Minimax Estimation of a nonlinear functional on a structured high-dimensional model

Eric Tchetgen Tchetgen

Professor of Biostatistics and Epidemiologic Methods, Harvard U.

Outline

- Heuristics

Outline

- Heuristics
- First order Influence functions: $\sqrt{n}$ -regular case

Outline

- Heuristics
- First order Influence functions: $\sqrt{n}$ -regular case
- Higher order influence functions non-regular rates $\ll \sqrt{n}$

Outline

- Heuristics
- First order Influence functions: $\sqrt{n}$ -regular case
- Higher order influence functions non-regular rates $\ll \sqrt{n}$
- Application to quadratic functional and missing at random data functional

Outline

- Heuristics
- First order Influence functions: $\sqrt{n}$ -regular case
- Higher order influence functions non-regular rates $\ll \sqrt{n}$
- Application to quadratic functional and missing at random data functional
- This is joint work with Robins, Li, Mukherjee and van der Vaart.

Heuristics

- $X_1, ..., X_n \overset{i.i.d}{\sim} p$ where p belongs to a collection $\mathcal{P}$ of densities and we aim to estimate the value $\chi(p)$ of a functional

$$\chi : \mathcal{P} \to \mathbb{R}$$

Heuristics

- $X_1, ..., X_n \overset{i.i.d}{\sim} p$ where p belongs to a collection $\mathcal{P}$ of densities and we aim to estimate the value $\chi(p)$ of a functional

$$\chi : \mathcal{P} \rightarrow \mathbb{R}$$

- We are particularly interested in settings of semiparametric or nonparametric model such that $\mathcal{P}$ is infinite dimensional.

Heuristics

- $X_1, ..., X_n \overset{i.i.d}{\sim} p$ where p belongs to a collection $\mathcal{P}$ of densities and we aim to estimate the value $\chi(p)$ of a functional

$$\chi : \mathcal{P} \rightarrow \mathbb{R}$$

- We are particularly interested in settings of semiparametric or nonparametric model such that $\mathcal{P}$ is infinite dimensional.
- We will give special attention to setting where the semiparametric model is described through parameters of low regularity, in which case $\sqrt{n}$ convergence may not be attainable.

Heuristics

- For estimation, consider the class of "one-step" estimators of the form

$$\hat{\chi} = \chi\left(\hat{p}_n\right) + \mathbb{P}_n \chi_{\hat{p}_n}$$

for $\hat{p}_n$ an initial estimator of p and $x \longrightarrow \chi_p\left(x\right)$ is a measurable function for each p and $\mathbb{P}_n f = n^{-1} \sum_{i=1}^{n} f\left(X_i\right)$.

Heuristics

- For estimation, consider the class of "one-step" estimators of the form

$$\hat{\chi} = \chi\left(\hat{p}_n\right) + \mathbb{P}_n \chi_{\hat{p}_n}$$

for $\hat{p}_n$ an initial estimator of p and $x \longrightarrow \chi_p(x)$ is a measurable function for each p and $\mathbb{P}_n f = n^{-1} \sum_{i=1}^{n} f(X_i)$.

- One possible choice for $\chi_{\hat{p}_n} = 0$ leading to the plug-in estimator, this is typically suboptimal.

Heuristics

- For estimation, consider the class of "one-step" estimators of the form

$$\hat{\chi} = \chi\left(\hat{p}_n\right) + \mathbb{P}_n \chi_{\hat{p}_n}$$

  for $\hat{p}_n$ an initial estimator of p and $x \longrightarrow \chi_p(x)$ is a measurable function for each p and $\mathbb{P}_n f = n^{-1} \sum_{i=1}^{n} f(X_i)$.
- One possible choice for $\chi_{\hat{p}_n} = 0$ leading to the plug-in estimator, this is typically suboptimal.
- $\hat{p}_n$ constructed from an independent sample, i.e. sample splitting. This is not required in the regular regime through use of empirical process theory. Such theory does not apply in non-regular regime.

Heuristics

- To motivate a good choice of $\chi_{\hat{p}_n}$ consider

$$\hat{\chi} - \chi(p) = \left[\chi(\hat{p}_n) - \chi(p) + P\chi_{\hat{p}_n}\right] + (\mathbb{P}_n - P)\chi_{\hat{p}_n}$$

where $P\chi_{\hat{p}_n} = \int \chi_{\hat{p}_n}(x)\, dP(x)$.

Heuristics

- To motivate a good choice of $\chi_{\hat{p}_n}$ consider

$$\hat{\chi} - \chi(p) = \left[\chi(\hat{p}_n) - \chi(p) + P\chi_{\hat{p}_n}\right] + (\mathbb{P}_n - P)\chi_{\hat{p}_n}$$

  where $P\chi_{\hat{p}_n} = \int \chi_{\hat{p}_n}(x)\, dP(x)$.
- The second term is standard normal, a good choice of $\chi_{\hat{p}_n}$ is to make sure that the term in brackets has "small bias", i.e. no larger than $O_P\left(n^{-1/2}\right)$

Heuristics

- To motivate a good choice of $\chi_{\hat{p}_n}$ consider

$$\hat{\chi} - \chi(p) = \left[\chi(\hat{p}_n) - \chi(p) + P\chi_{\hat{p}_n}\right] + (\mathbb{P}_n - P)\chi_{\hat{p}_n}$$

  where $P\chi_{\hat{p}_n} = \int \chi_{\hat{p}_n}(x)\, dP(x)$.
- The second term is standard normal, a good choice of $\chi_{\hat{p}_n}$ is to make sure that the term in brackets has "small bias", i.e. no larger than $O_P\left(n^{-1/2}\right)$
- This can be achieved if $P\chi_{\hat{p}_n}$ acts like minus the derivative of the functional χ in the $(\hat{p}_n - p)$ direction.

Heuristics

- To motivate a good choice of $\chi_{\hat{p}_n}$ consider

$$\hat{\chi} - \chi(p) = \left[\chi(\hat{p}_n) - \chi(p) + P\chi_{\hat{p}_n}\right] + (\mathbb{P}_n - P)\chi_{\hat{p}_n}$$

  where $P\chi_{\hat{p}_n} = \int \chi_{\hat{p}_n}(x)\, dP(x)$.
- The second term is standard normal, a good choice of $\chi_{\hat{p}_n}$ is to make sure that the term in brackets has "small bias", i.e. no larger than $O_P\left(n^{-1/2}\right)$
- This can be achieved if $P\chi_{\hat{p}_n}$ acts like minus the derivative of the functional χ in the $(\hat{p}_n - p)$ direction.
- A function $\chi_{\hat{p}_n}$ with this property is known as a "first order" influence function.

Heuristics

- For a 1st order influence function the bias term is quadratic in the error $d(\hat{p}_n, p)$ for an appropriate distance d

Heuristics

- For a 1st order influence function the bias term is quadratic in the error $d(\hat{p}_n, p)$ for an appropriate distance d
- no bias condition essentially requires $d(\hat{p}_n, p) = o_p(n^{-1/4})$. To get this rate the model cannot be too large

Heuristics

- For a 1st order influence function the bias term is quadratic in the error $d(\hat{p}_n, p)$ for an appropriate distance d
- no bias condition essentially requires $d(\hat{p}_n, p) = o_p(n^{-1/4})$. To get this rate the model cannot be too large
- For instance regression or density can be estimated at rate $n^{-1/4}$ if a-priori known to have at least $d/2$ derivatives

Heuristics

- For a 1st order influence function the bias term is quadratic in the error $d(\hat{p}_n, p)$ for an appropriate distance d
- no bias condition essentially requires $d(\hat{p}_n, p) = o_p(n^{-1/4})$. To get this rate the model cannot be too large
- For instance regression or density can be estimated at rate $n^{-1/4}$ if a-priori known to have at least $d/2$ derivatives
- Our main concern will be for settings where the model is very large or has very low regularity, such that $n^{-1/4}$ is not attainable for $d(\hat{p}_n, p)$.

Heuristics

- The one step estimator $\hat{\chi} = \chi(\hat{p}_n) + \mathbb{P}_n \chi_{\hat{p}_n}$ is suboptimal in such cases because it does not strike the right balance between bias and variance; that is

$$\hat{\chi} - \chi(p) = \left[\chi(\hat{p}_n) - \chi(p) + P\chi_{\hat{p}_n}\right] + (\mathbb{P}_n - P)\chi_{\hat{p}_n}$$

the magnitude of the first term dominates and one cannot obtain valid inferences about $\chi(p)$, i.e. CIs not centered properly to achieve nominal coverage.

Heuristics

- The one step estimator $\hat{\chi} = \chi(\hat{p}_n) + \mathbb{P}_n \chi_{\hat{p}_n}$ is suboptimal in such cases because it does not strike the right balance between bias and variance; that is

$$\hat{\chi} - \chi(p) = \left[\chi(\hat{p}_n) - \chi(p) + P\chi_{\hat{p}_n}\right] + (\mathbb{P}_n - P)\chi_{\hat{p}_n}$$

  the magnitude of the first term dominates and one cannot obtain valid inferences about $\chi(p)$, i.e. CIs not centered properly to achieve nominal coverage.
- We will replace the first order IF $\mathbb{P}_n \chi_{\hat{p}_n}$ by a higher order IF $\mathbb{U}_n \chi_{\hat{p}_n}$ which is an *mth* order U-statistic, such that

$$\hat{\chi}_n = \chi(\hat{p}_n) + \mathbb{U}_n \chi_{\hat{p}_n}$$

  and

$$\hat{\chi}_n - \chi(p) = \left[\chi(\hat{p}_n) - \chi(p) + P^m \chi_{\hat{p}_n}\right] + (\mathbb{U}_n - P^m)\chi_{\hat{p}_n}$$

Heuristics

- The one step estimator $\hat{\chi} = \chi(\hat{p}_n) + \mathbb{P}_n \chi_{\hat{p}_n}$ is suboptimal in such cases because it does not strike the right balance between bias and variance; that is

$$\hat{\chi} - \chi(p) = \left[\chi(\hat{p}_n) - \chi(p) + P\chi_{\hat{p}_n}\right] + (\mathbb{P}_n - P)\chi_{\hat{p}_n}$$

  the magnitude of the first term dominates and one cannot obtain valid inferences about $\chi(p)$, i.e. CIs not centered properly to achieve nominal coverage.
- We will replace the first order IF $\mathbb{P}_n \chi_{\hat{p}_n}$ by a higher order IF $\mathbb{U}_n \chi_{\hat{p}_n}$ which is an *mth* order U-statistic, such that

$$\hat{\chi}_n = \chi(\hat{p}_n) + \mathbb{U}_n \chi_{\hat{p}_n}$$

  and

$$\hat{\chi}_n - \chi(p) = \left[\chi(\hat{p}_n) - \chi(p) + P^m \chi_{\hat{p}_n}\right] + (\mathbb{U}_n - P^m)\chi_{\hat{p}_n}$$

- This in fact suggest choosing the HOIF such that $-P^m \chi_{\hat{p}_n}$ behaves like the first m terms of the Taylor expansion of $\chi(\hat{p}_n) - \chi(p)$.

Heuristics

- Exact HOIFs exist only for special functionals.

Heuristics

- Exact HOIFs exist only for special functionals.
- For general functionals, we find a "good approximate" functional in which case the convergence rate obtained by trading-off estimation bias+approximation bias with variance.

Heuristics

- Exact HOIFs exist only for special functionals.
- For general functionals, we find a "good approximate" functional in which case the convergence rate obtained by trading-off estimation bias+approximation bias with variance.
- In some cases the optimal rate of the HOIF may still be $n^{1/2}$ in which case the semiparametric efficiency bound may be achieved, the first order IF determines the variance and HOIFs correct the bias.

Heuristics

- Exact HOIFs exist only for special functionals.
- For general functionals, we find a "good approximate" functional in which case the convergence rate obtained by trading-off estimation bias+approximation bias with variance.
- In some cases the optimal rate of the HOIF may still be $n^{1/2}$ in which case the semiparametric efficiency bound may be achieved, the first order IF determines the variance and HOIFs correct the bias.
- Typically the rate will be slower than $n^{1/2}$. What is the optimal (i.e. minimax) rate of estimation? can we achieve it?

Heuristics

- Exact HOIFs exist only for special functionals.
- For general functionals, we find a "good approximate" functional in which case the convergence rate obtained by trading-off estimation bias+approximation bias with variance.
- In some cases the optimal rate of the HOIF may still be $n^{1/2}$ in which case the semiparametric efficiency bound may be achieved, the first order IF determines the variance and HOIFs correct the bias.
- Typically the rate will be slower than $n^{1/2}$. What is the optimal (i.e. minimax) rate of estimation? can we achieve it?
- Two cases emerge:

Heuristics

- Exact HOIFs exist only for special functionals.
- For general functionals, we find a "good approximate" functional in which case the convergence rate obtained by trading-off estimation bias+approximation bias with variance.
- In some cases the optimal rate of the HOIF may still be $n^{1/2}$ in which case the semiparametric efficiency bound may be achieved, the first order IF determines the variance and HOIFs correct the bias.
- Typically the rate will be slower than $n^{1/2}$. What is the optimal (i.e. minimax) rate of estimation? can we achieve it?
- Two cases emerge:
  - case 1:smoothness of p is assumed known, we have obtained both lower and upper bounds

Heuristics

- Exact HOIFs exist only for special functionals.
- For general functionals, we find a "good approximate" functional in which case the convergence rate obtained by trading-off estimation bias+approximation bias with variance.
- In some cases the optimal rate of the HOIF may still be $n^{1/2}$ in which case the semiparametric efficiency bound may be achieved, the first order IF determines the variance and HOIFs correct the bias.
- Typically the rate will be slower than $n^{1/2}$. What is the optimal (i.e. minimax) rate of estimation? can we achieve it?
- Two cases emerge:
  - case 1:smoothness of p is assumed known, we have obtained both lower and upper bounds
  - case 2: smoothness of p is not known and therefore one must adapt to it; some recent results

Example 1: Quadratic density functional

- Quadratic functional of a density $p(X)$:

$$\begin{aligned} \chi &: p \to \chi(p) = \int_{\mathbb{R}^p} p(x)^2 \, dx \\ p &\in H^d(\alpha, C) \text{ is in a d-dimensional Hölder ball} \\ &\quad \text{with smoothness } \alpha \text{ and radius } C. \end{aligned}$$

Example 1: Quadratic density functional

- Quadratic functional of a density $p(X)$:

$$\begin{aligned} \chi &: p \to \chi(p) = \int_{\mathbb{R}^p} p(x)^2 \, dx \\ p &\in H^d(\alpha, C) \text{ is in a d-dimensional Hölder ball} \\ &\quad \text{with smoothness } \alpha \text{ and radius } C. \end{aligned}$$

- Of interest in bandwidth selection for optimal density smoothing.

Example 1: Quadratic density functional

- Quadratic functional of a density $p(X)$:

$$\begin{aligned} \chi &: p \rightarrow \chi(p) = \int_{\mathbb{R}^p} p(x)^2 \, dx \\ p &\in H^d(\alpha, C) \text{ is in a d-dimensional Hölder ball} \\ &\quad \text{with smoothness } \alpha \text{ and radius } C. \end{aligned}$$

- Of interest in bandwidth selection for optimal density smoothing.
- Well studied problem e.g Bickel and Ritov (1988), Laurent (1996, 1997), Cai and Low (2005), Cai, Low et al (2006), Donoho and Nussbaum (1990), Efromovitch and Low et al (1994, 1996), Giné and Nickl (2008), Laurent and Massart (2000).

Example 1: Quadratic density functional

- Quadratic functional of a density $p(X)$:

$$\begin{aligned} \chi &: p \to \chi(p) = \int_{\mathbb{R}^p} p(x)^2 \, dx \\ p &\in H^d(\alpha, C) \text{ is in a d-dimensional Hölder ball} \\ &\quad \text{with smoothness } \alpha \text{ and radius } C. \end{aligned}$$

- Of interest in bandwidth selection for optimal density smoothing.
- Well studied problem e.g Bickel and Ritov (1988), Laurent (1996, 1997), Cai and Low (2005), Cai, Low et al (2006), Donoho and Nussbaum (1990), Efromovitch and Low et al (1994, 1996), Giné and Nickl (2008), Laurent and Massart (2000).
- Exhibits elbow phenomenon in that $n^{1/2}$-estimable only if $\alpha/d \geq 1/4$. Below the threshold, minimax lower bound is $n^{-\frac{8\alpha}{d+4\alpha}}$ in squared error norm. (Birgé and Massart, 1995)

Example 2:Nonlinear density functional

- nonlinear functional of a density $p(X)$:

$$\chi \quad : \quad p \rightarrow \chi(p) = \int_{\mathbb{R}^p} T(p(x))\, dx$$

where T is a smooth mapping

$p \in H^d(\alpha, C)$ is in a d-dimensional Hölder ball with smoothness α and radius C.

Example 2:Nonlinear density functional

- nonlinear functional of a density $p(X)$:

$$\chi \quad : \quad p \to \chi(p) = \int_{\mathbb{R}^p} T(p(x))\, dx$$

where T is a smooth mapping

$$p \quad \in \quad H^d(\alpha, C)$$ is in a d-dimensional Hölder ball with smoothness α and radius C.

- Large body on estimating the entropy of an underlying distribution. Beirlant et al. (1997), more recent works include estimation of Renyi and Tsallis entropies (Leonenko and Seleznjev, 2010; Pál, Póczos and Szepesvári, 2010). Also see Kandasamy et al. (2014).

Example 2:Nonlinear density functional

- nonlinear functional of a density $p(X)$:

$$\chi \quad : \quad p \rightarrow \chi(p) = \int_{\mathbb{R}^p} T(p(x))\, dx$$

where T is a smooth mapping

$$p \quad \in \quad H^d(\alpha, C)$$ is in a d-dimensional Hölder ball with smoothness α and radius C.

- Large body on estimating the entropy of an underlying distribution. Beirlant et al. (1997), more recent works include estimation of Renyi and Tsallis entropies (Leonenko and Seleznjev, 2010; Pál, Póczos and Szepesvári, 2010). Also see Kandasamy et al. (2014).
- Exhibit elbow phenomenon in that $n^{1/2}$-estimable only if $\alpha/d \geq 1/4$. Below the threshold, minimax lower bound is $n^{-\frac{8\alpha}{d+4\alpha}}$ in squared error norm, Birgé and Massart,1995).

Example 2:Nonlinear density functional

- nonlinear functional of a density $p(X)$:

$$\chi \quad : \quad p \rightarrow \chi(p) = \int_{\mathbb{R}^p} T(p(x))\, dx$$

where T is a smooth mapping

$$p \quad \in \quad H^d(\alpha, C)$$ is in a d-dimensional Hölder ball with smoothness α and radius C.

- Large body on estimating the entropy of an underlying distribution. Beirlant et al. (1997), more recent works include estimation of Renyi and Tsallis entropies (Leonenko and Seleznjev, 2010; Pál, Póczos and Szepesvári, 2010). Also see Kandasamy et al. (2014).
- Exhibit elbow phenomenon in that $n^{1/2}$-estimable only if $\alpha/d \geq 1/4$. Below the threshold, minimax lower bound is $n^{-\frac{8\alpha}{d+4\alpha}}$ in squared error norm, Birgé and Massart,1995).
- The case $T(p) = p^3$ studied when $d = 1$ by Kerkyacharian, Picard and Tsybakov (1998), general solution for $d > 1$ given by Tchetgen et al (2008).

Example 3:Missing at random data

- Suppose full data is (Y, Z) where Y is a binary response and Z a d-dimensional vector of covariates;however one observes (YA, A, Z) from which we wish to estimate $\chi = E(Y) = \Pr(Y = 1)$

Example 3:Missing at random data

- Suppose full data is (Y, Z) where Y is a binary response and Z a d-dimensional vector of covariates;however one observes (YA, A, Z) from which we wish to estimate $\chi = E(Y) = \Pr(Y = 1)$
- χ is nonparametrically identified under MAR $A \amalg Y | Z$ provided $\Pr(A = 1|Z) > 0$ a.s.

Example 3:Missing at random data

- Suppose full data is (Y, Z) where Y is a binary response and Z a d-dimensional vector of covariates;however one observes (YA, A, Z) from which we wish to estimate $\chi = E(Y) = \Pr(Y = 1)$
- χ is nonparametrically identified under MAR $A \amalg Y|Z$ provided $\Pr(A = 1|Z) > 0$ a.s.
- Note: isomorphic to estimating the average treatment effect of of a binary intervention A on Y under the assumption of no unmeasured confounding given Z.*i.e.* $\chi = E(Y_a)$ where Y_a is the counterfactual response under regime a, full data is (Y_a, X), observed data is (Y_a, X) if $A = a$ and X otherwise. Identified under $A \amalg Y_a|Z$

Example 3:Missing at random data

- The functional can be expressed as an observed data functional in terms of the density f of Z (relative to a measure ν), the probability $b(z) = \Pr(Y = 1|Z = z) = \Pr(Y = 1|Z = z, A = 1)$:

$$\chi : p_{b,f} \rightarrow \chi\left(p_{b,f}\right) = \int bf d\nu$$

Example 3:Missing at random data

- The functional can be expressed as an observed data functional in terms of the density f of Z (relative to a measure ν), the probability $b(z) = \Pr(Y = 1|Z = z) = \Pr(Y = 1|Z = z, A = 1)$:

$$\chi : p_{b,f} \rightarrow \chi\left(p_{b,f}\right) = \int bf d\nu$$

- Alternative parametrization in terms of $a(z)^{-1} = \Pr(A = 1|Z = z)$ and $g = f/a$ (proportional to the density of Z given $A = 1$)

$$\chi : p_{a,b,g} \rightarrow \chi\left(p_{a,b,g}\right) = \int abgd\nu$$

Example 3:Missing at random data

- Estimators of $\chi(p)$ that are $n^{1/2}-$consistent and asymptotically efficient in the semiparametric sense have been constructed w a variety of methods, but only if a or b or both parameters are restricted to sufficiently small regularity classes.

Example 3:Missing at random data

- Estimators of $\chi(p)$ that are $n^{1/2}-$consistent and asymptotically efficient in the semiparametric sense have been constructed w a variety of methods, but only if a or b or both parameters are restricted to sufficiently small regularity classes.
- Formally, suppose that a and b belong to $H^d(\alpha, C_\alpha)$ and $H^d(\beta, C_\beta)$ respectively, then $n^{1/2}-$consistent estimators have been obtained for α and β large enough that

$$\frac{\alpha}{2\alpha + d} + \frac{\beta}{2\beta + d} \geq \frac{1}{2} \tag{1}$$

Example 3:Missing at random data

- Estimators of $\chi(p)$ that are $n^{1/2}-$consistent and asymptotically efficient in the semiparametric sense have been constructed w a variety of methods, but only if a or b or both parameters are restricted to sufficiently small regularity classes.
- Formally, suppose that a and b belong to $H^d(\alpha, C_\alpha)$ and $H^d(\beta, C_\beta)$ respectively, then $n^{1/2}-$consistent estimators have been obtained for α and β large enough that

$$\frac{\alpha}{2\alpha + d} + \frac{\beta}{2\beta + d} \geq \frac{1}{2} \tag{1}$$

- In our work, we show that (1) is more stringent than required to achieve root-n-consistency which can be attained using HOIF provided

$$\frac{\alpha + \beta}{2d} \geq \frac{1}{4}$$

Example 3:Missing at random data

- For moderate to large dimensions d this is still a restrictive requirement, likewise if regularity is low even if d is small.

Example 3:Missing at random data

- For moderate to large dimensions d this is still a restrictive requirement, likewise if regularity is low even if d is small.
- In fact in Robins, Li, Tchetgen, van der Vaart (2009) we establish that even if g were known, the minimax lower bound when $(\alpha+\beta)/2d < \frac{1}{4}$ is

$$n^{-\frac{2(\alpha+\beta)}{2(\alpha+\beta)+d}}$$

Example 3:Missing at random data

- For moderate to large dimensions d this is still a restrictive requirement, likewise if regularity is low even if d is small.
- In fact in Robins, Li, Tchetgen, van der Vaart (2009) we establish that even if g were known, the minimax lower bound when $(\alpha+\beta)/2d < \frac{1}{4}$ is

$$n^{-\frac{2(\alpha+\beta)}{2(\alpha+\beta)+d}}$$

- Our estimators are shown to achieve this minimax rate when α and β are apriori known, provided g is smooth enough.

First Order Semiparametric Theory

- Theory which goes back to Pfanzagl (1982), Newey (1990), van der Vaart (1991), Bickel et al (1993)

First Order Semiparametric Theory

- Theory which goes back to Pfanzagl (1982), Newey (1990), van der Vaart (1991), Bickel et al (1993)
- Given a functional $\chi : \mathcal{P} \rightarrow \mathbb{R}$ defined on the semiparametric model $\mathcal{P}$, an IF is a function $\chi_p : x \rightarrow \chi_p(x)$ of the observed data which satisfies the following equation.

First Order Semiparametric Theory

- Theory which goes back to Pfanzagl (1982), Newey (1990), van der Vaart (1991), Bickel et al (1993)
- Given a functional $\chi : \mathcal{P} \rightarrow \mathbb{R}$ defined on the semiparametric model $\mathcal{P}$, an IF is a function $\chi_p : x \rightarrow \chi_p(x)$ of the observed data which satisfies the following equation.
- For a sufficiently regular submodel $t \rightarrow p_t \subset \mathcal{P}$

$$\frac{d}{dt}\Big|_{t=0} \chi(p_t) = \frac{d}{dt}\Big|_{t=0} P_t \chi_p = P(\chi_p g)$$

where $g = (d/dt)\big|_{t=0} p_t/p$ is the score function of the model $t \rightarrow p_t$ at $t = 0$. The closed linear span of all scored attached to submodels $t \rightarrow p_t$ of $\mathcal{P}$ is called the tangent space.

First Order Semiparametric Theory

- Theory which goes back to Pfanzagl (1982), Newey (1990), van der Vaart (1991), Bickel et al (1993)
- Given a functional $\chi : \mathcal{P} \rightarrow \mathbb{R}$ defined on the semiparametric model $\mathcal{P}$, an IF is a function $\chi_p : x \rightarrow \chi_p(x)$ of the observed data which satisfies the following equation.
- For a sufficiently regular submodel $t \rightarrow p_t \subset \mathcal{P}$

$$\frac{d}{dt}\Big|_{t=0} \chi(p_t) = \frac{d}{dt}\Big|_{t=0} P_t \chi_p = P(\chi_p g)$$

where $g = (d/dt)\big|_{t=0} p_t / p$ is the score function of the model $t \rightarrow p_t$ at $t = 0$. The closed linear span of all scored attached to submodels $t \rightarrow p_t$ of $\mathcal{P}$ is called the tangent space.

- Therefore an influence function is an element of the Hilbert space $L_2(p)$ whose inner product with elements of the tangent space represent the derivative of the functional. This result is known in functional analysis as Riesz Representation Theorem.

IF of quadratic density functional

- To compute the IF of $\chi(p) = \int_{\mathbb{R}^p} p(x)^2 dx$ consider $\chi(p_t) = \int_{\mathbb{R}^p} p_t(x)^2 dx$ and computing $\frac{d}{dt}\big|_{t=0} \chi(p_t)$ one obtains

$$\frac{d}{dt}\Big|_{t=0} \chi(p_t) = P\{2(p(X) - \chi(p))g\}$$

Therefore

$$\chi_p = 2(p(X) - \chi(p))$$

IF of quadratic density functional

- To compute the IF of $\chi(p) = \int_{\mathbb{R}^p} p(x)^2 dx$ consider $\chi(p_t) = \int_{\mathbb{R}^p} p_t(x)^2 dx$ and computing $\frac{d}{dt}\big|_{t=0} \chi(p_t)$ one obtains

$$\frac{d}{dt}\Big|_{t=0} \chi(p_t) = P\{2(p(X) - \chi(p))g\}$$

Therefore

$$\chi_p = 2(p(X) - \chi(p))$$

- Furthermore $\hat{\chi} = \chi(\hat{p}_n) + \mathbb{P}_n \chi^{(1)}_{\hat{p}_n} = 2\mathbb{P}_n \hat{p}_n(X) - \chi(\hat{p}_n)$ so that

$$P(\hat{\chi} - \chi(p)) = \int (p - \hat{p}_n)^2(x)\, dx$$

which is quadratic in the preliminary estimator as expected.

IF of quadratic density functional

- To compute the IF of $\chi(p) = \int_{\mathbb{R}^p} p(x)^2 dx$ consider $\chi(p_t) = \int_{\mathbb{R}^p} p_t(x)^2 dx$ and computing $\frac{d}{dt}\big|_{t=0} \chi(p_t)$ one obtains

$$\frac{d}{dt}\Big|_{t=0} \chi(p_t) = P\{2(p(X) - \chi(p))g\}$$

Therefore

$$\chi_p = 2(p(X) - \chi(p))$$

- Furthermore $\hat{\chi} = \chi(\hat{p}_n) + \mathbb{P}_n \chi^{(1)}_{\hat{p}_n} = 2\mathbb{P}_n \hat{p}_n(X) - \chi(\hat{p}_n)$ so that

$$P(\hat{\chi} - \chi(p)) = \int (p - \hat{p}_n)^2(x)\, dx$$

which is quadratic in the preliminary estimator as expected.

- The variance of $\hat{\chi}$ is of order $1/n$; if the preliminary estimator $\hat{p}_n$ can be constructed so that the squared bias is smaller than the variance, the estimator is rate optimal; otherwise the bias dominates and a higher order IF is needed.

Missing Data IF

- To compute IF in the MAR example, consider the score functions of the one dimensional submodel $p_t = p_{a_t,b_t,f_t}$ induced by paths of the form

$$\begin{aligned} a_t &= a + t\underline{\mathfrak{a}} \\ b_t &= b + t\underline{\mathfrak{b}} \\ f_t &= f + t\underline{\mathfrak{f}} \end{aligned}$$

given measurable functions $\underline{\mathfrak{a}}, \underline{\mathfrak{b}}, \underline{\mathfrak{f}} : \mathcal{Z} \to \mathbb{R}$.

Missing Data IF

- To compute IF in the MAR example, consider the score functions of the one dimensional submodel $p_t = p_{a_t,b_t,f_t}$ induced by paths of the form

$$\begin{aligned} a_t &= a + t\underline{\mathfrak{a}} \\ b_t &= b + t\underline{\mathfrak{b}} \\ f_t &= f + t\underline{\mathfrak{f}} \end{aligned}$$

  given measurable functions $\underline{\mathfrak{a}}, \underline{\mathfrak{b}}, \underline{\mathfrak{f}} : \mathcal{Z} \to \mathbb{R}$.
- This gives rise to corresponding score equations

$$\frac{Aa(Z) - 1}{a(Z)(a-1)(Z)}\underline{\mathfrak{a}}(Z) \qquad a - \text{score}$$

$$\frac{A(Y - b(Z))}{b(Z)(1-b)(Z)}\underline{\mathfrak{b}}(Z) \qquad b - \text{score}$$

$$\underline{\mathfrak{f}}(Z) \qquad f - \text{score}$$

IF for Missing Data Functional

- Let $\mathfrak{B} = \ a-\text{score}+ \ b-\text{score}+f-\text{score}$. Then,

$$\frac{d}{dt|}_{t=0} \chi_p\left(p_t\right) = P\left(\chi_p^{(1)}\mathfrak{B}\right)$$

for all $\mathfrak{B}$ in the tangent space of the model $\mathcal{P}$, where

$$\chi_p^{(1)} = Aa\left(Z\right)\left(Y - b\left(Z\right)\right) + b\left(Z\right) - \chi\left(p\right)$$

IF for Missing Data Functional

- Let $\mathfrak{B} = \ a-\text{score}+ \ b-\text{score}+f-\text{score}$. Then,

$$\frac{d}{dt}\Big|_{t=0} \chi_p(p_t) = P\left(\chi_p^{(1)}\mathfrak{B}\right)$$

for all $\mathfrak{B}$ in the tangent space of the model $\mathcal{P}$, where

$$\chi_p^{(1)} = Aa(Z)(Y - b(Z)) + b(Z) - \chi(p)$$

- This is the well-known doubly robust influence function due to Robins, Rotnitzky and Zhao (1994).

IF for Missing Data Functional

- Suppose one has obtained rate optimal nonparametric estimators $\widehat{a}$, $\widehat{b}$ (optimal kernel smoothing or series estimation),e.g $\widehat{a}$ converges at rate $n^{-\frac{2\alpha}{2\alpha+d}}$ in mean squared error. Then, the one-step estimator

$$\begin{aligned}\hat{\chi} &= \chi\left(\hat{p}_n\right) + \mathbb{P}_n \chi^{(1)}_{\hat{p}_n} \\ &= \mathbb{P}_n A\widehat{a}\left(Z\right)\left(Y - \widehat{b}\left(Z\right)\right) + \widehat{b}\left(Z\right)\end{aligned}$$

has 2nd order bias

$$\left|E\left(\hat{\chi} - \chi\left(p\right)\right)\right| \lesssim ||\widehat{a} - a||_2 ||\widehat{b} - b||_2$$

which is quadratic in the preliminary estimators as expected. Note that $\widehat{f}$ does not enter the bias (provided the empirical measure is used).

IF for Missing Data Functional

- Suppose one has obtained rate optimal nonparametric estimators $\widehat{a}, \widehat{b}$ (optimal kernel smoothing or series estimation),e.g $\widehat{a}$ converges at rate $n^{-\frac{2\alpha}{2\alpha+d}}$ in mean squared error. Then, the one-step estimator

$$\begin{aligned}\hat{\chi} &= \chi(\hat{p}_n) + \mathbb{P}_n \chi^{(1)}_{\hat{p}_n} \\ &= \mathbb{P}_n A\widehat{a}(Z)\left(Y - \widehat{b}(Z)\right) + \widehat{b}(Z)\end{aligned}$$

has 2nd order bias

$$|E(\hat{\chi} - \chi(p))| \lesssim ||\widehat{a} - a||_2 ||\widehat{b} - b||_2$$

which is quadratic in the preliminary estimators as expected. Note that $\widehat{f}$ does not enter the bias (provided the empirical measure is used).

- The variance of $\hat{\chi}$ is of order $1/n$. If the preliminary estimators $\widehat{a}$ and $\widehat{b}$ can be constructed so that the squared bias is smaller than the variance, the estimator is rate optimal; otherwise the bias dominates and a higher order IF is needed.

HOIFs

- Given a collection of smooth submodels $t \longrightarrow p_t$ in $\mathcal{P}$ an *mth* order influence function χ_{p_t} of the functional $\chi(p)$ if it satisfies

$$\begin{aligned} P^m \chi_p &= 0 \text{ and} \\ \frac{d^j}{dt^j_{|t=0}} \chi(p_t) &= -\frac{d^j}{dt^j_{|t=0}} P^m \chi_{p_t} \\ &= \frac{d^j}{dt^j_{|t=0}} P_t^m \chi_p \qquad j = 1, ..., m \end{aligned}$$

for every submodel through p in $\mathcal{P}$.

HOIFs

- Given a collection of smooth submodels $t \longrightarrow p_t$ in $\mathcal{P}$ an *mth* order influence function χ_{p_t} of the functional $\chi(p)$ if it satisfies

$$\begin{aligned} P^m \chi_p &= 0 \text{ and} \\ \frac{d^j}{dt^j_{|t=0}} \chi(p_t) &= -\frac{d^j}{dt^j_{|t=0}} P^m \chi_{p_t} \\ &= \frac{d^j}{dt^j_{|t=0}} P_t^m \chi_p \qquad j = 1, ..., m \end{aligned}$$

  for every submodel through p in $\mathcal{P}$.

- The third equation implies a Taylor expansion of $t \longrightarrow \chi(p_t)$ at $t = 0$ of order m. but in addition requires that the derivatives of this map can be represented as expectations involving a function χ_p.

HOIFs

- Given a collection of smooth submodels $t \longrightarrow p_t$ in $\mathcal{P}$ an *mth* order influence function χ_{p_t} of the functional $\chi(p)$ if it satisfies

$$\begin{aligned} P^m \chi_p &= 0 \text{ and} \\ \frac{d^j}{dt^j_{|t=0}} \chi(p_t) &= -\frac{d^j}{dt^j_{|t=0}} P^m \chi_{p_t} \\ &= \frac{d^j}{dt^j_{|t=0}} P_t^m \chi_p \qquad j = 1, ..., m \end{aligned}$$

  for every submodel through p in $\mathcal{P}$.
- The third equation implies a Taylor expansion of $t \longrightarrow \chi(p_t)$ at $t = 0$ of order m. but in addition requires that the derivatives of this map can be represented as expectations involving a function χ_p.
- In order to exploit derivatives up to the *mth* order, groups of m observations can be used to match the expectation P^m, this leads to U statistics of order m.

HOIFs

- Thus, we must solve for an *mth* order U-statistic with mean zero that satisfies

$$\frac{d^j}{dt^j_{|t=0}}\chi\left(p_t\right) = \frac{d^j}{dt^j_{|t=0}}P_t^m\chi_p \qquad j = 1, ..., m$$

for every submodel through p in $\mathcal{P}$.

HOIFs

- Thus, we must solve for an *mth* order U-statistic with mean zero that satisfies

$$\frac{d^j}{dt^j_{|t=0}}\chi\left(p_t\right) = \frac{d^j}{dt^j_{|t=0}}P_t^m\chi_p \qquad j = 1, ..., m$$

  for every submodel through p in $\mathcal{P}$.

- This equation involves multiple derivatives and many paths, directly solving for χ_p challenging.

HOIFs

- Thus, we must solve for an *mth* order U-statistic with mean zero that satisfies

$$\frac{d^j}{dt^j_{|t=0}}\chi\left(p_t\right) = \frac{d^j}{dt^j_{|t=0}}P_t^m\chi_p \qquad j = 1, ..., m$$

  for every submodel through p in $\mathcal{P}$.
- This equation involves multiple derivatives and many paths, directly solving for χ_p challenging.
- For actual computation of influence functions we have shown that it is usually easier to derive higher order influence functions as influence functions of lower order ones.

HOIFs

- Thus, we must solve for an *mth* order U-statistic with mean zero that satisfies

$$\frac{d^j}{dt^j_{|t=0}}\chi\left(p_t\right) = \frac{d^j}{dt^j_{|t=0}}P_t^m\chi_p \qquad j = 1, ..., m$$

  for every submodel through p in $\mathcal{P}$.
- This equation involves multiple derivatives and many paths, directly solving for χ_p challenging.
- For actual computation of influence functions we have shown that it is usually easier to derive higher order influence functions as influence functions of lower order ones.
- In fact, the Hoeffding Decomposition Theorem states that one can decompose any *mth* order U-statistic as the sum of m degenerate U-statistics of orders $1, 2, ..., m$.

HOIFs

- Thus by HDT

$$\mathbb{U}_n \chi_p = \mathbb{U}_n \chi_p^{(1)} + \mathbb{U}_n \frac{1}{2} \chi_p^{(2)} + \ldots + \mathbb{U}_n \frac{1}{m!} \chi_p^{(m)}$$

where $\chi_p^{(j)}$ is a degenerate (symmetric) kernel of j arguments defined uniquely as a projection of χ_p

HOIFs

- Thus by HDT

$$\mathbb{U}_n \chi_p = \mathbb{U}_n \chi_p^{(1)} + \mathbb{U}_n \frac{1}{2} \chi_p^{(2)} + \ldots + \mathbb{U}_n \frac{1}{m!} \chi_p^{(m)}$$

  where $\chi_p^{(j)}$ is a degenerate (symmetric) kernel of j arguments defined uniquely as a projection of χ_p

- Suitable functions $\chi_p^{(j)}$ can be found by the algorithm :

HOIFs

- Thus by HDT

$$\mathbb{U}_n\chi_p = \mathbb{U}_n\chi_p^{(1)} + \mathbb{U}_n\frac{1}{2}\chi_p^{(2)} + ... + \mathbb{U}_n\frac{1}{m!}\chi_p^{(m)}$$

  where $\chi_p^{(j)}$ is a degenerate (symmetric) kernel of j arguments defined uniquely as a projection of χ_p

- Suitable functions $\chi_p^{(j)}$ can be found by the algorithm :

  1. Let $x_1 \rightarrow \chi_p^{(1)}(x_1)$ be a first order influence function of the functional $p \rightarrow \chi(p)$

HOIFs

- Thus by HDT

$$\mathbb{U}_n\chi_p = \mathbb{U}_n\chi_p^{(1)} + \mathbb{U}_n\frac{1}{2}\chi_p^{(2)} + ... + \mathbb{U}_n\frac{1}{m!}\chi_p^{(m)}$$

where $\chi_p^{(j)}$ is a degenerate (symmetric) kernel of j arguments defined uniquely as a projection of χ_p

- Suitable functions $\chi_p^{(j)}$ can be found by the algorithm :

1. Let $x_1 \rightarrow \chi_p^{(1)}(x_1)$ be a first order influence function of the functional $p \rightarrow \chi(p)$
2. Let $x_j \rightarrow \widetilde{\chi}_p^{(j)}(x_1, x_2, ..., x_j)$ be a first order influence function of the functional $p \rightarrow \widetilde{\chi}_p^{(j-1)}(x_1, x_2, ..., x_{j-1})$ for each $x_1, x_2, ..., x_{j-1}$ and $j = 1, ..., m$.

HOIFs

- Thus by HDT

$$\mathbb{U}_n \chi_p = \mathbb{U}_n \chi_p^{(1)} + \mathbb{U}_n \frac{1}{2} \chi_p^{(2)} + ... + \mathbb{U}_n \frac{1}{m!} \chi_p^{(m)}$$

where $\chi_p^{(j)}$ is a degenerate (symmetric) kernel of j arguments defined uniquely as a projection of χ_p

- Suitable functions $\chi_p^{(j)}$ can be found by the algorithm :

1. Let $x_1 \rightarrow \chi_p^{(1)}(x_1)$ be a first order influence function of the functional $p \rightarrow \chi(p)$
2. Let $x_j \rightarrow \widetilde{\chi}_p^{(j)}(x_1, x_2, ..., x_j)$ be a first order influence function of the functional $p \rightarrow \widetilde{\chi}_p^{(j-1)}(x_1, x_2, ..., x_{j-1})$ for each $x_1, x_2, ..., x_{j-1}$ and $j = 1, ..., m$.
3. Let $\chi_p^{(j)}$ be the degenerate part of $\widetilde{\chi}_p^{(j)}$ relative to P, *i.e.* $\int \chi_p^{(j)}(X_1, ..., x_s, ..., X_j)\, dP(x_s) = 0$ for all $s = 1, ...j$.

HOIFs

- Thus, HIOFs are constructed as first order influence functions of an influence function.

HOIFs

- Thus, HIOFs are constructed as first order influence functions of an influence function.
- Although the order m is fixed at a suitable value, we suppress it in the notation χ_p

HOIFs

- Thus, HIOFs are constructed as first order influence functions of an influence function.
- Although the order m is fixed at a suitable value, we suppress it in the notation χ_p
- In abuse of language, we refer to $\chi_p^{(j)}$ as the *jth* order IF.

HOIFs

- Thus, HIOFs are constructed as first order influence functions of an influence function.
- Although the order m is fixed at a suitable value, we suppress it in the notation χ_p
- In abuse of language, we refer to $\chi_p^{(j)}$ as the *jth* order IF.
- The starting IF $\chi_p^{(1)}$ in step 1 of the algorithm may be any first order IF, it does not need to be an element of the tangent space (i.e. efficient) and does not need to have mean zero.

HOIFs

- Thus, HIOFs are constructed as first order influence functions of an influence function.
- Although the order m is fixed at a suitable value, we suppress it in the notation χ_p
- In abuse of language, we refer to $\chi_p^{(j)}$ as the *jth* order IF.
- The starting IF $\chi_p^{(1)}$ in step 1 of the algorithm may be any first order IF, it does not need to be an element of the tangent space (i.e. efficient) and does not need to have mean zero.
- The same remark applies in step 2 of the algorithm, it is only in step 3 that we make the IFs degenerate.

HOIFs of quadratic functional

- Recall the uncentered (1st order) IF of $p \longrightarrow \int_{\mathbb{R}^p} p(x)^2 dx$ is $\widetilde{\chi}_p^{(1)}(x) = 2p(x)$, where recall $p(X)$ is the joint density of X. Applying step 2 for $j = 2$, we seek $\widetilde{\chi}_p^{(2)}(x_1, x_2)$ that solves

$$\frac{d}{dt}\Big|_{t=0} \widetilde{\chi}_{p_t}^{(1)}(x_1) = P\left\{\widetilde{\chi}_p^{(2)}(x_1, X_2) g(X_2)\right\}$$

for all x_1 and smooth paths $t \longrightarrow p_t$; however $\frac{d}{dt}\big|_{t=0} \widetilde{\chi}_{p_t}^{(1)}(x_1) = 2\frac{d}{dt}\big|_{t=0} p_t(x_1)$;

HOIFs of quadratic functional

- Recall the uncentered (1st order) IF of $p \longrightarrow \int_{\mathbb{R}^p} p(x)^2 dx$ is $\widetilde{\chi}_p^{(1)}(x) = 2p(x)$, where recall $p(X)$ is the joint density of X. Applying step 2 for $j = 2$, we seek $\widetilde{\chi}_p^{(2)}(x_1, x_2)$ that solves

$$\frac{d}{dt}\Big|_{t=0} \widetilde{\chi}_{p_t}^{(1)}(x_1) = P\left\{\widetilde{\chi}_p^{(2)}(x_1, X_2) g(X_2)\right\}$$

  for all x_1 and smooth paths $t \longrightarrow p_t$; however $\frac{d}{dt}\big|_{t=0} \widetilde{\chi}_{p_t}^{(1)}(x_1) = 2\frac{d}{dt}\big|_{t=0} p_t(x_1)$;
- however, $p(x)$ as a continuous density is not pathwise differentiable so that such representation does not exist!

Approximate functionals

- HOIF will generally fail to exist if the first order IF depends on a parameter that is not pathwise differentiable, such as a conditional expectation or a density at a point.

Approximate functionals

- HOIF will generally fail to exist if the first order IF depends on a parameter that is not pathwise differentiable, such as a conditional expectation or a density at a point.
- To make progress, we propose to replace $\chi(p)$ with an approximate functional $\chi(\widetilde{p})$ for a given mapping $p \longrightarrow \widetilde{p}$ such that $\chi(p)$ and $\chi(\widetilde{p})$ are close, and $\chi(\widetilde{p})$ is higher order pathwise differentiable.

Approximate functionals

- HOIF will generally fail to exist if the first order IF depends on a parameter that is not pathwise differentiable, such as a conditional expectation or a density at a point.
- To make progress, we propose to replace $\chi(p)$ with an approximate functional $\chi(\widetilde{p})$ for a given mapping $p \longrightarrow \widetilde{p}$ such that $\chi(p)$ and $\chi(\widetilde{p})$ are close, and $\chi(\widetilde{p})$ is higher order pathwise differentiable.
- This leads to representation bias. To ensure that this bias is negligible will require the projection of p on $\widetilde{p}$ involves models of high dimension # of parms $\gg n$.

Approximate functionals

- HOIF will generally fail to exist if the first order IF depends on a parameter that is not pathwise differentiable, such as a conditional expectation or a density at a point.
- To make progress, we propose to replace $\chi(p)$ with an approximate functional $\chi(\widetilde{p})$ for a given mapping $p \longrightarrow \widetilde{p}$ such that $\chi(p)$ and $\chi(\widetilde{p})$ are close, and $\chi(\widetilde{p})$ is higher order pathwise differentiable.
- This leads to representation bias. To ensure that this bias is negligible will require the projection of p on $\widetilde{p}$ involves models of high dimension # of parms $\gg n$.
- This will typically result in variance of HOIF $\gg n^{-1}$ and nonparametric rates of convergence apply.

Projection Kernel

- An orthogonal projection $K_p : L_2(p) \longmapsto L_2(p)$ on a k-dimensional linear subspace of $L_2(p)$ is given by a kernel operator, with kernel denoted by the same symbol as the operator: $K_p t(x) = \int K_p(x, y)\, t(y)\, dP(y)$. The projection property $K_p^2 = K_p$ holds.

Projection Kernel

- An orthogonal projection $K_p : L_2(p) \longmapsto L_2(p)$ on a k-dimensional linear subspace of $L_2(p)$ is given by a kernel operator, with kernel denoted by the same symbol as the operator: $K_p t(x) = \int K_p(x, y)\, t(y)\, dP(y)$. The projection property $K_p^2 = K_p$ holds.
- We also require that

$$K_p(u, u) \precsim k \quad \text{and} \quad \|K_p t\|_\infty \precsim \|t\|_\infty$$

Projection Kernel

- An orthogonal projection $K_p : L_2(p) \longmapsto L_2(p)$ on a k-dimensional linear subspace of $L_2(p)$ is given by a kernel operator, with kernel denoted by the same symbol as the operator: $K_p t(x) = \int K_p(x, y)\, t(y)\, dP(y)$. The projection property $K_p^2 = K_p$ holds.
- We also require that

$$K_p(u, u) \precsim k \quad \text{and} \quad \|K_p t\|_\infty \precsim \|t\|_\infty$$

- Moreover, it is desirable that the projection operator is suitable for approximation in Hölder space: for every k,and $\|\cdot\|_\alpha$ a norm for $H^\alpha[0,1]^d$,

$$\sup_{\|t\|_\alpha \leq 1} \|t - K_p t\|_\infty \precsim \left(\frac{1}{k}\right)^{\alpha/d}$$

Approximate quadratic functional

- The approximate functional $\chi\left(\widetilde{p}\right)=\int\widetilde{p}\left(x\right)^{2}dx$

$$\widetilde{p}\left(x\right)=K_{\mu}p\left(x\right),\ \mu \text{ p-fold Lebesgue measure}$$

Approximate quadratic functional

- The approximate functional $\chi\left(\widetilde{p}\right) = \int \widetilde{p}\left(x\right)^2 dx$

$$\widetilde{p}\left(x\right) = K_{\mu} p\left(x\right), \ \mu \text{ p-fold Lebesgue measure}$$

- Note that whereas $p(x)$ is not pathwise differentiable $\widetilde{p}\left(x\right)$ is!

Approximate quadratic functional

- The approximate functional $\chi(\widetilde{p}) = \int \widetilde{p}(x)^2 \, dx$

$$\widetilde{p}(x) = K_\mu p(x), \ \mu \text{ p-fold Lebesgue measure}$$

- Note that whereas $p(x)$ is not pathwise differentiable $\widetilde{p}(x)$ is!
-

$$\begin{aligned} \frac{d}{dt}\Big|_{t=0} \chi(\widetilde{p}_t) &= P\{2(\widetilde{p}(X))g\} \\ \frac{d}{dt}\Big|_{t=0} \widetilde{\chi}^{(1)}_{p_t}(x_1) &= 2\frac{d}{dt}\Big|_{t=0} \widetilde{p}_t(x_1) \\ &= P\{2K_\mu(x_1, X_2)g(X_2)\} \end{aligned}$$

therefore $\widetilde{\chi}^{(2)}_{p_t}(x_1, x_2) = K_\mu(x_1, x_2)$, $\widetilde{\chi}^{(j)}_{p_t}(x_1, x_2) = 0$, $j > 2$

Approximate quadratic functional

- Step 3 of the algorithm yields a 2nd order IF
  $\mathbb{U}_n \chi_p = \mathbb{U}_n \chi_p^{(1)} + \mathbb{U}_n \frac{1}{2} \chi_p^{(2)}$

$$\begin{aligned} \chi_p^{(1)}(x_1) &= 2\left(\widetilde{p}\left(x_1\right) - \chi\left(\widetilde{p}\right)\right); \\ \chi_p^{(2)}(x_1, x_2) &= K_\mu\left(x_1, x_2\right) - K_\mu p\left(x_1\right) - K_\mu p\left(x_2\right) + \chi\left(\widetilde{p}\right) \end{aligned}$$

Approximate quadratic functional

- Step 3 of the algorithm yields a 2nd order IF
  $\mathbb{U}_n \chi_p = \mathbb{U}_n \chi_p^{(1)} + \mathbb{U}_n \frac{1}{2} \chi_p^{(2)}$

$$\begin{aligned} \chi_p^{(1)}(x_1) &= 2\left(\widetilde{p}(x_1) - \chi(\widetilde{p})\right); \\ \chi_p^{(2)}(x_1, x_2) &= K_\mu(x_1, x_2) - K_\mu p(x_1) - K_\mu p(x_2) + \chi(\widetilde{p}) \end{aligned}$$

- We further have

$$\hat{\chi}_n = \chi(\hat{p}_n) + \mathbb{U}_n \chi_{\hat{p}_n}$$

and

$$\begin{aligned} P\left(\hat{\chi}_n - \chi(\widetilde{p})\right) &= 0 \\ var(\hat{\chi}_n) &\asymp \max\left(var\left(\mathbb{U}_n \chi_{\widehat{p}}^{(1)}\right), var\left(\mathbb{U}_n \frac{1}{2} \chi_p^{(2)}\right)\right) \\ &= \max\left(\frac{1}{n}, \frac{k}{n^2}\right) \end{aligned}$$

Approximate quadratic functional

- Step 3 of the algorithm yields a 2nd order IF
  $\mathbb{U}_n \chi_p = \mathbb{U}_n \chi_p^{(1)} + \mathbb{U}_n \frac{1}{2} \chi_p^{(2)}$

$$\begin{aligned} \chi_p^{(1)}(x_1) &= 2\left(\widetilde{p}(x_1) - \chi(\widetilde{p})\right); \\ \chi_p^{(2)}(x_1, x_2) &= K_\mu(x_1, x_2) - K_\mu p(x_1) - K_\mu p(x_2) + \chi(\widetilde{p}) \end{aligned}$$

- We further have

$$\hat{\chi}_n = \chi(\hat{p}_n) + \mathbb{U}_n \chi_{\hat{p}_n}$$

and

$$\begin{aligned} P\left(\hat{\chi}_n - \chi(\widetilde{p})\right) &= 0 \\ var(\hat{\chi}_n) &\asymp \max\left(var\left(\mathbb{U}_n \chi_{\widehat{p}}^{(1)}\right), var\left(\mathbb{U}_n \frac{1}{2} \chi_p^{(2)}\right)\right) \\ &= \max\left(\frac{1}{n}, \frac{k}{n^2}\right) \end{aligned}$$

- The representation bias is $\chi(\widetilde{p}) - \chi(p) = \left(\frac{1}{k}\right)^{2\alpha/d}$

Approximate quadratic functional

- Therefore if $\frac{\alpha}{d} \geq \frac{1}{4}$, approximation bias can be made $\precsim \frac{1}{n}$ for $k < n$ so that $var(\hat{\chi}_n) \asymp \max\left(\frac{1}{n}, \frac{k}{n^2}\right) = \frac{1}{n}$, and the functional is estimable at rate $n^{1/2}$.

Approximate quadratic functional

- Therefore if $\frac{\alpha}{d} \geq \frac{1}{4}$, approximation bias can be made $\precsim \frac{1}{n}$ for $k < n$ so that $var(\hat{\chi}_n) \asymp \max\left(\frac{1}{n}, \frac{k}{n^2}\right) = \frac{1}{n}$, and the functional is estimable at rate $n^{1/2}$.
- However whenever $\frac{\alpha}{d} < \frac{1}{4}$ optimal MSE is attained by $k_{opt} = n^{\frac{2d}{4\alpha+d}} > n$ achieving the minimax rate $n^{-\frac{8\alpha}{4\alpha+d}}$

Approximate quadratic functional

- Therefore if $\frac{\alpha}{d} \geq \frac{1}{4}$, approximation bias can be made $\precsim \frac{1}{n}$ for $k < n$ so that $var(\hat{\chi}_n) \asymp \max\left(\frac{1}{n}, \frac{k}{n^2}\right) = \frac{1}{n}$, and the functional is estimable at rate $n^{1/2}$.
- However whenever $\frac{\alpha}{d} < \frac{1}{4}$ optimal MSE is attained by $k_{opt} = n^{\frac{2d}{4\alpha+d}} > n$ achieving the minimax rate $n^{-\frac{8\alpha}{4\alpha+d}}$
- Note that because there is no estimation bias, the optimal rate is obtained by trading off variance and representation bias.

Approximate functional

- Interestingly, in this example the approximate functional is locally equivalent to the true functional in the sense that their first order IFs match.

Approximate functional

- Interestingly, in this example the approximate functional is locally equivalent to the true functional in the sense that their first order IFs match.
- Although this is a desirable property because it leads to parsimounous HOIFs, this will not generally be the case for a general functional, but can be ensured by construction.

Approximate functional

- Interestingly, in this example the approximate functional is locally equivalent to the true functional in the sense that their first order IFs match.
- Although this is a desirable property because it leads to parsimounous HOIFs, this will not generally be the case for a general functional, but can be ensured by construction.
- To do so, we propose to define the approximate functional to satisfy

$$\frac{d}{dt_{|t=0}}\left(\chi\left(\widetilde{p}_t\right)+P\chi^{(1)}_{\widetilde{p}_t}\right)=0$$

Approximate functional

- Interestingly, in this example the approximate functional is locally equivalent to the true functional in the sense that their first order IFs match.
- Although this is a desirable property because it leads to parsimounous HOIFs, this will not generally be the case for a general functional, but can be ensured by construction.
- To do so, we propose to define the approximate functional to satisfy

$$\frac{d}{dt_{|t=0}}\left(\chi\left(\widetilde{p}_t\right)+P\chi^{(1)}_{\widetilde{p}_t}\right)=0$$

- In fact then $\widetilde{\chi}^{(1)}_p=\chi^{(1)}_{\widetilde{p}}$, therefore step 1 of the algorithm returns the same 1st order IF.

Approximate functional for missing data functional

- To define approximate functional, for user-specified $\underline{a}$ we let $\widetilde{a}$ be the function such that $(\widetilde{a} - \widehat{a}) / \underline{a}$ is the orthogonal projection of $(a - \widehat{a}) / \underline{a}$ onto a "large" linear subspace of $L_2(g)$, with projection kernel K_g, i.e.

$$\frac{\widetilde{a}}{\underline{a}} = \frac{\widehat{a}}{\underline{a}} + K_g \left(\frac{a}{\underline{a}} - \frac{\widehat{a}}{\underline{a}} \right)$$

  and define $\widetilde{b}$ similarly.

Approximate functional for missing data functional

- To define approximate functional, for user-specified $\underline{a}$ we let $\widetilde{a}$ be the function such that $(\widetilde{a}-\widehat{a})/\underline{a}$ is the orthogonal projection of $(a-\widehat{a})/\underline{a}$ onto a "large" linear subspace of $L_2(g)$, with projection kernel K_g, i.e.

$$\frac{\widetilde{a}}{\underline{a}} = \frac{\widehat{a}}{\underline{a}} + K_g\left(\frac{a}{\underline{a}} - \frac{\widehat{a}}{\underline{a}}\right)$$

  and define $\widetilde{b}$ similarly.
- The corresponding approximate functional is $\chi(\widetilde{p}) = \int \widetilde{a}\widetilde{b}gd\nu$ satisfies $\widetilde{\chi}_p^{(1)} = \chi_{\widetilde{p}}^{(1)}$ and its representation bias is

$$\left|\int \widetilde{a}\widetilde{b}gd\nu - \int abgd\nu\right|^2 \leq \int \left|\frac{(a-\widetilde{a})}{\underline{a}}\right|^2 \underline{ab}gd\nu \int \left|\frac{\left(b-\widetilde{b}\right)}{\underline{b}}\right|^2 \underline{ab}gd\nu$$

$$\precsim \left(\frac{1}{k}\right)^{(\alpha+\beta)/d}$$

which can be made small by chosing the range of K_g large enough.

Approximate functional for missing data functional

- Step 1 of the algorithm applied to $\chi\left(\widetilde{p}\right)$ gives $\widetilde{\chi}_{\widetilde{p}}^{(1)} = a_1 \widetilde{a}\left(z_1\right)\left(y_1 - \widetilde{b}\left(z_1\right)\right) + \widetilde{b}\left(z_1\right)$

Approximate functional for missing data functional

- Step 1 of the algorithm applied to $\chi(\widetilde{p})$ gives $\widetilde{\chi}_{\widetilde{p}}^{(1)} = a_1 \widetilde{a}(z_1)\left(y_1 - \widetilde{b}(z_1)\right) + \widetilde{b}(z_1)$
- The corresponding bias of $\hat{\chi}_n = \chi(\hat{p}_n) + \mathbb{U}_n \chi^1_{\hat{p}_n}$ is

$$-\int (\widehat{a} - a)\left(\widehat{b} - b\right) g d\nu = O_p\left(n^{-\frac{\alpha}{2\alpha+d} - \frac{\beta}{2\beta+d}}\right)$$

while the corresponding variance is order $1/n$, this requires $\frac{\alpha}{2\alpha+d} + \frac{\beta}{2\beta+d} \geq 1/2$ to recover regular estimator when possible.

Approximate functional for missing data functional

- Step 1 of the algorithm applied to $\chi(\widetilde{p})$ gives $\widetilde{\chi}_{\widetilde{p}}^{(1)} = a_1 \widetilde{a}(z_1)\left(y_1 - \widetilde{b}(z_1)\right) + \widetilde{b}(z_1)$
- The corresponding bias of $\hat{\chi}_n = \chi(\hat{p}_n) + \mathbb{U}_n \chi^1_{\hat{p}_n}$ is

$$-\int (\widehat{a} - a)\left(\widehat{b} - b\right) g d\nu = O_p\left(n^{-\frac{\alpha}{2\alpha+d} - \frac{\beta}{2\beta+d}}\right)$$

  while the corresponding variance is order $1/n$, this requires $\frac{\alpha}{2\alpha+d} + \frac{\beta}{2\beta+d} \geq 1/2$ to recover regular estimator when possible.
- One may be able to relax this somewhat by undersmoothing or exploiting the fact that the bias has integral representation. We do not wish to do so here as we assume all nuisance parameters are rate optimal.

Approximate functional for missing data functional

- Whereas $\widetilde{\chi}_{p}^{(1)}(x_1)$ is not pathwise diff., $\widetilde{\chi}_{\widetilde{p}}^{(1)}(x_1)$ is.

Approximate functional for missing data functional

- Whereas $\widetilde{\chi}_{p}^{(1)}(x_1)$ is not pathwise diff., $\widetilde{\chi}_{\widetilde{p}}^{(1)}(x_1)$ is.
- Step 2 of the algorithm applied to $\widetilde{\chi}_{\widetilde{p}}^{(1)}(x_1)$ gives

$$\widetilde{\chi}_{\widetilde{p}}^{(2)}(x_1, X_2) = -2a_1(y_1 - \widetilde{b}(z_1))K_g(z_1, Z_2)(A_2\widetilde{a}(Z_2) - 1)$$

Approximate functional for missing data functional

- Whereas $\widetilde{\chi}_{p}^{(1)}(x_1)$ is not pathwise diff., $\widetilde{\chi}_{\widetilde{p}}^{(1)}(x_1)$ is.
- Step 2 of the algorithm applied to $\widetilde{\chi}_{\widetilde{p}}^{(1)}(x_1)$ gives

$$\widetilde{\chi}_{\widetilde{p}}^{(2)}(x_1, X_2) = -2a_1(y_1 - \widetilde{b}(z_1))K_g(z_1, Z_2)(A_2\widetilde{a}(Z_2) - 1)$$

- Step 3 symmetrizes this IF and makes it degenerate U-statistic.

Approximate functional for missing data functional

- Whereas $\widetilde{\chi}_{p}^{(1)}(x_1)$ is not pathwise diff., $\widetilde{\chi}_{\widetilde{p}}^{(1)}(x_1)$ is.
- Step 2 of the algorithm applied to $\widetilde{\chi}_{\widetilde{p}}^{(1)}(x_1)$ gives

$$\widetilde{\chi}_{\widetilde{p}}^{(2)}(x_1, X_2) = -2a_1(y_1 - \widetilde{b}(z_1))K_g(z_1, Z_2)(A_2\widetilde{a}(Z_2) - 1)$$

- Step 3 symmetrizes this IF and makes it degenerate U-statistic.
- Note that because K_g will depend on g the 2nd order IF is not unbiased and therefore higher order IFs may be required to further reduce estimation bias depending on smoothness of g.

Minimax estimation with 2nd order IF

- Let $\hat{\chi}_n = \chi(\hat{p}_n) + \mathbb{U}_n \chi^1_{\hat{p}_n} + \mathbb{U}_n \chi^2_{\hat{p}_n}$ and let γ denote the Hölder coeff of g.

Minimax estimation with 2nd order IF

- Let $\hat{\chi}_n = \chi(\hat{p}_n) + \mathbb{U}_n \chi^1_{\hat{p}_n} + \mathbb{U}_n \chi^2_{\hat{p}_n}$ and let γ denote the Hölder coeff of g.

Theorem

Suppose $\frac{\gamma}{d+2\gamma} \geq \frac{2(\beta+\alpha)}{d+2(\beta+\alpha)} - \left(\frac{\beta}{d+2\beta} + \frac{\alpha}{d+2\alpha}\right)$; *set* $k = n^{\frac{2}{d+2(\beta+\alpha)}}$, *then we have the following result:*
If $\frac{(\beta+\alpha)}{2d} \geq \frac{1}{4}$, *then*

$$\sup_{\theta \in \Theta(\beta,\alpha,\gamma)} (\hat{\chi}_n - \chi(p)) = O_p\left(\frac{1}{\sqrt{n}}\right)$$

If $\frac{(\beta+\alpha)}{2d} \leq \frac{1}{4}$, *then*

$$\sup_{\theta \in \Theta(\beta,\alpha,\gamma)} (\hat{\chi}_n - \chi(p)) = O_p\left(n^{-\frac{2(\beta+\alpha)}{d+2(\beta+\alpha)}}\right)$$

Minimax estimation

- This result is new!!! All existing methods break down once $(\beta + \alpha)/2d \leq \frac{1}{4}$, i.e. once the effective smoothness $(\beta + \alpha)/2d$ is too small.

Minimax estimation

- This result is new!!! All existing methods break down once $(\beta + \alpha)/2d \leq \frac{1}{4}$, i.e. once the effective smoothness $(\beta + \alpha)/2d$ is too small.
- We have likewise constructed minimax estimators when $\frac{\gamma}{d+2\gamma} < \frac{2(\beta+\alpha)}{d+2(\beta+\alpha)} - \left(\frac{\beta}{d+2\beta} + \frac{\alpha}{d+2\alpha}\right)$

Minimax estimation

- This result is new!!! All existing methods break down once $(\beta+\alpha)/2d \leq \frac{1}{4}$, i.e. once the effective smoothness $(\beta+\alpha)/2d$ is too small.
- We have likewise constructed minimax estimators when $\frac{\gamma}{d+2\gamma} < \frac{2(\beta+\alpha)}{d+2(\beta+\alpha)} - \left(\frac{\beta}{d+2\beta} + \frac{\alpha}{d+2\alpha}\right)$
- These higher order IFs are higher order U statistics for further bias reduction without increasing the variance from the second order rate $\frac{k}{n^2}$, this requires tremendous care. (see Robins et al, 2016, Annals of Statistics)

Honest confidence Intervals

- Our approach can be used to obtain honest CIs for $\chi(p)$ given by

$$\mathcal{C}_n = \hat{\chi}_n \pm m_{1-\alpha} \widehat{\sigma}(\hat{\chi}_n)$$

so that

$$\lim_{n \to \infty} \Pr\{\chi(p) \notin \mathcal{C}_n\} \leq \alpha$$

uniformly in $p \in \mathcal{P}$ with appropriate constant $m_{1-\alpha}$

Honest confidence Intervals

- Our approach can be used to obtain honest CIs for $\chi(p)$ given by

$$\mathcal{C}_n = \hat{\chi}_n \pm m_{1-\alpha} \widehat{\sigma}(\hat{\chi}_n)$$

so that

$$\lim_{n \to \infty} \Pr\{\chi(p) \notin \mathcal{C}_n\} \leq \alpha$$

uniformly in $p \in \mathcal{P}$ with appropriate constant $m_{1-\alpha}$

- This is because our HOIFs are asymptotically normally distributed (Robins et al, 2016).

Simulations of honest confidence Intervals

Smoothness	Coverage Rate	
β/d	$\mathscr{C}_1$	$\mathscr{C}_{\text{new}}$
0.6	89	91
0.5	85	90
0.3	51	89
0.25	16	86
0.125	0	84

- Nominal coverage 90%. Sample size 1000. 200 iterations, effective smoothnes of $g = .2$.

Final remarks

- Our results appy to a general class of functionals which includes the semilinear model, several IV models, MAR monotone missing data functionals, nonignorable missing data functionals among others.

Final remarks

- Our results appy to a general class of functionals which includes the semilinear model, several IV models, MAR monotone missing data functionals, nonignorable missing data functionals among others.
- One will rarely know the smoothness of nuisance parameters, ideally one would like to adapt to such smoothness.

Final remarks

- Our results appy to a general class of functionals which includes the semilinear model, several IV models, MAR monotone missing data functionals, nonignorable missing data functionals among others.
- One will rarely know the smoothness of nuisance parameters, ideally one would like to adapt to such smoothness.
- We have recently developed such adaptive estimators for a large class of functionals by extending and applying a technique proposed by Lepski.(Mukherjee,Tchetgen Tchetgen, Robins, 2016)