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

**META THEOREM**

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

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

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

*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 $\varepsilon$ with symmetric density $\eta$, Fisher information $I_\eta < \infty$

Observe $X = \theta + \varepsilon$

THEOREM
There exists $T_n = T_n(X_1, \ldots, X_n)$ with, for all $(\theta, \eta)$,

$$\sqrt{n}(T_n - \theta) \xrightarrow{d} N(0, I_\eta^{-1})$$
**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 $(\theta, \lambda, f)$,

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

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

Covariates \((W, Z)\), density \(f\)

Error \(\varepsilon\), such that \(\varepsilon | (W, Z)\) has density \(g(\cdot | W, Z)\) with \(E(\varepsilon | W, Z) = 0\)

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 \((\theta, f, g, \eta)\) such that \(g\) and \(\eta\) are sufficiently smooth,

\[
\sqrt{n}(T_n - \theta) \leadsto N\left(0, \sigma_{\theta,f,g,\eta}^2\right)
\]
Example: missing data (Robins and Rotnitzky, ...)

Covariate $Z$, $\sim f$

Response $Y$, with $Y|Z \sim \text{binomial}(1, b(Z))$

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

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

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

We wish to estimate mean response $\chi(a, b, f) = \int b f \, d\nu = EY$. 
Example: missing data \cite{Robins+Rotnitzky}

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

We wish to estimate mean response $\chi(a, b, f) = \int b f \, d\nu = EY$.

$Y$ is observed only if $A = 1$
$Z$ is included to make the assumption $Y \perp \perp A \mid Z$ realistic
Example: missing data (Robins and Rotnitzky, ...)

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

We wish to estimate mean response $\chi(a, b, f) = \int b f \, dv = EY$.

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

$$\sqrt{n}(T_n - \chi(a, b, f)) \overset{d}{\to} N\left(0, \sigma_{a,b,f}^2\right)$$
Classical semiparametrics — minimal variance

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

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

**META THEOREM**
If \( \mathcal{P} \) and \( \chi \) are nice, then there exist \( T_n = T_n(X_1, \ldots, X_n) \), with

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

What is the minimal variance \( \sigma_p^2 \)?
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 $p_0 = p$
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 $p_0 = p$

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).)

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

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

for $\Pi_p$ orthogonal projection onto closed linear span of tangent set.
Example: missing data

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

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

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

\[
\frac{Aa(Z) - 1}{a(Z)(a - 1)(Z)} \alpha(Z) \quad \text{a-score, } a_t = a + t \alpha
\]

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

\[
\frac{\phi(Z)}{\phi(Z)} \quad \text{f-score, } f_t = f(1 + t \phi), \quad \int \phi f = 0
\]
Example: missing data

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

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

\[
\frac{Aa(Z) - 1}{a(Z)(a - 1)(Z)} \alpha(Z) \quad \text{a-score, } a_t = a + t\alpha
\]
\[
\frac{A(Y - b(Z))}{b(Z)(1 - b)(Z)} \beta(Z) \quad \text{b-score, } b_t = b + t\beta
\]
\[
\frac{\phi(Z)}{\phi(Z)} \quad \text{f-score, } f_t = f(1 + t\phi)
\]

Parameter of interest $\chi(p) = \int bf = EY$
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 $f(Z)(1/a)(Z)^A(1 - 1/a(Z))^{1-A}b(Z)^YA(1 - b(Z))^{(1-Y)A}$

Parameter of interest $\chi(p) = \int bf = EY$

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

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

$E_p\chi_p(X)[b\text{-score}] = \frac{\partial}{\partial t}|_{t=0} \int b_tf d\nu = \int \beta f d\nu$

$E_p\chi_p(X)[f\text{-score}] = \frac{\partial}{\partial t}|_{t=0} \int bf_t d\nu = \int b\phi f d\nu$
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

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

(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

$$\hat{\theta} = \chi(\hat{p}) + U_n \chi(\hat{p}),$$

for

$$U_n f = \frac{(n - m)!}{n!} \sum \cdots \sum f(X_{i_1}, \ldots, X_{i_m})$$

where

$$f(x_1, \ldots, x_m) = \frac{1}{n} \sum \cdots \sum f(X_{i_1}, \ldots, X_{i_m})$$

(Classical semiparametrics: $m = 1$ and $U_n = P_n$.)

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

$$\hat{\theta} = \chi(\hat{p}) + U_n \chi(\hat{p}),$$

for

$$U_n f = \frac{(n - m)!}{n!} \sum_{1 \leq i_1 \neq \ldots \neq i_m \leq n} f(X_{i_1}, \ldots, X_{i_m}).$$

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

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

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

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

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

$\chi_p = 0$ gives plug-in $\chi(\hat{p})$. Not good!
If \( \theta = \chi(p) \) is estimated by \( \hat{\theta}_n = \chi(\hat{p}) + U_n \chi \hat{p} \), then

\[
\hat{\theta}_n - \chi(p) = \left[ \chi(\hat{p}_n) - \chi(p) + P^m \chi \hat{p}_n \right] + (U_n - P^n) \chi \hat{p}_n.
\]
If $\theta = \chi(p)$ is estimated by $\hat{\theta}_n = \chi(\hat{p}) + U_n \chi \hat{p}$, then

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

Construct $\chi_p$ such that $-P^m \chi \hat{p}_n$ “represents” the first $m$ terms of the Taylor expansion of $\chi(\hat{p}_n) - \chi(p)$:

$$P^m \chi_p = 0$$

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

for “smooth” one-dimensional submodels $t \mapsto p_t$ with $p_0 = p$. 
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 $\chi_p$ such that $-P^m \chi_{\hat{p}_n}$ “represents” the first $m$ terms of the Taylor expansion of $\chi(\hat{p}_n) - \chi(p)$:

$$P^m \chi_p = 0$$

$$\frac{d^j}{dt^j} \chi(p_t) = -\frac{d^j}{dt^j} P^m \chi_{p_t}, \quad j = 1, \ldots, m$$

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 $\chi_p$ from classical semiparametrics.

**META THEOREM**

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

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

$$= O_P \left( \frac{1}{\sqrt{n}} \right) + O_P \left( \|\hat{p} - p\|^2 \right)$$

(Worst case scenario for bias)
Example: missing data

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

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

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

$$\hat{\theta} - \chi(p) = (\mathbb{U}_n - P)\chi\hat{p} - \left[ \int (\hat{a} - a)(\hat{b} - b) \frac{f}{a} \right]$$

$$= O_P\left(\frac{1}{\sqrt{n}}\right) + O_P\left(\|\hat{a} - a\|\|\hat{b} - b\|\right)$$

<table>
<thead>
<tr>
<th>$a$</th>
<th>$b$</th>
<th>$f$</th>
<th>no bias if</th>
</tr>
</thead>
<tbody>
<tr>
<td>$O_P(n^{-1/2})$</td>
<td>$o_P(1)$</td>
<td>$-\cdot$</td>
<td>$\dim(a) &lt; \infty$</td>
</tr>
<tr>
<td>$o_P(1)$</td>
<td>$O_P(n^{-1/2})$</td>
<td>$-\cdot$</td>
<td>$\dim(b) &lt; \infty$</td>
</tr>
<tr>
<td>$n^{-\alpha/(2\alpha+d)}$</td>
<td>$n^{-\alpha/(2\alpha+d)}$</td>
<td>$-\cdot$</td>
<td>$\alpha &gt; d/2$</td>
</tr>
<tr>
<td>$\ldots$</td>
<td>$\ldots$</td>
<td>$-\cdot$</td>
<td>$\ldots$</td>
</tr>
</tbody>
</table>

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 $t \mapsto p_t$

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

(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 $t \mapsto p_t$

$$\frac{d^j}{dt^j} \chi(p_t) \bigg|_{t=0} = \frac{d^j}{dt^j} \bigg|_{t=0} P^m \chi_p g, \quad j = 1, \ldots, m$$
Higher-order influence function — computation

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

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

- $\chi_p^{(1)}$ 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 $p \mapsto \chi_p^{(j-1)}(x_1, \ldots, x_{j-1}) \quad (j = 2, \ldots, m)$

*(Optimal version may need projection in tangent space)*
Higher-order estimator

Estimator $\hat{\theta} = \chi(\hat{p}) + U_n \chi \hat{p}$ with $\chi_p$ an $m$th order influence function

$$\hat{\theta} - \chi(p) = (U_n - P^n) \chi \hat{p} + \left[ \chi(\hat{p}) - \chi(p) - (\hat{P}^m - P^m) \chi \hat{p} \right]$$

$$= O_P \left( \frac{1}{\sqrt{n}} \right) + O_P \left( \|\hat{p} - p\|^{m+1} \right)$$
Higher-order estimator

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

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

$$= O_P \left( \frac{1}{\sqrt{n}} \right) + O_P \left( \|\hat{p} - p\|^{m+1} \right)$$

Free lunch??
Higher-order estimator

Estimator \( \hat{\theta} = \chi(\hat{p}) + \mathbb{U}_n \chi_{\hat{p}} \) with \( \chi_p \) an \( m \)th order influence function

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

\[
= O_P \left( \frac{1}{\sqrt{n}} \right) + O_P \left( \|\hat{p} - p\|^{m+1} \right)
\]

Free lunch??

No! Higher order influence functions may not exist.
Higher-order estimator

Estimator $\hat{\theta} = \chi(\hat{p}) + U_n \chi_{\hat{p}}$ with $\chi_{\hat{p}}$ an $m$th order influence function

$$\hat{\theta} - \chi(p) = (U_n - P^n) \chi_{\hat{p}} + [\chi(\hat{p}) - \chi(p) - (\hat{P}^m - P^m) \chi_{\hat{p}}]$$

$$= O_P \left( \frac{1}{\sqrt{n}} \right) + O_P \left( \|\hat{p} - p\|^{m+1} \right)$$

Free lunch??
No! Higher order influence functions may not exist.

Must use approximations (representing derivative in selected directions).

**META META THEOREM**

$m$th-order estimator $\hat{\theta} = \chi(\hat{p}) + U_n \chi_{\hat{p}}$ satisfies

$$\hat{\theta} - \chi(p) = (U_n - P^n) \chi_{\hat{p}} + O_P \left( \|\hat{p} - p\|^{m+1} \right) + \text{approximation bias}$$

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

\[
\tilde{\chi}(p) = \chi(\tilde{p}) + P\chi^{(1)}_{\tilde{p}}
\]

Definition of \( \chi^{(1)}_{\tilde{p}} \) suggests

\[
\tilde{\chi}(p) - \chi(p) = O(\|\tilde{p} - p\|^2)
\]

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

\[
\frac{d}{dt} \bigg|_{t=0} \left( \chi(\tilde{p}_t) + P_0\chi^{(1)}_{\tilde{p}_t} \right) = 0
\]

Then \( \tilde{\chi}^{(1)}_p = \chi^{(1)}_{\tilde{p}} \) and \( \tilde{\chi} \) ought to have influence functions to any order.
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 \( L \subset L_2(g) \), \( \tilde{Y} = A(Y - \tilde{b}(Z)) \), \( \tilde{A} = A\tilde{a}(Z) - 1 \),

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

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

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

\[
\tilde{\chi}_p^{(4)}(X_1, X_2, X_3, X_4) = -24[\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]
\]

etc.
Projection kernel — von Mises calculus

A projection $\Pi_g: L_2(g) \to L$ 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. $\Pi_g$ is the “Dirac kernel on the diagonal”, but this does not exist.
Example: missing data – parametric rate – $m$th-order

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

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

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

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 $\sqrt{n} \left(\hat{\chi}_n - \chi(p) - \mathbb{P}_n \chi_p^{(1)}\right) \xrightarrow{P} 0$. 

24 / 36
Example: missing data – parametric rate – \( m \)th-order

\[ \text{THEOREM} \]

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

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

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

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 \( \sqrt{n} (\hat{\chi}_n - \chi(p) - \mathbb{P}_n \chi_p^{(1)}) \overset{P}{\to} 0 \).

**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 $k_R \sim l_S \sim k$,

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

and replace preceding display by

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

$$k_r \sim n2^{r/\alpha}, \quad r = 0, \ldots, R,$$

$$l_s \sim n2^{s/\beta}, \quad s = 0, \ldots, S.$$
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

$$\sum_{i=1}^{j-2} j! (-1)^{j-1} \tilde{Y}_1 \Pi_{1,2}^{(0,n]} A_2 \times \cdots \times A_{i-1} \Pi_{i-1,i}^{(0,n]} A_i \times$$

$$\times \left[ \sum_{(r,s):r+s\leq D} \sum_{\forall r=0 \forall s=0} \Pi_{i,i+1}^{(k_r-1,k_r]} A_{i+1} \Pi_{i+1,i+2}^{(l_s-1,l_s]} \right] \times$$

$$\times A_{i+2} \Pi_{i+2,i+3}^{(0,n]} \times \cdots \times A_{j-1} \Pi_{j-1,j}^{(0,n]} \tilde{A}_j.$$
THEOREM
For $\sup_x \Pi_p^{(0,l]}(x, x) \lesssim l$, the pruned $m$th order estimator satisfies (for $r \geq 2$)

$$\hat{\var} \hat{\chi}_n - \chi(p) = O\left(\|\hat{a} - a\|_r \|\hat{b} - b\|_r \|\hat{g} - g\|_{\frac{m-1}{m r-2}}\right)$$

$$+ O\left(\left\|\left(I - \Pi_p^{(0,k]}(\hat{a} - a)\right)\left(I - \Pi_p^{(0,k]}(\hat{b} - b)\right)\right\|_2\right),$$

$$+ O\left(\sum_{r=1}^{R} \left\|\left(I - \Pi_p^{(0,k, r-1]}(\hat{a} - a)\right)\left(I - \Pi_p^{(0,lD-r]}(\hat{b} - b)\right)\right\|_r \|\hat{g} - g\|_{\frac{r-2}{r-2}}\right),$$

$$+ O\left(R \left\|\left(I - \Pi_p^{(0,n]}(\hat{a} - a)\right)\left(I - \Pi_p^{(0,n]}(\hat{b} - b)\right)\right\|_r \|\hat{g} - g\|_{\frac{m r}{m r-2}}\right),$$

$$\var \hat{\chi}_n \lesssim \frac{1}{n} + \frac{k}{n^2} + \frac{D2^{\left(\frac{1}{\alpha} \lor \frac{1}{\beta}\right)D}}{n}.$$

If $\phi > \phi(\alpha, \beta)$ obtain rate $n^{-\left(\frac{2\alpha + 2\beta}{2\alpha + 2\beta + d}\right)}$,

with sufficiently large $m$, suitable $D$ and suitable initial estimators.

$$\phi(\alpha, \beta) = (\alpha/d \lor \beta/d)(d - 2\alpha - 2\beta)/(d + 2\alpha + 2\beta)$$
Lower Bounds
In classical semiparametrics the rate of estimation is $\sqrt{n}$.

The best limit distribution of $\sqrt{n}(T_n - \chi(p))$ 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 $\{P^n: p \in \mathcal{P}, \chi(p) \leq 0\}$ and $\{P^n: p \in \mathcal{P}, \chi(p) \geq \varepsilon_n\}$, and

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

then the rate is not faster than $\varepsilon_n$.

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

**Theorem** If $np_j(1 \lor a \lor b) \lesssim 1$ and $0 \lesssim p_\lambda \lesssim 1$, then

$$
\rho \left( \int P_\lambda^n d\pi(\lambda), \int Q_\lambda^n d\pi(\lambda) \right) \geq 1 - Cn^2 \left( \max_j p_j \right) (b^2 + ab) - C'nd.
$$

\[ a = \max_j \sup_{\lambda} \int_{\mathcal{X}_j} \frac{(p_\lambda - p)^2}{p_\lambda} \frac{d\nu}{p_j}, \]
\[ b = \max_j \sup_{\lambda} \int_{\mathcal{X}_j} \frac{(q_\lambda - p_\lambda)^2}{p_\lambda} \frac{d\nu}{p_j}, \]
\[ d = \max_j \sup_{\lambda} \int_{\mathcal{X}_j} \frac{(q - p)^2}{p_\lambda} \frac{d\nu}{p_j}. \]
Example: missing data

Covariate $Z$, $\sim f$
Response $Y$, with $Y|Z \sim \text{binomial } (1, b(Z))$
Missingness indicator $A$, with $A|Z \sim \text{binomial } (1, 1/a(Z))$
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 $\chi(a, b, f) = \int b f \, d\nu = EY$.

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

- \( H : \mathbb{R}^d \to \mathbb{R}, \ C^\infty, \) 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 \( k^{-1/d}[0, 1/2]^d. \)
- \( \pi \) uniform on \( \Lambda := \{0, 1\}^k. \)

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

\[
\begin{align*}
    a_\lambda(z) &= 2 + \left(\frac{1}{k}\right)^{\alpha/d} \sum_{j=1}^{k} \lambda_j H((z - z_j)k^{1/d}), \\
    b_\lambda(z) &= \frac{1}{2} + \left(\frac{1}{k}\right)^{\beta/d} \sum_{j=1}^{k} \lambda_j H((z - z_j)k^{1/d}).
\end{align*}
\]
\( \alpha \leq \beta: \) \( p_\lambda \leftrightarrow (a_\lambda, 1/2, 1/2) \) and \( q_\lambda \leftrightarrow (a_\lambda, b_\lambda, 1/2) \).

\( \alpha \geq \beta: \) \( p_\lambda \leftrightarrow (2, b_\lambda, 1/2) \) 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 \( \int P^n_\lambda \ d\pi(\lambda) \) perturbs only the coarsest of the two parameters \( a \) and \( b \).
- the second mixture \( \int Q^n_\lambda \ d\pi(\lambda) \) perturbs both parameters.

(The third parameter \( g \) is always taken 1/2.)
Concluding remarks
Outlook

Adaptation to $\alpha$ and $\beta$.
Implementation.
Prior parameter classes defined by sparsity.
Other models, e.g. semiparametric regression.