta}_k^T \hat{\bar{z}}_k(X) \\ \bar{\theta}_k &= E \left[\psi(X) \hat{\bar{z}}_k(X) \right]\end{aligned}$$

where

$$k = \max \left(n^{\frac{2}{1+4\beta/d}}, n^{2 - \frac{4\beta/d}{2\beta/d+1} - \frac{2\beta_g/d}{2\beta_g/d+1}} \right)$$

If $\beta = \beta_g = 1$,

$$\begin{aligned}k &= \max(192, 420, 353, 550) \\ &= 550\end{aligned}$$

$$C_{\bar{\theta}_k} (.95)$$

$$= \left\{ \bar{\theta}_k : \left(\hat{\bar{\theta}}_{adapt,k} - \bar{\theta}_k \right)^T \left(\hat{\bar{\theta}}_{adapt,k} - \bar{\theta}_k \right) < Q^2 \right\}$$

$$C_{\psi} (.95)$$

$$= \left\{ \psi (\cdot) : \hat{E} \left[\left\{ \psi (X) - \hat{\psi}_{adapt,k} (X) \right\}^2 \right] < Q^2 \right\}$$

$$C_{dop} (.95)$$

$$= \left\{ d (\cdot) : d (X) = I (\psi (X) > 0), \psi (X) \in C_{\psi} (.95) \right\}$$

What is Q^2 ?

Let $v(X) = p(X) \{1 - p(X)\}$

Let $\widehat{\bar{\theta}}_{\beta, M}$ solve

$$0 = \sum_{i=1}^{5000} \left\{ Y_i - A_i \widehat{\bar{z}}_M(X)^T \bar{\theta}_{\beta, M} \right\} (A_i - P_i) \widehat{\bar{z}}_M(X) v(X)^{-1}$$

where $M = n^{\frac{1}{2\beta/d+1}}$.

If $\beta = 1$, we use $n^{\frac{1}{1.2}} = 5000^{(1/2)} = 1209$ tensor basis functions

$$\begin{aligned} \widehat{\bar{\theta}}_{\beta, k} &= \left(\widehat{\bar{\theta}}_{\beta, M}, 0 \right) \\ \widehat{\psi}_k(X) &= \widehat{\bar{\theta}}_k^T \widehat{\bar{z}}_k(X) \end{aligned}$$

Let

$$\begin{aligned}
& R_k \left(\widehat{\theta}_{adap,k} \right) \\
&= \sum_{i=1}^n 2 \left(\widehat{\theta}_{adap,k} - \widehat{\theta}_k^T \right)^T \widehat{\bar{z}}_k (X_i) \left\{ Y_i - A_i \widehat{\bar{z}}_k (X_i)^T \widehat{\theta}_k^T \right\} \times \\
& \quad (A_i - P_i) v (X_i)^{-1} \\
& \quad - \left(\widehat{\theta}_{adap,k} - \widehat{\theta}_k^T \right)^T \times \\
& \quad \left\{ \widehat{\bar{z}}_k (X_i) A_i (A_i - P_i) v (X_i)^{-1} \widehat{\bar{z}}_k (X_i)^T - I_{k \times k} \right\} \times \\
& \quad \widehat{\bar{z}}_k (X_j)^T \left\{ Y_j - A_j \widehat{\bar{z}}_k (X_j)^T \widehat{\theta}_k^T \right\} \times \\
& \quad (A_j - P_j) v (X_j)^{-1} \widehat{\bar{z}}_k (X_j)^T \\
& Q^2 = R_k \left(\widehat{\theta}_{adap,k} \right) + z_\alpha \left\{ \widehat{var} \left[R_k \left(\widehat{\theta}_{adap,k} \right) \right] \right\}^{1/2}
\end{aligned}$$

Then

EB .

$$= c_{EB} E \left[\begin{array}{c} \left[\left\{ f(X_i) - \hat{f}(X_i) \right\} + \left\{ p(X_i) - \hat{p}(X_i) \right\} \right] \\ \times \left\{ b(X_i) - \hat{b}(X_i) \right\} \left\{ p(X_i) - \hat{p}(X_i) \right\} \end{array} \right]$$

$$= O \left(\max \left[n^{-\frac{2\beta_f^*}{2\beta_f^*+d}}, n^{-\frac{2\beta_p^*}{2\beta_p^*+d}} \right] n^{\frac{-2\beta_p^*}{2\beta_p^*+d}} n^{\frac{-2\beta_b^*}{2\beta_b^*+d}} \right)$$

When $(\beta_b^* + \beta_p^*) / d < 1/2$

Then $k = n \Rightarrow$

$$TB^2 = n^{-2(\beta_b^* + \beta_p^*)/d} > n^{-1} = \text{var} \left\{ IF_2(\hat{\theta}) \right\}$$

so need $k > n$

CI does not shrink at $n^{-1/2}$

Price of valid intervals

When $(\beta_b^* + \beta_p^*) / d > 1/2$
then $\Rightarrow k < n$ and we get first order efficiency
provided $\beta_f^* > \beta_p^*$
since then $EB^2 < n^{-1}$

Otherwise we may need $k > n$ and CI does not shrink at n^{-1}

This is unlike $\int f(Z)^2$ where IF_2 is always efficient if
 $(2\beta) / d > 1/2$

We will see that we are efficient for some IF_m whatever
be $\beta_f^* > 0$ whenever $(\beta_b^* + \beta_p^*) / d > 1/2$.

However when $(\beta_b^* + \beta_p^*) / d < 1/2$, the (conjectured)
optimal rate of convergence depends on β_f^*

whenever $\beta_f^* > \beta_p^*$.

$IF_{2,2}(\hat{\theta})$ is AN (when properly standardized).

When $k > n$, standard martingale CLTs fail. Need to argue in an new way, by proving conditional AN given knowing which subjects had their X in the support of various compact wavelengths. Aad van der vaart proving this will work with any basis , splines Fourier even not local. That is subtle.

Indeed all $IF_{m,m}(\hat{\theta})$ are AN

Consistent variance estimator is the empirical:

Under law $\hat{\theta}$, all $n(n-1)$

$$h_2(O_i, O_j; \hat{\theta}) = \frac{A_i}{\hat{p}(X_i)} (Y_i - \hat{b}(X_i)) \hat{Z}_{ki}^T \hat{Z}_{kj} \times \left(\frac{A_j - \hat{p}(X_j)}{\hat{p}(X_j)} \right)$$

are mean zero and uncorrelated (even with $h_2(O_{ji}, O_i; \hat{\theta})$) and also uncorrelated with

$$n^{-1} \sum_i \frac{A_i}{\hat{p}(X_i)} (Y_1 - \hat{b}(X_i)) + \hat{b}(X_i) - \psi(\hat{\theta})$$

So under $\hat{\theta}$, unbiased estimate of variance is

$$n^{-2} \sum_i \left\{ \frac{A_i}{\hat{p}(X_i)} (Y_1 - \hat{b}(X_i)) + \hat{b}(X_i) - \psi(\hat{\theta}) \right\}^2 \\ + [n(n-1)]^{-2} \sum_{i \neq j} \left\{ h_2(O_i, O_j; \hat{\theta}) \right\}^2$$

Properly standardized it is consistent for variance under $\hat{\theta}$ and thus θ

since $\hat{\theta} \rightarrow_p \theta$

Problem with our intervals so far.

We equalized rates of variance and bias as to rate

So we have a problem with constants since we need bias shrinking to 0 for no asym bias and correct coverage

Natural solution: choose k as above multiplied by n^ϵ or $\ln(n)$. Good math solution but in actual data analysis not a

solution to get correct finite sample coverage.

.

Honest solution is as follows. If $h(X)$ is a regression function or density in Holder (β, c) in principle we can find an estimator $\hat{h}(X)$ that satisfies for every sample size $n \geq N$

$$\sup_{F_{\epsilon,X}} \sup_{\mathcal{F}_{\epsilon,X}} \sup_{h \in H(\beta,c)} E \left[\left\| \hat{h}_n(X) - h(X) \right\|_p \right] \leq c_{+,p} \left(\beta^*, c, N, \mathcal{F}_{\epsilon} \right).$$

Now by Holder with $p = 3$

$$\begin{aligned} & E \left[\left\| \begin{bmatrix} \left\{ f(X_i) - \hat{f}(X_i) \right\} \\ \left\{ b(X_i) - \hat{b}(X_i) \right\} \left\{ p(X_i) - \hat{p}(X_i) \right\} \end{bmatrix} \right\| \right] \\ & \leq E \left[\left\| f(X_i) - \hat{f}(X_i) \right\|_p \right] E \left[\left\| b(X_i) - \hat{b}(X_i) \right\|_p \right] \\ & \quad \times E \left[\left\| p(X_i) - \hat{p}(X_i) \right\|_p \right] \end{aligned}$$

which can be bounded as as above with $N^* = n$. Also we can bound c_{EB} and c_{TB} . This will allow us to bound by some Q_n

$$T_{ot}B < TB + EB$$

and we use

$$\hat{\psi}_m \pm \left\{ \widehat{var} \left[IF_m \left(\hat{\theta} \right) \right]^{1/2} + Q_n \right\} z_{\alpha}$$

Now coverage is guaranteed

Model:

Average of $100 * 1.645 * \text{empirical s.d.}$

C_b	<i>Unconditional</i>			<i>Conditional</i>		
	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$
6	0.94	2.95	5.43	0.94	2.97	5.63
8	9.43	2.95	5.44	0.94	2.96	5.61
9	0.94	2.95	5.44	0.94	2.96	5.62
10	0.94	2.95	5.44	0.94	2.97	5.62
12	0.94	2.95	5.44	0.94	2.96	5.62
20	0.94	2.95	5.42	0.94	2.97	5.62

Coverage rate at nominal level 90%.

	<i>Unconditional</i>			<i>Conditional</i>		
C_b	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$
6	91	87	95	91	88	90
8	89	86	93	85	86	80
9	80	85	92	81	84	90
10	67	82	92	70	82	92
12	35	78	92	40	78	92
20	0	48	78	0	47	75

Missing Data $A = 1$:

Average of $100 * 1.645 * \text{empirical s.d.}$

C_b	<i>Unconditional</i>			<i>Conditional</i>		
	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$
6	2.82	10.2	26.98	2.73	9.02	22.71
8	2.81	10.1	27.46	2.71	8.80	22.14
9	2.81	10.10	26.86	2.71	8.74	21.95
10	2.80	9.98	25.58	2.69	8.73	22.09
12	3.43	10.58	26.5	2.67	8.62	21.5
20	2.74	9.55	24.31	2.6	8.28	20.3

Coverage rate at nominal level 90%.

C_b	<i>Unconditional</i>			<i>Conditional</i>		
	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$
6	92	87	86	92	90	86
8	90	88	84	90	91	87
9	83	88	86	82	89	87
10	80	88	86	79	89	85
12	69	85	90	70	88	88
20	3	72	81	2	67	80

Missing Data $A = 0$:

Average of $100 * 1.645 * \text{empirical s.d.}$

C_b	<i>Unconditional</i>			<i>Conditional</i>		
	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$
6	2.86	10.66	30.97	2.88	10.22	28.7
8	2.86	10.61	30.36	2.87	10	27.46
9	2.87	10.65	30.96	2.88	10.08	27.68
10	2.87	10.72	31.47	2.9	10.35	28.9
12	2.87	10.64	30.18	2.92	10.91	31.69
20	2.89	10.7	29.18	2.97	11.34	32.56

Coverage rate at nominal level 90%.

C_b	<i>Unconditional</i>			<i>Conditional</i>		
	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$
6	92	84	86	86	91	83
8	85	85	87	80	87	86
9	85	83	87	87	89	84
10	80	83	87	76	88	82
12	68	82	85	62	88	83
20	2	77	81	1	77	84

$$\begin{aligned}
IF_3\left(\widehat{\theta}\right) &= IF_2\left(\widehat{\theta}\right) + IF_{3,3}\left(\widehat{\theta}\right) \\
IF_{3,3}\left(\widehat{\theta}\right) &= [n\left(n-1\right)\left(n-2\right)]^{-1} \sum_{i \neq j \neq s} h_3\left(O_i, O_j, O_s; \widehat{\theta}\right) \\
&h_3\left(O_i, O_j, O_s; \widehat{\theta}\right) \\
&= \frac{A_i}{\widehat{p}\left(X_i\right)}\left(Y_i-\widehat{b}\left(X_i\right)\right) \overline{Z}_{k i}^T \times \\
&\left\{\frac{A_s}{\widehat{p}\left(X_s\right)} \overline{Z}_{k s} \overline{Z}_{k s}^T-I\right\} \times \\
&\overline{Z}_{k j}\left(\frac{A_j-\widehat{p}\left(X_j\right)}{\widehat{p}\left(X_j\right)}\right)
\end{aligned}$$

Must choose $k_3 = k_3(n)$ such that

$$\begin{aligned}
 & \text{var} \left\{ IF_3 \left(\hat{\theta} \right) \right\} \\
 &= O \left(\frac{1}{n} \frac{k}{n} \frac{k}{n} \right) \\
 &= \\
 & E \left[IF_{3,3} \left(\hat{\theta} \right) - \psi \right]^2 \\
 &= \max \left\{ \left[\max \left(n^{-\frac{-2\beta_f^*}{4\beta_f^*+d}}, n^{\frac{-2\beta_p^*}{4\beta_p^*+d}} \right) \right]^2, n^{\frac{-2\beta_p^*}{2\beta_p^*+d}} n^{\frac{-2\beta_b^*}{2\beta_b^*+d}}, k^{-2(\beta_b^*+\beta)} \right\}
 \end{aligned}$$

EB .

$$= cE \left[\frac{\left\{ b(X_i) - \hat{b}(X_i) \right\} \left\{ p(X_i) - \hat{p}(X_i) \right\} \times}{\left[\left\{ f(X_i) - \hat{f}(X_i) \right\} + \left\{ p(X_i) - \hat{p}(X_i) \right\} \right]^2} \right]$$

The advantage of $IF_3 \left(\hat{\theta} \right)$ is that if $IF_2 \left(\hat{\theta} \right)$ has estimation bias dominate truncation bias

then $IF_3 \left(\hat{\theta} \right)$ with smaller EB

can improve rate of convergence.

For example when $(\beta_b^* + \beta_p^*)/d > 1/2$ and $\beta_f^* > \beta_p^*$
 $\widehat{\psi}_3$ may be efficient but $\widehat{\psi}_2$ not.

As $\beta_f^* \rightarrow 0$, $m \rightarrow \infty$ for $\widehat{\psi}_m$ to be efficient (practical sample size problems)

Mapping from smoothness assumptions to optimal CI?

What smoothness or other size controlling assumptions.

$$IF_m\left(\widehat{\theta}\right)=IF_{m-1}\left(\widehat{\theta}\right)+IF_{m,m}\left(\widehat{\theta}\right)$$

$$IF_{m,m}\left(\widehat{\theta}\right)=(-1)^m\frac{1}{n\times...\times(n-m+1)}\times\sum_{r_1\neq...\neq r_{m-2}}\sum_{i\neq j\neq}$$

$$h_3\left(O_{r_1},...,O_{r_m},;\widehat{\theta}\right)=$$

$$=\frac{A_{r_1}}{\widehat{p}\left(X_{r_1}\right)}\left(Y_{r_1}-\widehat{b}\left(X_{r_1}\right)\right)\times\overline{Z}_{kr_1}^T$$

$$\prod_{s=2}^{m-1}\left\{\frac{A_{r_s}}{\widehat{p}\left(X_{r_s}\right)}\overline{Z}_{kr_s}\overline{Z}_{kr_s}^T-I\right\}\times$$

$$\left(\frac{A_{r_m}-\widehat{p}\left(X_{r_m}\right)}{\widehat{p}\left(X_{r_m}\right)}\right)\overline{Z}_{kr_m}$$

Must choose $k_m = k_m(n)$ such that

$$\begin{aligned}
 & \text{var} \left\{ IF_m \left(\hat{\theta} \right) \right\} \\
 &= O \left(\frac{1}{n} \left(\frac{k}{n} \right)^{m-1} \right) \\
 &= \\
 & E \left[IF_{m,m} \left(\hat{\theta} \right) - \psi \right]^2 \\
 &= \max \left\{ \left[\max \left(n^{-\frac{-2\beta_f^*}{4\beta_f^*+d}}, n^{\frac{-2\beta_p^*}{4\beta_p^*+d}} \right) \right]^{m-1} n^{\frac{-2\beta_p^*}{2\beta_p^*+d}} n^{\frac{-2\beta_b^*}{2\beta_b^*+d}}, k^{-2(\beta_f + \beta_p + \beta_b)} \right\}
 \end{aligned}$$

EB .

$$= cE \left[\left[\left\{ b(X_i) - \hat{b}(X_i) \right\} \left\{ p(X_i) - \hat{p}(X_i) \right\} \times \left[\left\{ f(X_i) - \hat{f}(X_i) \right\} + \left\{ p(X_i) - \hat{p}(X_i) \right\} \right]^{m-1} \right] \right]$$

Given a sufficiently smooth p – dimensional parametric submodel $\tilde{\theta}(\varsigma)$ mapping $\varsigma \in A^p$ injectively into Θ , define

$$\psi_{\setminus i_1 \dots i_m}(\theta) = \left(\psi \circ \tilde{\theta} \right)_{\setminus i_1 \dots i_m}(\varsigma) \big|_{\varsigma = \tilde{\theta}^{-1}(\theta)}$$

and

$$f_{\setminus i_1 \dots i_m}(\mathbf{O}; \theta) = \left(f \circ \tilde{\theta} \right)_{\setminus i_1 \dots i_m}(\varsigma) \big|_{\varsigma = \tilde{\theta}^{-1}(\theta)}$$

where each $i_s \in \{1, \dots, p\}$

$$f(\mathbf{O}; \theta) \triangleq \prod_{i=1}^n f(O_i; \theta)$$

Canonical (Hoeffding) Representation of Order 1 and 2 Mean 0 U-stat-

$$U_1(\theta) = \sum_{i \neq j} u_1(O_i), E[u_1(O_i)] = 0 :$$

$u(\cdot, \cdot)$ not necc sym

$$U_2(\theta) = \sum_{i \neq j} u(O_i, O_j), E[u(O_i, O_j)] = 0 :$$

$u(\cdot, \cdot)$ not necc sym

$$U_2(\theta) = \sum_i d(O_i, \theta) + \sum_{i \neq j} m(O_i, O_j),$$

$$E_\theta[d(O_i, \theta)] = 0,$$

$$E[m(O_i, O_j) | O_i] = E[m(O_i, O_j) | O_j] = 0,$$

$m(\cdot, \cdot)$ not necc sym

$\sum_i d(O_i, \theta)$ and $\sum_{i \neq j} m(O_i, O_j)$ uncorr

Canonical Representation of Order 3 Mean 0 U-stat-

$$U_3(\theta) = U_2(\theta) + \sum_{i \neq j \neq X} t(O_i, O_j, O_X)$$

$$\begin{aligned} E[t(O_i, O_j, O_X) | O_i, O_j] &= E[t(O_i, O_j, O_X) | O_i, O_X] \\ &= E[t(O_i, O_j, O_X) | O_j, O_X] = 0, \end{aligned}$$

t(·, ·, ·) not necc sym

$$\sum_{i \neq j \neq X} m(O_i, O_j, O_X) \text{ and } U_2(\theta) \text{ uncorr}$$

Formula for higher order scores associated with $\tilde{\theta}(\varsigma)$

$$S_{i_1 \dots i_m}(\theta) = f_{/i_1 \dots i_m}(\mathbf{O}; \theta) / f(\mathbf{O}; \theta)$$

$$f(\mathbf{O}; \theta) = \prod_{i=1}^n f(O_i; \theta)$$

of order m in terms of the subject specific scores (Waterman and Lindsay (1996)

$$S_{i_1 \dots i_m, j}(\theta) = f_{/i_1 \dots i_m, j}(O_j; \theta) / f_j(O_j; \theta), j = 1, \dots, n$$

(Waterman and Lindsay (1996) .

$$S_{i_1} = \sum_j S_{i_1, j}$$

$$S_{i_1 i_2} = \sum_j S_{i_1 i_2, j} + \sum_{X \neq j} S_{i_1, j} S_{i_2, X}$$

$$S_{i_1 i_2, j}(\theta) = S_{i_1, j}(\theta) S_{i_2, j}(\theta) + \partial S_{i_1, j}(\theta(\varsigma)) / \partial \varsigma_{i_2} |_{\tilde{\theta}(\varsigma) = \theta}$$

$$\begin{aligned}
& S_{i_1 i_2 i_3} \\
&= \sum_j S_{i_1 i_2 i_3, j} + \sum_{X \neq j} S_{i_1 i_2, j} S_{i_3, X} + S_{i_3 i_2, j} S_{i_1, X} + S_{i_1 i_3, j} S_{i_2, X} \\
&\quad \sum_{X \neq j \neq t} S_{i_1, j} S_{i_2, X} S_{i_3, t}
\end{aligned}$$

Definition of a kth order influence function: A U-statistic $U_k(\theta) = u_k(\mathbf{O}; \theta)$ of order k , dimension p and finite variance is said to be an k th order influence function for $\psi(\theta)$ if (i)

$$E_{\theta}[U_k(\theta)] = 0, \theta \in \Theta$$

(ii) for $m = 1, 2, \dots, k$, and every $\tilde{\theta}(\varsigma)$, $p = 1, 2, \dots$

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

$p = k$ sufficient. We say that $\psi(\theta)$ is k th order pathwise differentiable

Theorem: If the model is nonparametric .then there is at most one mth order estimation influence function $IF_m^{est}(\theta)$, the efficient mth order IF.

Lemma: $IF_m^{est}(\theta) = IF_{m-1}^{est}(\theta) + IF_{mm}^{est}(\theta)$,

$IF_{mm}^{est}(\theta) = \sum_{\{i_1 \neq i_2 \neq \dots \neq i_m; i_X \in \{1, 2, \dots, n\}, X \in \{1, \dots, m\}\}} d_m(O_{i_1}, \dots, O_{i_m})$,
where $d_m(O_{i_1}, O_{i_2}, \dots, O_{i_m})$ is canonical

$\text{Var}[IF_m^{est}(\theta)]$ increases with m

$\text{Var}[IF_m^{est}(\theta)] / \text{Var}[IF_m^{est}(\hat{\theta})] = 1 + o(1)$

The following Extended Information Theorem is closely related to result in McLeish and Small (1994).

Theorem: Given $U_k(\theta)$, for all $\tilde{\theta}(\varsigma)$ for $s \leq k$

$$\begin{aligned} & \partial^s E_{\theta} [U_k(\theta(\varsigma))] / \partial \varsigma_{i_1} \dots \partial \varsigma_{i_s} \\ &= -E_{\theta} [U_k(\theta) S_{i_1 \dots i_s}(\theta)] \\ &= -\psi_{\setminus i_1 \dots i_s}(\theta) \end{aligned}$$

$$E_{\theta} [U_k(\hat{\theta})] = -[\psi(\hat{\theta}) - \psi(\theta)] + O_p(\|\hat{\theta} - \theta\|^{k+1})$$

since as functions of $\hat{\theta}$, the functions $E_{\theta} [U_k(\hat{\theta})]$ and $-\left[\psi(\hat{\theta}) - \psi(\theta)\right]$ have the same Taylor expansion around θ up to order k

$$\hat{\psi}_m = \psi(\hat{\theta}) + IF_m(\hat{\theta})$$

where $\hat{\theta}$ is an initial estimator of θ . from a separate sample
(no Donsker like needed).

But, by extended info equality

$$E_{\theta} [IF_m(\hat{\theta})] = -[\psi(\hat{\theta}) - \psi(\theta)] + O_p(\|\hat{\theta} - \theta\|^{m+1})$$

so (conditional) bias of $\hat{\psi}_m$ is

$$\{\psi(\hat{\theta}) + E_{\theta} [IF_m(\hat{\theta})] - \psi(\theta)\} = O_p(\|\hat{\theta} - \theta\|^{m+1}), \downarrow m$$

$\text{Var}[\hat{\psi}_m]$ increases with m

$IF_m(\hat{\theta})$ and $\hat{\psi}_m = \psi(\hat{\theta}) + IF_m(\hat{\theta})$ are AN given $\hat{\theta}$ often normal.

Shortest conservative uniform asymptotic confidence intervals based on

$$\hat{\psi}_{m_{conf}} \pm \text{var} \left[IF_{m_{conf}}(\hat{\theta}) | \hat{\theta} \right]^{1/2} z_\alpha$$

where k_{conf} is the smallest k with $\text{var}[U_k(\theta)]$ higher order (or equal if constants dealt with) than the squared bias.

Example: Problem: If $IF_1(\theta)$ depends on θ through a nonparametric $\rho(\theta)$ where $\rho(\theta)$ infinite dimensional

$IF_m(\theta)$ does not exist for $m \geq 2$

Example:

$$IF_1(\hat{\theta}) = n^{-1} \sum_i \frac{A_i}{\hat{p}(X_i)} (Y_1 - \hat{b}(X_i)) + \hat{b}(X_i) - \psi(\hat{\theta})$$

Use sieves ie $k = k(n)$ dimensional submodels for $b(X)$, $p(X_i)$. Then IF_m exists for all m .

But then truncation bias

$$TB_k = E \left[\left\{ b(X_i) - \bar{b}_k(X_i) \right\} \left\{ p(X_i) - \bar{p}_k(X_i) \right\} / \bar{p}_k(X_i) \right]$$

added to estimation bias $\left\| \hat{\theta} - \theta \right\|^{m+1}$

where $\bar{b}_k(X_i)$ is the limit of the model

Specifically $b(X) = b^* \left(Z_k^T B_k \right), p(X) = p^* \left(Z_k^T \alpha_k \right)$

where Z_k a basis in A^d as $k \rightarrow \infty$