Semiparametric estimation in very high-dimensional models

James Robins Lingling Li Eric Tchetgen Aad van der Vaart

Harvard University Universiteit Leiden

Four facets of statistics and a Surprise, Zürich, November 2015

Semiparametrics

Semiparametric Inference

A semiparametric model is a model of infinite dimension

Interest is in estimating a finite-dimensional parameter defined through the structure of the model, e.g.

- a relative risk
- a coefficient of a particular regression variable
- a mean response

Classical semiparametrics (1980/90s): the combination of parameter and model is such that the “bias” is small relative to “variance” and estimation is possible at the “parametric rate” $\sqrt{n}$.

Modern semiparametrics: what if the model is too “large” for a parametric rate?

Classical semiparametrics

$X_1, \ldots, X_n$ i.i.d. with density $p \in \mathcal{P}$

We want to estimate $\chi(p)$, for $\chi\colon \mathcal{P} \to \mathbb{R}$.

META THEOREM

If $\mathcal{P}$ and χ are nice, then there exist $T_n = T_n(X_1, \ldots, X_n)$, with

$$\sqrt{n}\big(T_n - \chi(p)\big) \rightsquigarrow N(0, \sigma_p^2).$$

Classical semiparametrics (1980/90s) was concerned with finding T_n with minimal σ_p^2

General methods such as (semiparametric, penalized, sieved) maximum likelihood or Bayes work well for many semiparametric models.

Example: symmetric location (Stein (1956), Stone (1975), Bickel (1981), ...)

Error ε with symmetric density η, Fisher information $I_\eta < \infty$

Observe $X = \theta + \varepsilon$

THEOREM

There exists $T_n = T_n(X_1, \ldots, X_n)$ with, for all (θ, η),

$$\sqrt{n}(T_n - \theta) \rightsquigarrow N(0, I_\eta^{-1})$$

Example: Cox model (Cox (1975), Tsiatis, Gill, Wellner,...)

Covariate Z, $\sim f$
Survival time T with conditional hazard $\lambda(t)e^{\theta^T z}$
Observe (Z, T)

THEOREM
There exists $T_n = T_n(X_1, \ldots, X_n)$ with, for all (θ, λ, f),

$$\sqrt{n}(T_n - \theta) \rightsquigarrow N\big(0, \sigma^2_{\theta,\lambda,f}\big)$$

Maximum likelihood estimator and certain Bayes estimators attain minimal $\sigma^2_{\theta,\lambda,f}$.

Example: semiparametric regression

Covariates (W, Z), density f
Error ε, such that $\varepsilon|\,(W, Z)$ has density $g(\cdot|\,W, Z)$ with $\mathrm{E}(\varepsilon|\,W, Z) = 0$
Outcome $Y = \theta W + \eta(Z) + \varepsilon$
Observe $X = (W, Z, Y)$

THEOREM
There exists $T_n = T_n(X_1, \ldots, X_n)$ with, for all (θ, f, g, η) such that g and η are sufficiently smooth,

$$\sqrt{n}(T_n - \theta) \rightsquigarrow N\big(0, \sigma^2_{\theta,f,g,\eta}\big)$$

Example: missing data (Robins and Rotnitzky, ...)

Covariate Z, $\quad \sim f$
Response Y, with $Y|\,Z \sim$ binomial $\big(1, b(Z)\big)$
Missingness indicator A, with $A|\,Z \sim$ binomial $\big(1, 1/a(Z)\big)$
Missing at random: $Y \perp\!\!\!\perp A|\,Z$
Observe $X = (YA, A, Z) \in \{0,1\} \times \{0,1\} \times [0,1]^d$

We wish to estimate mean response $\chi(a,b,f) = \int b\, f\, d\nu = \mathrm{E}Y$.

Example: missing data (Robins and Rotnitzky, ...)

Covariate Z, $\quad\sim f$
Response Y, with $Y|\,Z \sim$ binomial $\big(1, b(Z)\big)$
Missingness indicator A, with $A|\,Z \sim$ binomial $\big(1, 1/a(Z)\big)$
Missing at random: $Y \perp\!\!\!\perp A|\,Z$
Observe $X = (YA, A, Z) \in \{0,1\} \times \{0,1\} \times [0,1]^d$

We wish to estimate mean response $\chi(a,b,f) = \int b\, f\, d\nu = \mathrm{E}Y$.

Y is observed only if $A = 1$
Z is included to make the assumption $Y \perp\!\!\!\perp A|\,Z$ realistic

Example: missing data (Robins and Rotnitzky, ...)

Covariate Z, $\quad \sim f$

Response Y, with $Y|\,Z \sim$ binomial $\big(1, b(Z)\big)$

Missingness indicator A, with $A|\,Z \sim$ binomial $\big(1, 1/a(Z)\big)$

Missing at random: $Y \perp\!\!\!\perp A|\,Z$

Observe $X = (YA, A, Z) \in \{0,1\} \times \{0,1\} \times [0,1]^d$

We wish to estimate mean response $\chi(a,b,f) = \int b\, f\, d\nu = \mathrm{E}Y$.

Y is observed only if $A = 1$

Z is included to make the assumption $Y \perp\!\!\!\perp A|\,Z$ realistic

THEOREM

There exists $T_n = T_n(X_1, \ldots, X_n)$ with, for all (a,b,f) such that a and b are sufficiently smooth,

$$\sqrt{n}\big(T_n - \chi(a,b,f)\big) \rightsquigarrow N\big(0, \sigma^2_{a,b,f}\big)$$

Classical semiparametrics — minimal variance

$X_1, \ldots, X_n$ i.i.d. with density $p \in \mathcal{P}$

We want to estimate $\chi(p)$, for $\chi\colon \mathcal{P} \to \mathbb{R}$.

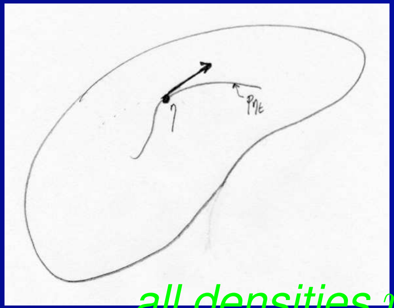
META THEOREM

If $\mathcal{P}$ and χ are nice, then there exist $T_n = T_n(X_1, \ldots, X_n)$, with

$$\sqrt{n}\big(T_n - \chi(p)\big) \rightsquigarrow N(0, \sigma_p^2).$$

What is the minimal variance σ_p^2?

First order tangent space and influence functions (Koshevnik and Levit (1976), Pfanzagl (1983), vdV (1988).)

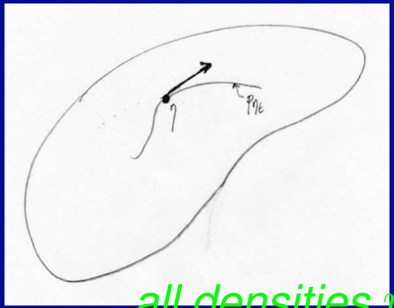


all densities p

Tangent set (at p): all score functions $g = \frac{d}{dt}_{|t=0} \log p_t$ of one-dimensional submodels $t \mapsto p_t$ with $p_0 = p$

First order tangent space and influence functions (Koshevnik and Levit (1976), Pfanzagl (1983), vdV (1988).)

Tangent set (at p): all score functions $g = \frac{d}{dt}_{|t=0} \log p_t$ of one-dimensional submodels $t \mapsto p_t$ with $p_0 = p$

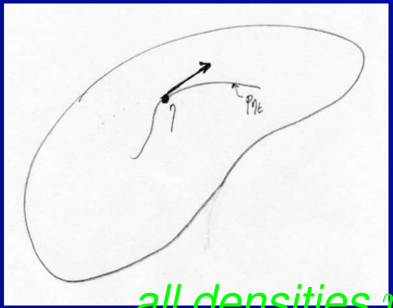


all densities p

Influence function of $p \mapsto \chi(p)$ is map $x \mapsto \chi_p(x)$ with for all $t \mapsto p_t$

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

First order tangent space and influence functions (Koshevnik and Levit (1976), Pfanzagl (1983), vdV (1988).)



all densities p

Tangent set (at p): all score functions $g = \frac{d}{dt}_{|t=0} \log p_t$ of one-dimensional submodels $t \mapsto p_t$ with $p_0 = p$

Influence function of $p \mapsto \chi(p)$ is map $x \mapsto \chi_p(x)$ with for all $t \mapsto p_t$

$$\frac{d}{dt}\chi(p_t)_{|t=0} = Pg\chi_p$$

THEOREM If $\sqrt{n}\big(T_n - \chi(p)\big) \rightsquigarrow L$, locally uniformly in p, then for some M

$$L = N\big(0, P(\Pi_p\chi_p)^2\big) * M,$$

for Π_p orthogonal projection onto closed linear span of tangent set.

Example: missing data

Observe $X = (YA, A, Z)$
Parameter $p \leftrightarrow (a, b, f)$

Likelihood $f(Z)(1/a)(Z)^A\big(1 - 1/a(Z)\big)^{1-A} b(Z)^{YA}\big(1 - b(Z)\big)^{(1-Y)A}$

$$\frac{Aa(Z) - 1}{a(Z)(a-1)(Z)}\alpha(Z) \qquad a\text{-score, } a_t = a + t\alpha$$

$$\frac{A(Y - b(Z))}{b(Z)(1-b)(Z)}\beta(Z) \qquad b\text{-score, } b_t = b + t\beta$$

$$\phi(Z) \qquad f\text{-score, } f_t = f(1 + t\phi), \quad \int \phi f = 0$$

Example: missing data

Observe $X = (YA, A, Z)$

Parameter $p \leftrightarrow (a, b, f)$

Likelihood $f(Z)(1/a)(Z)^A\big(1 - 1/a(Z)\big)^{1-A} b(Z)^{YA}\big(1 - b(Z)\big)^{(1-Y)A}$

$$\frac{Aa(Z) - 1}{a(Z)(a-1)(Z)}\alpha(Z) \qquad a\text{-score, } a_t = a + t\alpha$$

$$\frac{A(Y - b(Z))}{b(Z)(1-b)(Z)}\beta(Z) \qquad b\text{-score, } b_t = b + t\beta$$

$$\phi(Z) \qquad f\text{-score, } f_t = f(1 + t\phi)$$

Parameter of interest $\chi(p) = \int bf = \mathrm{E}Y$

Influence function $\chi_p(X) = Aa(Z)(Y - b(Z)) + b(Z) - \chi(p)$

Example: missing data

Observe $X = (YA, A, Z)$

Parameter $p \leftrightarrow (a, b, f)$

Likelihood $f(Z)(1/a)(Z)^A \big(1 - 1/a(Z)\big)^{1-A} b(Z)^{YA} \big(1 - b(Z)\big)^{(1-Y)A}$

$$\frac{Aa(Z) - 1}{a(Z)(a-1)(Z)}\alpha(Z) \qquad a\text{-score, } a_t = a + t\alpha$$

$$\frac{A(Y - b(Z))}{b(Z)(1-b)(Z)}\beta(Z) \qquad b\text{-score, } b_t = b + t\beta$$

$$\phi(Z) \qquad f\text{-score, } f_t = f(1 + t\phi)$$

Parameter of interest $\chi(p) = \int bf = \mathrm{E}Y$

Influence function $\chi_p(X) = Aa(Z)(Y - b(Z)) + b(Z) - \chi(p)$

$$\mathrm{E}_p \chi_p(X)[a\text{-score}] = 0$$

$$\mathrm{E}_p \chi_p(X)[b\text{-score}] = \frac{\partial}{\partial t}_{|t=0} \int b_t f \, d\nu = \int \beta f \, d\nu$$

$$\mathrm{E}_p \chi_p(X)[f\text{-score}] = \frac{\partial}{\partial t}_{|t=0} \int b f_t \, d\nu = \int b\phi f \, d\nu$$

Corrected plug-in estimators

Heuristics — plug in and bias correction

Estimate $\theta := \chi(p) \in \mathbb{R}$ from iid $X_1, \ldots, X_n \sim p$.

Given $\hat{p}$ and "influence function" $(x_1, \ldots, x_m) \mapsto \chi_p(x_1, \ldots, x_m)$ use

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

for

$$\mathbb{U}_n f = \frac{(n-m)!}{n!} \sum_{1 \leq i_1 \neq \cdots \neq i_m \leq n} \cdots \sum f(X_{i_1}, \ldots, X_{i_m})$$

(Classical semiparametrics: $m = 1$ and $\mathbb{U}_n = \mathbb{P}_n$.)

Heuristics — plug in and bias correction

Estimate $\theta := \chi(p) \in \mathbb{R}$ from iid $X_1, \ldots, X_n \sim p$.

Given $\hat{p}$ and "influence function" $(x_1, \ldots, x_m) \mapsto \chi_p(x_1, \ldots, x_m)$ use

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

for

$$\mathbb{U}_n f = \frac{(n-m)!}{n!} \sum \cdots \sum_{1 \le i_1 \ne \cdots \ne i_m \le n} f(X_{i_1}, \ldots, X_{i_m})$$

(Classical semiparametrics: $m = 1$ and $\mathbb{U}_n = \mathbb{P}_n$.)

What is a good influence function?

Heuristics — plug in and bias correction

Estimate $\theta := \chi(p) \in \mathbb{R}$ from iid $X_1, \ldots, X_n \sim p$.

Given $\hat{p}$ and "influence function" $(x_1, \ldots, x_m) \mapsto \chi_p(x_1, \ldots, x_m)$ use

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

for

$$\mathbb{U}_n f = \frac{(n-m)!}{n!} \sum_{1 \leq i_1 \neq \cdots \neq i_m \leq n} \cdots \sum f(X_{i_1}, \ldots, X_{i_m})$$

(Classical semiparametrics: $m = 1$ and $\mathbb{U}_n = \mathbb{P}_n$.)

What is a good influence function that works with a general purpose p?

Heuristics — plug in and bias correction

Estimate $\theta := \chi(p) \in \mathbb{R}$ from iid $X_1, \ldots, X_n \sim p$.

Given $\hat{p}$ and "influence function" $(x_1, \ldots, x_m) \mapsto \chi_p(x_1, \ldots, x_m)$ use

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

for

$$\mathbb{U}_n f = \frac{(n-m)!}{n!} \sum_{1 \leq i_1 \neq \cdots \neq i_m \leq n} \cdots \sum f(X_{i_1}, \ldots, X_{i_m})$$

(Classical semiparametrics: $m = 1$ and $\mathbb{U}_n = \mathbb{P}_n$.)

What is a good influence function that works with a general purpose p?

$\chi_p = 0$ *gives* plug-in $\chi(\hat{p})$. *Not good!*

Heuristics — plug in and bias correction

If $\theta = \chi(p)$ is estimated by $\hat{\theta}_n = \chi(\hat{p}) + \mathbb{U}_n \chi_{\hat{p}}$, then

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

Heuristics — plug in and bias correction

If $\theta = \chi(p)$ is estimated by $\hat\theta_n = \chi(\hat p) + \mathbb{U}_n\chi_{\hat p}$, then

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

Construct χ_p such that $-P^m\chi_{\hat p_n}$ "represents" the first m terms of the Taylor expansion of $\chi(\hat p_n) - \chi(p)$:

$$P^m\chi_p = 0$$

$$\frac{d^j}{dt^j}_{|t=0}\chi(p_t) = -\frac{d^j}{dt^j}_{|t=0}P^m\chi_{p_t}, \qquad j = 1, \ldots, m$$

for "smooth" one-dimensional submodels $t \mapsto p_t$ with $p_0 = p$.

Heuristics — plug in and bias correction

If $\theta = \chi(p)$ is estimated by $\hat\theta_n = \chi(\hat p) + \mathbb{U}_n \chi_{\hat p}$, then

$$\hat\theta_n - \chi(p) = \big[\chi(\hat p_n) - \chi(p) + P^m \chi_{\hat p_n}\big] + (\mathbb{U}_n - P^n)\chi_{\hat p_n}.$$

Construct χ_p such that $-P^m \chi_{\hat p_n}$ "represents" the first m terms of the Taylor expansion of $\chi(\hat p_n) - \chi(p)$:

$$P^m \chi_p = 0$$

$$\frac{d^j}{dt^j}_{|t=0} \chi(p_t) = -\frac{d^j}{dt^j}_{|t=0} P^m \chi_{p_t}, \qquad j = 1, \ldots, m$$

for "smooth" one-dimensional submodels $t \mapsto p_t$ with $p_0 = p$.

This translates into inner products of influence function and scores.

Hasminskii and Ibragimov (1979), Nemirovski (2000) and others explored this in nonparametric models.

First-order estimator

For $m = 1$ we find the influence function χ_p from classical semiparametrics.

META THEOREM

First-order estimator $\hat{\theta} = \chi(\hat{p}) + \mathbb{U}_n \chi_{\hat{p}}$ satisfies

$$\hat{\theta} - \chi(p) = (\mathbb{U}_n - P)\chi_{\hat{p}} + \left[\chi(\hat{p}) - \chi(p) - (\hat{P} - P)\chi_{\hat{p}}\right]$$
$$= O_P\Big(\frac{1}{\sqrt{n}}\Big) + O_P\big(\|\hat{p} - p\|^2\big)$$

(Worst case scenario for bias)

Example: missing data

Observe $X = (YA, A, Z)$
Parameter $p \leftrightarrow (a, b, f)$

First-order estimator $\hat\theta = \chi(\hat p) + \mathbb{U}_n \chi_{\hat p}$ satisfies

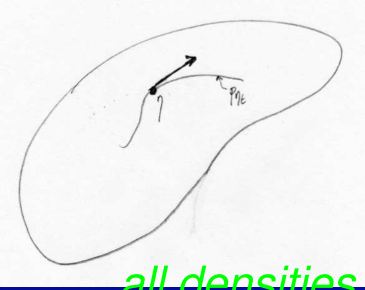
$$\hat\theta - \chi(p) = (\mathbb{U}_n - P)\chi_{\hat p} - \Big[\int (\hat a - a)(\hat b - b)\frac{f}{a}\Big]$$
$$= O_P\Big(\frac{1}{\sqrt{n}}\Big) + O_P\Big(\|\hat a - a\|\|\hat b - b\|\Big)$$

a	b	f	no bias if
$O_P(n^{-1/2})$	$o_P(1)$	$-$	$\dim(a) < \infty$
$o_P(1)$	$O_P(n^{-1/2})$	$-$	$\dim(b) < \infty$
$n^{-\alpha/(2\alpha+d)}$	$n^{-\alpha/(2\alpha+d)}$	$-$	$\alpha > d/2$
$\cdots$	$\cdots$	$-$	$\cdots$

If Z of high dimension, then bias of linear estimator dominates variance.

Higher-order tangent space and influence function

Tangent space of order m (at p): all higher order score functions of one-dimensional submodels $t \mapsto p_t$



all densities p

$$g(x_1,\ldots,x_m) = \frac{\frac{d^j}{dt^j}_{|t=0} \prod_{i=1}^m p_t(x_i)}{\prod_{i=1}^m p(x_i)}, \qquad j = 1,\ldots,m$$

(These are U-statistics)

Influence function of order m of $p \mapsto \chi(p)$ is map $(x_1,\ldots,x_m) \mapsto \chi_p(x_1,\ldots,x_m)$ with for all submodels $t \mapsto p_t$

$$\frac{d^j}{dt^j}_{|t=0} \chi(p_t) = \frac{d^j}{dt^j}_{|t=0} P^m \chi_p g, \qquad j = 1,\ldots,m$$

Higher-order influence function — computation

Influence function χ_p of parameter $p \mapsto \chi(p)$ can be computed recursively from its Hoeffding decomposition

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

- $\chi_p^{(1)}$ is a first order influence function of $p \mapsto \chi(p)$
- $x_j \mapsto \chi_p^{(j)}(x_1, \ldots, x_j)$ is a first order influence function of $p \mapsto \chi_p^{(j-1)}(x_1, \ldots, x_{j-1}) \qquad (j = 2, \ldots, m)$

(Optimal version may need projection in tangent space)

Higher-order estimator

Estimator $\hat\theta = \chi(\hat p) + \mathbb{U}_n \chi_{\hat p}$ with χ_p an mth order influence function

$$\begin{aligned}\hat\theta - \chi(p) &= (\mathbb{U}_n - P^n)\chi_{\hat p} + \big[\chi(\hat p) - \chi(p) - (\hat P^m - P^m)\chi_{\hat p}\big] \\ &= O_P\Big(\frac{1}{\sqrt{n}}\Big) + O_P\big(\|\hat p - p\|^{m+1}\big)\end{aligned}$$

Higher-order estimator

Estimator $\hat{\theta} = \chi(\hat{p}) + \mathbb{U}_n \chi_{\hat{p}}$ with χ_p an mth order influence function

$$\begin{aligned}\hat{\theta} - \chi(p) &= (\mathbb{U}_n - P^n)\chi_{\hat{p}} + \left[\chi(\hat{p}) - \chi(p) - (\hat{P}^m - P^m)\chi_{\hat{p}}\right] \\ &= O_P\Big(\frac{1}{\sqrt{n}}\Big) + O_P\big(\|\hat{p} - p\|^{m+1}\big)\end{aligned}$$

Free lunch??

Higher-order estimator

Estimator $\hat{\theta} = \chi(\hat{p}) + \mathbb{U}_n \chi_{\hat{p}}$ with χ_p an mth order influence function

$$\begin{aligned} \hat{\theta} - \chi(p) &= (\mathbb{U}_n - P^n)\chi_{\hat{p}} + \left[\chi(\hat{p}) - \chi(p) - (\hat{P}^m - P^m)\chi_{\hat{p}}\right] \\ &= O_P\Big(\frac{1}{\sqrt{n}}\Big) + O_P\big(\|\hat{p} - p\|^{m+1}\big) \end{aligned}$$

Free lunch??

No! *Higher order influence functions may not exist.*

Higher-order estimator

Estimator $\hat{\theta} = \chi(\hat{p}) + \mathbb{U}_n \chi_{\hat{p}}$ with χ_p an mth order influence function

$$\hat{\theta} - \chi(p) = (\mathbb{U}_n - P^n)\chi_{\hat{p}} + \left[\chi(\hat{p}) - \chi(p) - (\hat{P}^m - P^m)\chi_{\hat{p}}\right]$$
$$= O_P\Big(\frac{1}{\sqrt{n}}\Big) + O_P\big(\|\hat{p} - p\|^{m+1}\big)$$

Free lunch??

No! *Higher order influence functions may not exist.*

Must use *approximations* (representing derivative in selected directions).

META META THEOREM

mth-order estimator $\hat{\theta} = \chi(\hat{p}) + \mathbb{U}_n \chi_{\hat{p}}$ satisfies

$$\hat{\theta} - \chi(p) = (\mathbb{U}_n - P^n)\chi_{n,\hat{p}} + O_P\big(\|\hat{p} - p\|^{m+1}\big) + \text{approximation bias}$$

One variance term and two bias terms.

Approximate functional

Choose a map $p \mapsto \tilde{p}$ of the model onto a "smaller" model and consider

$$\tilde{\chi}(p) = \chi(\tilde{p}) + P\chi_{\tilde{p}}^{(1)}$$

Definition of $\chi_p^{(1)}$ suggests

$$\tilde{\chi}(p) - \chi(p) = O\big(\|\tilde{p} - p\|^2\big)$$

Choose $p \mapsto \tilde{p}$ such that, for any path $t \mapsto p_t$,

$$\frac{d}{dt}_{|t=0}\Big(\chi(\tilde{p}_t) + P_0\chi_{\tilde{p}_t}^{(1)}\Big) = 0$$

Then $\tilde{\chi}_p^{(1)} = \chi_{\tilde{p}}^{(1)}$ and $\tilde{\chi}$ ought to have influence functions to any order.

Example: missing data — approximate functional

Define $\tilde{\chi}(p) := \chi(\tilde{p})$, for $p \mapsto \tilde{p}$ given by projection onto $L \subset L_2(g)$:

$$(a, b, g) \mapsto (\tilde{a}, \tilde{b}, g) \in L \times L \times \{g\}$$

THEOREM

For $\Pi_{i,j} = \Pi_p(Z_i, Z_j)$ *projection kernel* on $L \subset L_2(g)$, $\tilde{Y} = A\big(Y - \tilde{b}(Z)\big)$, $\tilde{A} = A\tilde{a}(Z) - 1$,

$$\tilde{\chi}_p^{(1)}(X) = A\tilde{a}(Z)\big(Y - \tilde{b}(Z)\big) + \tilde{b}(Z) - \chi\big(\tilde{p}\big)$$

$$\tilde{\chi}_p^{(2)}(X_1, X_2) = -2[\tilde{Y}_1 \Pi_{1,2} \tilde{A}_2]$$

$$\tilde{\chi}_p^{(3)}(X_1, X_2, X_3) = 6\Big[\tilde{Y}_1 \Pi_{1,2} A_2 \Pi_{2,3} \tilde{A}_3 - \tilde{Y}_1 \Pi_{1,3} \tilde{A}_3\Big]$$

$$\tilde{\chi}_p^{(4)}(X_1, X_2, X_3, X_4) = -24\Big[\tilde{Y}_1 \Pi_{1,2} A_2 \Pi_{2,3} A_3 \Pi_{3,4} \tilde{A}_4 - \tilde{Y}_1 \Pi_{1,3} A_3 \Pi_{3,4} \tilde{A}_4 - \tilde{Y}_1 \Pi_{1,2} A_2 \Pi_{2,4} \tilde{A}_4 + \tilde{Y}_1 \Pi_{1,4} \tilde{A}_4\Big]$$

etc.

Projection kernel — von Mises calculus

A projection $\Pi_g \colon L_2(g) \to L$ onto a finite-dimensional space can be represented as

$$\Pi_g h(x) = \int h(y)\Pi_g(x,y)\, g(y)\, d\nu(y)$$

$y \mapsto \Pi_g(x,y)$ restricted to L works as Dirac kernel at x, because equivalent:

- $\Pi_g h = h$
- $h \in L$
- $h(x) = \int h(y)\Pi_g(x,y)\, d\nu(y)$ a.e. x

Representation would be exact if $L = L_2(g)$, i.e. Π_g is the "Dirac kernel on the diagonal", but this does not exist.

Example: missing data – parametric rate – mth-order

THEOREM

For $\sup_x \Pi_p(x,x) \lesssim k$,

$$\hat{\mathrm{E}}_p\hat\theta_n - \chi(p) = O\Big(\|\hat a - a\|_r \|\hat b - b\|_r \|\hat g - g\|_{(m-1)r/(r-2)}^{m-1}\Big) + O\Big(\Big\|(I-\Pi_p)(\hat a - a)\Big\|_2 \Big\|(I-\Pi_p)(\hat b - b)\Big\|_2\Big),$$

$$\widehat{\mathrm{var}}_p\hat\chi_n \le \sum_{j=1}^m \frac{1}{\binom{n}{j}} c^j k^{j-1}.$$

If $(\alpha+\beta)/2 \ge d/4$ obtain $\sqrt{n}$-rate by choosing

- large enough order m.
- L optimal for approximation in Hölder spaces, of dimension $k = n/(\log n)^2$.
- $\hat a$, $\hat b$, $\hat g$ that attain uniform minimax rates $(\log n/n)^{-\delta/(2\delta+d)}$.

If $(\alpha+\beta)/2 > d/4$ obtain even efficiency $\sqrt{n}\big(\hat\chi_n - \chi(p) - \mathbb{P}_n\chi_p^{(1)}\big) \xrightarrow{P} 0$.

Example: missing data – parametric rate – mth-order

THEOREM

For $\sup_x \Pi_p(x,x) \lesssim k$,

$$\hat{\mathrm{E}}_p \hat{\theta}_n - \chi(p) = O\Big(\|\hat{a}-a\|_r \|\hat{b}-b\|_r \|\hat{g}-g\|^{m-1}_{(m-1)r/(r-2)}\Big) + O\Big(\Big\|(I-\Pi_p)(\hat{a}-a)\Big\|_2 \Big\|(I-\Pi_p)(\hat{b}-b)\Big\|_2\Big),$$

$$\hat{\mathrm{var}}_p \hat{\chi}_n \leq \sum_{j=1}^m \frac{1}{\binom{n}{j}} c^j k^{j-1}.$$

If $(\alpha+\beta)/2 \geq d/4$ obtain $\sqrt{n}$-rate by choosing

- large enough order m.
- L optimal for approximation in Hölder spaces, of dimension $k = n/(\log n)^2$.
- $\hat{a}$, $\hat{b}$, $\hat{g}$ that attain uniform minimax rates $(\log n/n)^{-\delta/(2\delta+d)}$.

If $(\alpha+\beta)/2 > d/4$ obtain even efficiency $\sqrt{n}\big(\hat{\chi}_n - \chi(p) - \mathbb{P}_n \chi_p^{(1)}\big) \xrightarrow{P} 0$.

Linear estimator ($m=1$) works only if $(\alpha+\beta)/2 \geq d/2$.

Example: missing data — lower smoothness — 3rd-order IF

Leading part of 3rd-order part of 3rd-order influence function of $\tilde{\chi}$ is

$$6\tilde{A}_1 \Pi_p(Z_1, Z_2) A_2 \Pi_p(Z_2, Z_3) \tilde{Y}_3.$$

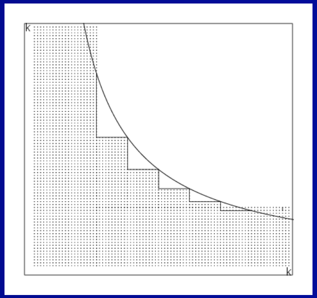
Decompose, for $k_{-1} = l_{-1} = 1$ and $k_R \sim l_S \sim k$,

$$\Pi_p = \sum_{r=0}^{R} \Pi_p^{(k_{r-1}, k_r]}, \qquad \Pi_p = \sum_{s=0}^{S} \Pi_p^{(l_{s-1}, l_s]}$$

and replace preceding display by

$$6 \sum_{\substack{(r,s): r+s \le D \\ \vee r=0 \vee s=0}} \sum \tilde{A}_1 \Pi_p^{(k_{r-1}, k_r]}(Z_1, Z_2) A_2 \Pi_p^{(l_{s-1}, l_s]}(Z_2, Z_3) \tilde{Y}_3.$$

$$k_r \sim n 2^{r/\alpha}, \qquad r = 0, \ldots, R,$$
$$l_s \sim n 2^{s/\beta}, \qquad s = 0, \ldots, S.$$



Example: missing data — lower smoothness — higher order

Replace

$$\tilde{Y}_1 \Pi_{1,2} A_2 \Pi_{2,3} A_3 \times \cdots \times A_{j-1} \Pi_{j-1,j} \tilde{A}_j$$

by

$$\sum_{i=1}^{j-2} j!(-1)^{j-1}\, \tilde{Y}_1 \Pi_{1,2}^{(0,n]} A_2 \times \cdots \times A_{i-1} \Pi_{i-1,i}^{(0,n]} A_i \times$$
$$\times \Big[\sum\sum_{\substack{(r,s): r+s \le D \\ \vee r=0 \vee s=0}} \Pi_{i,i+1}^{(k_{r-1},k_r]} A_{i+1} \Pi_{i+1,i+2}^{(l_{s-1},l_s]} \Big] \times$$
$$\times A_{i+2} \Pi_{i+2,i+3}^{(0,n]} \times \cdots \times A_{j-1} \Pi_{j-1,j}^{(0,n]} \tilde{A}_j .$$

Example: missing data — lower smoothness — higher order

THEOREM

For $\sup_x \Pi_{\hat p}^{(0,l]}(x,x) \lesssim l$, the pruned mth order estimator satisfies (for $r \geq 2$)

$$\hat{\mathrm{E}}_p \hat\chi_n - \chi(p) = O\Big(\|\hat a - a\|_r \|\hat b - b\|_r \|\hat g - g\|_{\frac{mr}{r-2}}^{m-1}\Big)$$
$$+ O\Big(\Big\|(I - \Pi_p^{(0,k]})(\hat a - a)\Big\|_2 \Big\|(I - \Pi_p^{(0,k]})(\hat b - b)\Big\|_2\Big),$$
$$+ O\Big(\sum_{r=1}^{R} \Big\|(I - \Pi_{\hat p}^{(0,k_{r-1}]})(\hat a - a)\Big\|_r \Big\|(I - \Pi_{\hat p}^{(0,l_{D-r}]})(\hat b - b)\Big\|_r \|\hat g - g\|_{\frac{r}{r-2}}\Big)$$
$$+ O\Big(R\Big\|(I - \Pi_{\hat p}^{(0,n]})(\hat a - a)\Big\|_r \Big\|(I - \Pi_{\hat p}^{(0,n]})(\hat b - b)\Big\|_r \|\hat g - g\|_{\frac{mr}{r-2}}^{2}\Big),$$
$$\hat{\mathrm{var}}_p \hat\chi_n \lesssim \frac{1}{n} + \frac{k}{n^2} + \frac{D 2^{(\frac{1}{\alpha} \vee \frac{1}{\beta})D}}{n}.$$

If $\phi > \phi(\alpha, \beta)$ obtain rate $n^{-(2\alpha+2\beta)/(2\alpha+2\beta+d)}$,
with sufficiently large m, suitable D and suitable initial estimators.

$\phi(\alpha, \beta) = (\alpha/d \vee \beta/d)(d - 2\alpha - 2\beta)/(d + 2\alpha + 2\beta)$

Lower Bounds

Classical semiparametrics

In classical semiparametrics the rate of estimation is $\sqrt{n}$.

The best limit distribution of $\sqrt{n}\big(T_n - \chi(p)\big)$ is normal, with variance equal to the inverse efficient Fisher information.

Slow rates — testing argument (Le Cam)

$X_1, X_2, \ldots, X_n$ i.i.d. sample from density $p \in \mathcal{P}$.

THEOREM
If P_n and Q_n are in the convex hulls of the sets of measures $\{P^n \colon p \in \mathcal{P}, \chi(p) \le 0\}$ and $\{P^n \colon p \in \mathcal{P}, \chi(p) \ge \varepsilon_n\}$, and

$$\rho(P_n, Q_n) \colon= \int \sqrt{dP_n}\sqrt{dQ_n} \gg 0$$

then the rate is not faster than ε_n.

Nontrivial details:

- find the least favourable P_n and Q_n.
- compute their Hellinger affinity $\rho(P_n, Q_n)$.

Affinity bound (Birgé and Massart (1995), Robins and vdV (2008))

- Partition $\mathcal{X} = \cup_{j=1}^k \mathcal{X}_j$.
- Perturbation parameter $\lambda = (\lambda_1, \ldots, \lambda_k)$ with prior $\pi = \pi_1 \otimes \cdots \otimes \pi_k$.
- P_λ and Q_λ probability measures on $\mathcal{X}$ such that restrictions $P_{\lambda|\mathcal{X}_j}$ and $Q_{\lambda|\mathcal{X}_j}$ depend on λ_j only and have equal mass p_j.

THEOREM If $np_j(1 \vee a \vee b) \lesssim 1$ and $0 \lesssim p_\lambda \lesssim 1$, then

$$\rho\Big(\int P_\lambda^n\, d\pi(\lambda), \int Q_\lambda^n\, d\pi(\lambda)\Big) \geq 1 - Cn^2(\max_j p_j)(b^2 + ab) - Cnd.$$

$$p = \int p_\lambda\, d\pi(\lambda)$$

$$q = \int q_\lambda\, d\pi(\lambda)$$

$$a = \max_j \sup_\lambda \int_{\mathcal{X}_j} \frac{(p_\lambda - p)^2}{p_\lambda} \frac{d\nu}{p_j},$$

$$b = \max_j \sup_\lambda \int_{\mathcal{X}_j} \frac{(q_\lambda - p_\lambda)^2}{p_\lambda} \frac{d\nu}{p_j},$$

$$d = \max_j \sup_\lambda \int_{\mathcal{X}_j} \frac{(q - p)^2}{p_\lambda} \frac{d\nu}{p_j}.$$

Example: missing data

Covariate Z, $\quad \sim f$
Response Y, with $Y|\,Z \sim$ binomial $\bigl(1, b(Z)\bigr)$
Missingness indicator A, with $A|\,Z \sim$ binomial $\bigl(1, 1/a(Z)\bigr)$
Missing at random: $Y \perp\!\!\!\perp A|\,Z$
Observe $X = (YA, A, Z) \in \{0,1\} \times \{0,1\} \times [0,1]^d$

We wish to estimate mean response $\chi(a,b,f) = \int b\, f\, d\nu = \mathrm{E}Y$.

THEOREM
If a, b, and g belong to Hölder classes $C^\alpha[0,1]^d$, $C^\beta[0,1]^d$, $C^\gamma[0,1]^d$, then the rate of estimation is not faster than $n^{-(\alpha+\beta)/(2\alpha+2\beta+d)}$.

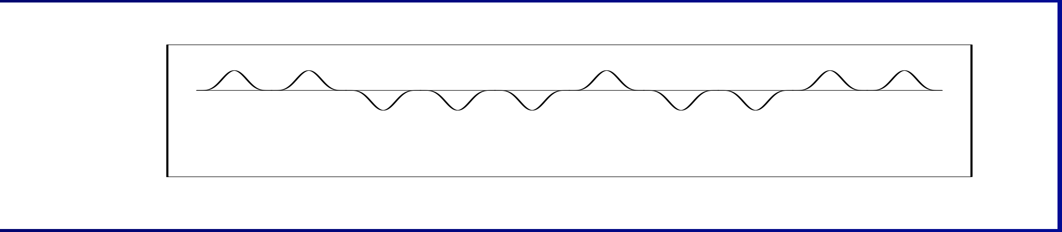
Perturbations

- $H\colon \mathbb{R}^d \to \mathbb{R}$, C^∞, support $\subset [0,1/2]^d$, $\int H\,d\nu = 0$.
- $k \sim n^{2d/(2\alpha+2\beta+d)}$.
- $\mathcal{X}_j = \{0,1\} \times \{0,1\} \times \mathcal{Z}_j$, for $\mathcal{Z}_j$ disjoint translates of $k^{-1/d}[0,1/2]^d$.
- π uniform on $\Lambda := \{0,1\}^k$.

For $\lambda = (\lambda_1, \ldots, \lambda_k) \in \Lambda$:

$$a_\lambda(z) = 2 + \Big(\frac{1}{k}\Big)^{\alpha/d} \sum_{j=1}^k \lambda_j H\big((z-z_j)k^{1/d}\big),$$

$$b_\lambda(z) = \frac{1}{2} + \Big(\frac{1}{k}\Big)^{\beta/d} \sum_{j=1}^k \lambda_j H\big((z-z_j)k^{1/d}\big).$$



Perturbations–2

- $\alpha \le \beta$: $p_\lambda \leftrightarrow (a_\lambda, 1/2, 1/2)$ and $q_\lambda \leftrightarrow (a_\lambda, b_\lambda, 1/2)$.
- $\alpha \ge \beta$: $p_\lambda \leftrightarrow (2, b_\lambda, 1/2)$ and $q_\lambda \leftrightarrow (a_\lambda, b_\lambda, 1/2)$.

This leads to comparing the functional $\chi(a, b, g)$ on two mixtures, where

- the first mixture $\int P_\lambda^n \, d\pi(\lambda)$ perturbs only the coarsest of the two parameters a and b.
- the second mixture $\int Q_\lambda^n \, d\pi(\lambda)$ perturbs both parameters.

(The third parameter g is always taken $1/2$.)

Concluding remarks

Outlook

Adaptation to α and β.

Implementation.

Prior parameter clases defined by sparsity.

Other models, e.g. semiparametric regression.