Semiparametric Bayesian estimation, with or without bias

Kolyan Ray and Aad van der Vaart
King's College London and Universiteit Leiden

BNP12, Oxford, June 2019

Outline

Questions

- on priors on semiparametric models

- on Bayesian inference versus “any” inference

Some results

Amine

Amine

Home Office

Web www.gov.uk/ukvi

VAF no: UKVS / 1398611

GWF no:

Date 21 May 2019

Dear MOHAMED AMINE HADJI,

Your application for a visit visa to the United Kingdom has been refused.

What this means for you

Any future UK visa applications you make will be considered on their individual merits, however you are likely to be refused unless the circumstances of your application change.

In relation to this decision, there is no right of appeal or right to administrative review.

The reasons for this decision are set out on the next page.

Yours sincerely,

GG
UKVS

Amine

REASONS FOR REFUSAL **NRA v 1.0**

You have applied for a visa to visit the UK.

In deciding whether you meet the requirements of Appendix V: of the Immigration Rules for visitors (https://www.gov.uk/guidance/immigration-rules/immigration-rules-appendix-v-visitor-rules), I have considered:

- your application and any additional relevant information you have provided with it
- your immigration history

The decision
I have refused your application for a visit visa because I am not satisfied that you meet the requirements of Appendix V: Immigration Rules for Visitors because:

You have applied for entry clearance for 7 days.

I note that you previously applied for a visit visa and that your application was refused on 21/06/2018.

I also note the reasons for that refusal and the documents and comments you have now submitted in support of your current application. It was noted on the previous refusal that any future UK visa applications you made would be considered on their individual merits, however you were likely to be refused unless the circumstances of your application changed.

I have considered your application carefully but I find that your circumstances have not changed since your last refusal and you have chosen not to address the reasons given by the Entry Clearance Officer for your previous refusal with this application.

I am not satisfied that you have accurately presented your circumstances or intentions in wishing to enter the United Kingdom. This means that I am not satisfied that you are genuinely seeking entry as a visitor or that you intend to leave the United Kingdom at the end of your proposed visit; or that you have access to sufficient funds to cover all reasonable costs in relation to your visit without working or accessing public funds (this includes, but is not exclusively, the cost of the return or onward journey, any costs relating to dependants, and the cost of planned activities such as private medical treatment) Paragraph V 4.2 (a) (c) (e) of the Immigration Rules.

Amine

NEXT STEPS **NRA v 1.0**

In relation to this decision, there is no right of appeal or right to administrative review.

https://www.homeofficesurveys.homeoffice.gov.uk/s/visasurveyuk

Questions

Missing data

Outcome Y, observed if $A = 1$, unobserved if $A = 0$.
Covariate Z such that

$$Y \perp\!\!\!\perp A \mid Z \qquad \textit{(missing at random)}$$

Observe sample of (YA, A, Z) satisfying

- $Y \mid Z \sim \text{binomial}\,\big(1, b(Z)\big)$
- $A \mid Z \sim \text{binomial}\,\big(1, a(Z)\big)$ [propensity score]
- $Z \sim F$

We wish to estimate

$$\mathrm{E}Y = \int b\, dF$$

Missing data

Outcome Y, observed if $A = 1$, unobserved if $A = 0$.
Covariate Z such that

$$Y \perp\!\!\!\perp A \,|\, Z \qquad\qquad \textit{(missing at random)}$$

Observe sample of (YA, A, Z) satisfying

- $Y \,|\, Z \sim$ binomial $\big(1, b(Z)\big)$
- $A \,|\, Z \sim$ binomial $\big(1, a(Z)\big)$ [propensity score]
- $Z \sim F$

We wish to estimate

$$\mathrm{E}Y = \int b \, dF$$

Assume a, b smooth, > 0 and < 1
Sometimes: assume F smooth, density $f > 0$

Connection to causal inference

Treatment indicator A
Outcome Y^0 if not treated
Outcome Y^1 if treated

Causal effect $\mathrm{E}Y^1 - \mathrm{E}Y^0$

Connection to causal inference

Treatment indicator A
Outcome Y^0 if not treated
Outcome Y^1 if treated

Causal effect $\mathrm{E}Y^1 - \mathrm{E}Y^0$

Covariate Z such that

$$Y^0, Y^1 \perp\!\!\!\perp A \mid Z \qquad (\textit{no unmeasured confounders})$$

Connection to causal inference

Treatment indicator A
Outcome Y^0 if not treated
Outcome Y^1 if treated

Causal effect $\mathrm{E}Y^1 - \mathrm{E}Y^0$

Covariate Z such that

$$Y^0, Y^1 \perp\!\!\!\perp A \,|\, Z \qquad (\textit{no unmeasured confounders})$$

$$Y := \begin{cases} Y^0, & \text{if } A = 0 \\ Y^1, & \text{if } A = 1 \end{cases}$$

Observe sample from (Y, A, Z), equivalently $(Y^1 A, Y^0(1 - A), A, Z)$

Connection to causal inference

Treatment indicator A
Outcome Y^0 if not treated
Outcome Y^1 if treated

Causal effect $\quad \mathrm{E}Y^1 - \mathrm{E}Y^0$

Covariate Z such that

$$Y^0, Y^1 \perp\!\!\!\perp A \,|\, Z \qquad (\textit{no unmeasured confounders})$$

$$Y := \begin{cases} Y^0, & \text{if } A = 0 \\ Y^1, & \text{if } A = 1 \end{cases}$$

Observe sample from (Y, A, Z), equivalently $(Y^1 A, Y^0(1-A), A, Z)$

We wish to estimate $\mathrm{E}Y^1$ from $(Y^1 A, A, Z)$
and $\mathrm{E}Y^0$ from $(Y^0(1-A), A, Z)$

Semiparametric regression

Outcome Y, covariates X and Z, error e

$$Y = \theta X + b(Z) + e$$

Observe sample of (Y, X, Z) satisfying

- $e \sim a$ such that $\mathrm{E}(e|\, X, Z) = 0$
- $(X, Z) \sim F$

We wish to estimate θ

Missing data

Outcome Y, observed if $A=1$, unobserved if $A=0$.
Covariate Z such that

$$Y \perp\!\!\!\perp A \,|\, Z \qquad \textit{(missing at random)}$$

Observe sample of (YA, A, Z) satisfying

- $Y|\,Z \sim$ binomial $\big(1, b(Z)\big)$
- $A|\,Z \sim$ binomial $\big(1, a(Z)\big)$ [propensity score]
- $Z \sim F$

We wish to estimate

$$\mathrm{E}Y = \int b\,dF$$

Assume a, b smooth, > 0 and < 1
Sometimes assume F smooth, density $f > 0$

Preliminary estimators

- $Y \perp\!\!\!\perp A \mid Z$
- $Z \sim F$
- $a(Z) = \mathrm{E}(A \mid Z)$
- $b(Z) = \mathrm{E}(Y \mid Z)$

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

- a and b can be estimated by any nonparametric regression estimator
- f can be estimated by any nonparametric density estimator,
  (and so can $f/a \propto f(\cdot \mid A = 1)$)

If a, b, f or f/a are Hölder smooth of orders α, β, γ on $[0,1]^d$, estimation errors will be of order

$$n^{-\delta/(2\delta+d)}, \qquad \delta \in \{\alpha, \beta, \gamma\}.$$

Preliminary estimators

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

- $Y \perp\!\!\!\perp A \mid Z$
- $Z \sim F$
- $a(Z) = \mathrm{E}(A \mid Z)$
- $b(Z) = \mathrm{E}(Y \mid Z)$

- a and b can be estimated by any nonparametric regression estimator
- f can be estimated by any nonparametric density estimator,
  (and so can $f/a \propto f(\cdot \mid A = 1)$)

If a, b, f or f/a are Hölder smooth of orders α, β, γ on $[0,1]^d$, estimation errors will be of order

$$n^{-\delta/(2\delta+d)}, \qquad \delta \in \{\alpha, \beta, \gamma\}.$$

Plug-in estimator $\int \hat{b}\, d\hat{F}$ has rate $n^{-\beta\wedge\gamma/(2\beta\wedge\gamma+d)} \gg n^{-1/2}$

Likelihood

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

- $Y \perp\!\!\!\perp A \mid Z$
- $Z \sim F$
- $a(Z) = \mathrm{E}(A \mid Z)$
- $b(Z) = \mathrm{E}(Y \mid Z)$

Likelihood

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

- $Y \perp\!\!\!\perp A \mid Z$
- $Z \sim F$
- $a(Z) = \mathrm{E}(A \mid Z)$
- $b(Z) = \mathrm{E}(Y \mid Z)$

Likelihood

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

Likelihood

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

- $Y \perp\!\!\!\perp A \mid Z$
- $Z \sim F$
- $a(Z) = \mathrm{E}(A \mid Z)$
- $b(Z) = \mathrm{E}(Y \mid Z)$

Likelihood

$$(a, b, f) \mapsto \Big[f(Z)\Big] \times \Big[a(Z)^A \big(1 - a(Z)\big)^{1-A}\Big] \times \Big[b(Z)^{YA} \big(1 - b(Z)\big)^{(1-Y)A}\Big]$$

Factorizes over f, a, b

Likelihood

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

- $Y \perp\!\!\!\perp A \mid Z$
- $Z \sim F$
- $a(Z) = \mathrm{E}(A \mid Z)$
- $b(Z) = \mathrm{E}(Y \mid Z)$

Likelihood

$$(a, b, f) \mapsto \Big[f(Z)\Big] \times \Big[a(Z)^{A}\big(1 - a(Z)\big)^{1-A}\Big] \times \Big[b(Z)^{YA}\big(1 - b(Z)\big)^{(1-Y)A}\Big]$$

Factorizes over f, a, b

For independent priors, posterior factorizes too

Likelihood

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

- $Y \perp\!\!\!\perp A \mid Z$
- $Z \sim F$
- $a(Z) = \mathrm{E}(A \mid Z)$
- $b(Z) = \mathrm{E}(Y \mid Z)$

Likelihood

$$(a, b, f) \mapsto \Big[f(Z)\Big] \times \Big[a(Z)^{A}\big(1 - a(Z)\big)^{1-A}\Big] \times \Big[b(Z)^{YA}\big(1 - b(Z)\big)^{(1-Y)A}\Big]$$

Factorizes over f, a, b

When using independent priors, posterior factorizes too

Is that good?

Horvitz-Thompson estimator

Assume a known and consider

$$T_a = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i A_i}{a(Z_i)}$$

- $Y \perp\!\!\!\perp A \mid Z$
- $Z \sim F$
- $a(Z) = \mathrm{E}(A \mid Z)$
- $b(Z) = \mathrm{E}(Y \mid Z)$

Thm $\mathrm{E}T_a = \mathrm{E}Y = \theta = \int b \, dF$ and

$$\sqrt{n}(T_a - \theta) \rightsquigarrow N(0, \sigma_a^2)$$

Horvitz-Thompson estimator

Assume a known and consider

$$T_a = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i A_i}{a(Z_i)}$$

- $Y \perp\!\!\!\perp A \mid Z$
- $Z \sim F$
- $a(Z) = \mathrm{E}(A \mid Z)$
- $b(Z) = \mathrm{E}(Y \mid Z)$

Thm $\mathrm{E}T_a = \mathrm{E}Y = \theta = \int b\, dF$ and

$$\sqrt{n}(T_a - \theta) \rightsquigarrow N(0, \sigma_a^2)$$

True for any b or f (unlike nonparametric plug-in $\int \hat{b}\, d\hat{F}$)

Horvitz-Thompson estimator

Assume a known and consider

$$T_a = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i A_i}{a(Z_i)}$$

- $Y \perp\!\!\!\perp A \mid Z$
- $Z \sim F$
- $a(Z) = \mathrm{E}(A \mid Z)$
- $b(Z) = \mathrm{E}(Y \mid Z)$

Thm $\mathrm{E}T_a = \mathrm{E}Y = \theta = \int b\,dF$ and

$$\sqrt{n}(T_a - \theta) \rightsquigarrow N(0, \sigma_a^2)$$

True for any b or f (unlike nonparametric plug-in $\int \hat{b}\,d\hat{F}$)

Crucially utilizes a

Independent priors on a and b would not involve a to estimate θ

Is there a Bayesian analogue?

Horvitz-Thompson estimator (2)

Assume a known, but is estimated by fitting a correct parametric model.

$$T_{\hat{a}} = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i A_i}{\hat{a}(Z_i)}$$

Thm

$$\sqrt{n}(T_a - \theta) \rightsquigarrow N(0, \sigma_a^2),$$
$$\sqrt{n}(T_{\hat{a}} - \theta) \rightsquigarrow N(0, \tau_a^2),$$

for

$$\tau_a^2 < \sigma_a^2$$

Horvitz-Thompson estimator (2)

Assume a known, but is estimated by fitting a correct parametric model.

$$T_{\hat{a}} = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i A_i}{\hat{a}(Z_i)}$$

Thm

$$\sqrt{n}(T_a - \theta) \rightsquigarrow N(0, \sigma_a^2),$$
$$\sqrt{n}(T_{\hat{a}} - \theta) \rightsquigarrow N(0, \tau_a^2),$$

for

$$\tau_a^2 < \sigma_a^2$$

Is there a Bayesian analogue?

Not with independent priors on a and b

Dependent priors

Is it reasonable to model a and b dependent?

$$a(z) = \mathrm{P}(A = 1 | Z = z) \qquad b(z) = \mathrm{E}(Y | Z = z)$$

Dependent priors

Is it reasonable to model a and b dependent?

$$a(z) = \mathrm{P}(A = 1 \mid Z = z) \qquad b(z) = \mathrm{E}(Y \mid Z = z)$$

Chris Sims in blog discussion with Larry Wasserman, Jamie Robins:

IV. Conclusion

Robins and Wasserman have presented not a single case where likelihood based inference, or Bayesian inference in particular, leads one astray. They have presented examples where naive approaches to specifying priors on infinite dimensional spaces can unintentionally imply dogmatic beliefs about parameters of interest. Such examples are interesting and instructive, but they are not cases where a Bayesian approach to inference fails to give good results.

Dependent priors

Is it reasonable to model a and b dependent?

$$a(z) = \mathrm{P}(A = 1 | Z = z) \qquad b(z) = \mathrm{E}(Y | Z = z)$$

Chris Sims in blog discussion with Larry Wasserman, Jamie Robins:

IV. CONCLUSION

Robins and Wasserman have presented not a single case where likelihood based inference, or Bayesian inference in particular, leads one astray. They have presented examples where naive approaches to specifying priors on infinite dimensional spaces can unintentionally imply dogmatic beliefs about parameters of interest. Such examples are interesting and instructive, but they are not cases where a Bayesian approach to inference fails to give good results.

Likelihood

$$\prod_i \phi_i \, \alpha_i^{A_i} (1 - \alpha_i)^{1 - A_i} \, \beta_i^{Y_i A_i} (1 - \beta_i)^{(1 - Y_i) A_i}$$

$(\phi_i = f(Z_i),\ \alpha_i = a(Z_i),\ \beta_i = b(Z_i))$

Double robustness

[many authors, 1980s, 90s; Klaassen, 1987; Robins; vdV 1998]

Estimate a and b nonparametrically (with external data), and set

$$T = \int \hat{b}\, d\hat{F} + \frac{1}{n}\sum_{i=1}^{n} \chi_{\hat{a},\hat{b},\hat{f}}(X_i),$$

for

$$\chi_{a,b,f}(X) = \frac{A}{a(Z)}\big(Y - b(Z)\big) + b(Z) - \int b\, dF$$

Thm If $a \in C^\alpha[0,1]^d$ and $b \in C^\beta[0,1]^d$ for

$$\frac{\alpha}{2\alpha + d} + \frac{\beta}{2\beta + d} > \frac{1}{2},$$

then

$$\sqrt{n}(T - \theta) \rightsquigarrow N(0, \tau^2),$$

for minimal variance

$$\tau^2 < \tau_a^2 < \sigma_a^2$$

Double robustness (2)

Estimate a and b nonparametrically, and set

$$T = \int \hat{b}\, d\hat{F} + \frac{1}{n}\sum_{i=1}^{n} \chi_{\hat{a},\hat{b},\hat{f}}(X_i),$$

for

$$\chi_{a,b,f}(X) = \frac{A}{a(Z)}(Y - b(Z)) + b(Z) - \int b\, dF$$

Thm

$$\sqrt{n}\Big(T - \theta + \int (\hat{a} - a)(\hat{b} - b)\frac{dF}{\hat{a}}\Big) \rightsquigarrow N(0, \tau^2)$$

Double robustness (2)

Estimate a and b nonparametrically, and set

$$T = \int \hat{b}\, d\hat{F} + \frac{1}{n}\sum_{i=1}^{n} \chi_{\hat{a},\hat{b},\hat{f}}(X_i),$$

for

$$\chi_{a,b,f}(X) = \frac{A}{a(Z)}(Y - b(Z)) + b(Z) - \int b\, dF$$

Thm

$$\sqrt{n}\left(T - \theta + \int (\hat{a} - a)(\hat{b} - b)\frac{dF}{\hat{a}}\right) \rightsquigarrow N(0, \tau^2)$$

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

Double robustness (2)

Estimate a and b nonparametrically, and set

$$T = \int \hat{b}\, d\hat{F} + \frac{1}{n}\sum_{i=1}^{n} \chi_{\hat{a},\hat{b},\hat{f}}(X_i),$$

for

$$\chi_{a,b,f}(X) = \frac{A}{a(Z)}(Y - b(Z)) + b(Z) - \int b\, dF$$

Thm

$$\sqrt{n}\Big(T - \theta + \int (\hat{a} - a)(\hat{b} - b)\frac{dF}{\hat{a}}\Big) \rightsquigarrow N(0, \tau^2)$$

Is there a Bayesian analogue?

Double robustness (3) [Robins, Li, Tchetgen, vdV, 2017]

Estimate a, b and f/a nonparametrically, at optimal rates

$$T = \int \hat{b}\, d\hat{F} + \mathbb{U}_n \tilde{\chi}_{\hat{a},\hat{b},\hat{f}},$$

for a higher order influence function $(x_1, \ldots, x_k) \mapsto \tilde{\chi}_{a,b,f}(x_1, \ldots, x_k)$

Thm If $a \in C^\alpha[0,1]^d$, $b \in C^\beta[0,1]^d$, $f/a \in C^\gamma[0,1]^d$,

$$\sqrt{n}(T - \theta) \rightsquigarrow N(0, \tau^2), \qquad \text{if } \alpha + \beta > d/2$$

$$T - \theta = O_P\big(n^{-(2\alpha+2\beta)/(2\alpha+2\beta+d)}\big), \qquad \text{otherwise}$$

This is optimal

(provided γ not too small)

Double robustness (3) [Robins, Li, Tchetgen, vdV, 2017]

Estimate a, b and f/a nonparametrically, at optimal rates

$$T = \int \hat{b}\, d\hat{F} + \mathbb{U}_n \tilde{\chi}_{\hat{a},\hat{b},\hat{f}},$$

for a higher order influence function $(x_1, \ldots, x_k) \mapsto \tilde{\chi}_{a,b,f}(x_1, \ldots, x_k)$

Thm If $a \in C^\alpha[0,1]^d$, $b \in C^\beta[0,1]^d$, $f/a \in C^\gamma[0,1]^d$,

$$\sqrt{n}(T - \theta) \rightsquigarrow N(0, \tau^2), \qquad \text{if } \alpha + \beta > d/2$$

$$T - \theta = O_P(n^{-(2\alpha+2\beta)/(2\alpha+2\beta+d)}), \qquad \text{otherwise}$$

This is optimal

(provided γ not too small)

Is there a Bayesian analogue?

Bias-variance trade-off

In high-dimensional models the quality of estimation
is determined by bias-variance trade-off

Non-principled methods (e.g. estimating equations)
can make explicit bias corrections,
as opposed to Bayesian or other likelihood-based inference

Bias-variance trade-off

In high-dimensional models the quality of estimation
is determined by bias-variance trade-off

Non-principled methods (e.g. estimating equations)
can make explicit bias corrections,
as opposed to Bayesian or other likelihood-based inference

True?
Or can one use clever priors?
Should one?

Results

Bernstein-von Mises

Given priors on a, b, F consider posterior distribution on $\theta = \int b\, dF$.

We say the Bernstein-von Mises theorem holds if, under (a_0, b_0, F_0),

$$d\Big(\mathcal{L}\big(\sqrt{n}(\theta - \hat{\theta}_n)|\, X_1, \ldots, X_n\big), N(0, i_0^{-2})\Big) \xrightarrow{P} 0,$$

for estimators $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ such that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow N(0, i_0^{-2}).$$

The classical result uses the total variation distance d.

Here we use the weak distance (sufficient for justifying credible intervals).

Scores, influence function, and information

[Koshevnik and Levit (1976), Pfanzagl (1983), vdV (1991)]

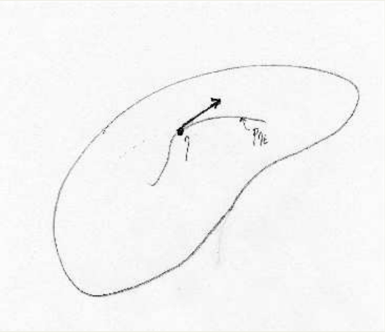
We wish to estimate $\theta = \chi(p)$, from sample from density $p \in \mathcal{P}$

Scores, influence function, and information

[Koshevnik and Levit (1976), Pfanzagl (1983), vdV (1991)]

We wish to estimate $\theta = \chi(p)$, from sample from density $p \in \mathcal{P}$

Tangent cone (at p) to $\mathcal{P}$: all score functions



$$g = \frac{d}{dt}_{|t=0} \log p_t$$

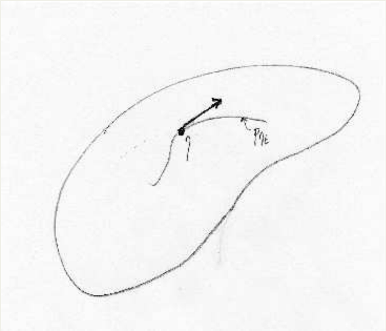
of 1-dimensional submodels $t \mapsto p_t$ with $p_0 = p$

Scores, influence function, and information

[Koshevnik and Levit (1976), Pfanzagl (1983), vdV (1991)]

We wish to estimate $\theta = \chi(p)$, from sample from density $p \in \mathcal{P}$

Tangent cone (at p) to $\mathcal{P}$: all score functions



$$g = \frac{d}{dt}_{|t=0} \log p_t$$

of 1-dimensional submodels $t \mapsto p_t$ with $p_0 = p$

Influence function is $\chi_p \in$ closed span of tangent cone with, $\forall\, t \mapsto p_t$

$$\frac{d}{dt}\chi(p_t)_{|t=0} = \mathrm{E}_p g(X) \chi_p(X)$$

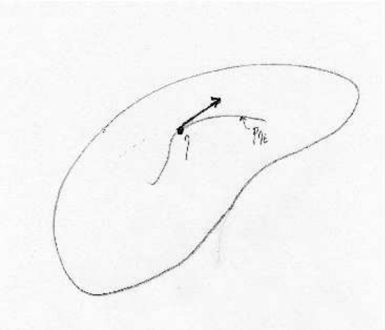
Scores, influence function, and information

[Koshevnik and Levit (1976), Pfanzagl (1983), vdV (1991)]

We wish to estimate $\theta = \chi(p)$, from sample from density $p \in \mathcal{P}$

Tangent cone (at p) to $\mathcal{P}$: all score functions

$$g = \frac{d}{dt}_{|t=0} \log p_t$$



of 1-dimensional submodels $t \mapsto p_t$ with $p_0 = p$

Influence function is $\chi_p \in$ closed span of tangent cone with, $\forall\, t \mapsto p_t$

$$\frac{d}{dt}\chi(p_t)_{|t=0} = \mathrm{E}_p g(X)\chi_p(X)$$

- $\hat\theta_n$ is asymptotically efficient for $\theta = \chi(p)$ if and only if

$$\hat\theta_n = \theta_0 + \frac{1}{n}\sum_{i=1}^n \chi_p(X_i) + o_{P_0}(n^{-1/2}).$$

- Minimal asymptotic variance is

$$i_0^{-2} = \mathrm{E}_p {\chi_p}^2(X).$$

Scores, influence function, and information — Missing data

Observe $X = (YA, A, Z)$
Parameter $p = p_{(a,b,f)}$

- $Y \perp\!\!\!\perp A \mid Z$
- $Z \sim F$
- $a(Z) = \mathrm{E}(A \mid Z)$
- $b(Z) = \mathrm{E}(Y \mid Z)$

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

For given directions (α, β, ϕ), and $\Psi(t) = 1/(1 + e^{-t})$

- a-score: $(A - a(Z))\alpha(Z)$ $\qquad a_t = \Psi\big(\Psi^{-1}(a) + t\alpha\big)$
- b-score: $\dfrac{A(Y - b(Z))}{b(Z)(1 - b)(Z)}\beta(Z)$ $\qquad b_t = \Psi\big(\Psi^{-1}(b) + t\beta\big)$
- f-score: $\phi(Z) - \int \phi\, dF$ $\qquad f_t \propto e^{\log f + t\phi}$

$\theta = \int b\, dF$ has Influence function

$$\chi_{a,b,F}(X) = \frac{A}{a(Z)}\big(Y - b(Z)\big) + b(Z) - \int b\, dF$$

Scores, influence function, and information — Missing data

Least favourable direction (at (a, b, f))

$$\xi := (\alpha^{lf}, \beta^{lf}, \phi^{lf}) = \Big(0, \frac{1}{a}, b - \int b \, dF\Big)$$

gives submodels (a_t, b_t, f_t) with score the efficient influence function:

$$\frac{\partial}{\partial t}_{|t=0} \log p_{a_t, b_t, f_t}(X) = \chi_{a,b,F}(X).$$

Priors

(1) Put prior on $\eta := (\eta^a, \eta^b, \eta^f)$ and set

$$a = \Psi(\eta^a), \qquad b = \Psi(\eta^b), \qquad f \propto e^{\eta^f}$$

(2) Put prior on $\eta := (\eta^a, \eta^b) \perp\!\!\!\perp F \sim \mathrm{DP}$ and set

$$a = \Psi(\eta^a), \qquad b = \Psi(\eta^b)$$

(3) Put prior on $w \perp\!\!\!\perp F \sim \mathrm{DP} \perp\!\!\!\perp \lambda \sim N(0, \sigma_n^2)$, estimate a, and set

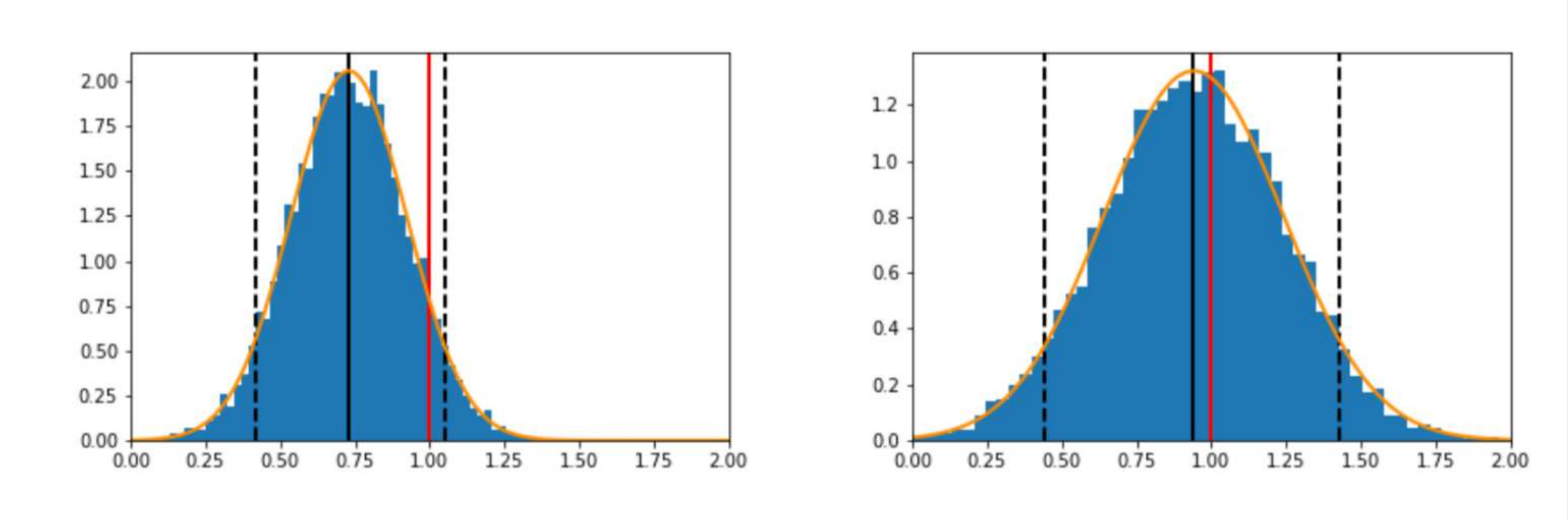
$$b(z) = \Psi\left(w(z) + \frac{\lambda}{\hat{a}(z)}\right)$$

(4) Put prior on $w \perp\!\!\!\perp a \perp\!\!\!\perp F \sim \mathrm{DP} \perp\!\!\!\perp \lambda \sim N(0, \sigma_n^2)$, and set

$$b(z) = \Psi\left(w(z) + \frac{\lambda}{a(z)}\right)$$

Numerical results [Ray and Szabó, 2019]

Simulation results for causal effect in Gaussian response model



independent priors bias corrected with dependent priors

posterior of θ; black line: posterior mean, red line: true value

Numerical results [Ray and Szabó, 2019]

Simulation results for causal effect in Gaussian response model

Method	Abs. error ± sd	Size CI ± sd	Coverage	Type II error
GP	0.317 ± 0.031	0.516 ± 0.021	0.06	**0.00**
GP (noRand)	0.317 ± 0.031	0.359 ± 0.014	0.00	**0.00**
GP PS	**0.063 ± 0.040**	0.742 ± 0.030	**1.00**	**0.00**
GP PS (noRand)	**0.063 ± 0.040**	**0.640 ± 0.028**	**1.00**	**0.00**
BART	0.225 ± 0.168	1.705 ± 0.474	0.99	0.47
BART (PS)	0.139 ± 0.096	0.747 ± 0.080	0.98	**0.00**
CF (AIPW)	0.138 ± 0.098	0.697 ± 0.103	0.96	**0.00**
CF (TMLE)	0.136 ± 0.100	0.891 ± 0.152	0.99	0.01
OLS	0.715 ± 0.166	0.363 ± 0.035	0.00	0.24
CB (IPW)	0.607 ± 0.332	1.504 ± 0.425	0.71	0.01
PSM	0.218 ± 0.173	1.281 ± 0.165	0.97	0.06

Numerical results [Ray and Szabó, 2019]

Simulation results for causal effect in Gaussian response model

Method	Abs. error $\pm$ sd	Size CI $\pm$ sd	Coverage	Type II error
GP	0.246 ± 0.398	1.250 ± 1.014	0.92	0.01
GP (noRand)	0.246 ± 0.398	1.014 ± 0.877	0.88	0.01
GP + PS	0.189 ± 0.234	1.286 ± 1.107	0.99	0.01
GP +PS (noRand)	0.189 ± 0.234	1.040 ± 0.897	0.94	0.01
BART	0.256 ± 0.332	0.925 ± 0.707	0.87	**0.00**
BART PS	0.249 ± 0.353	0.876 ± 0.644	0.87	**0.00**
CF (AIPW)	0.243 ± 0.265	1.030 ± 0.778	0.90	0.01
CF (TMLE)	0.244 ± 0.274	1.057 ± 0.782	0.90	0.01
OLS	0.134 ± 0.112	$\mathbf{0.801 \pm 0.526}$	0.97	**0.00**
CB (IPW)	0.238 ± 0.199	1.200 ± 0.850	0.90	**0.00**
CB (CM)	0.138 ± 0.120	0.967 ± 0.786	0.94	**0.00**
PSM	$\mathbf{0.133 \pm 0.105}$	2.002 ± 1.681	**1.00**	0.01

BvM (1) — general prior

Prior on $\eta := (\eta^a, \eta^b, \eta^f)$

Thm Assume $\exists \mathcal{H}_n$ with $\Pi(\eta \in \mathcal{H}_n \mid X_1 \ldots X_n) \to^{P_0} 1$,

$$\sup_{b,f:\eta\in\mathcal{H}_n} \|b - b_0\|_{L^2(F_0)} + \|f - f_0\|_1 \to 0$$

$$\sup_{b:\eta\in\mathcal{H}_n} \left|\mathbb{G}_n[b - b_0]\right| \to^{P_0} 0$$

$$\sup_{b,f:\eta\in\mathcal{H}_n} \left|\sqrt{n}\int (b - b_0)(f - f_0)\, dz\right| \to 0$$

Then BvM theorem holds if $\forall t$

$$\frac{\int_{\mathcal{H}_n} \prod_{i=1}^n p_{\eta - t\xi_0/\sqrt{n}}(X_i)\, d\Pi(\eta)}{\int_{\mathcal{H}_n} \prod_{i=1}^n p_\eta(X_i)\, d\Pi(\eta)} \to^{P_0} 1$$

Modelled after Castillo (2012) and Castillo & Rousseau (2015)

BvM (2) — with Dirichlet prior

Prior on η: $= (\eta^a, \eta^b) \perp\!\!\!\perp F \sim \mathrm{DP}$

Thm Assume $\exists \mathcal{H}_n^{a,b}$ with $\Pi(\eta \in \mathcal{H}_n^{a,b} | X_1, \ldots, X_n) \to^{P_0} 1$,

$$\sup_{b:\eta \in \mathcal{H}_n^{a,b}} \|b - b_0\|_{L^2(F_0)} \to 0$$

$$\sup_{b:\eta \in \mathcal{H}_n^{a,b}} |\mathbb{G}_n[b - b_0]| \to^{P_0} 0$$

Then BvM theorem holds if $\forall t$

$$\frac{\int_{\mathcal{H}_n^{a,b}} \prod_{i=1}^n p_{\eta - t\tilde{\xi}_0/\sqrt{n}}(X_i)\, d\Pi(\eta)}{\int_{\mathcal{H}_n^{a,b}} \prod_{i=1}^n p_\eta(X_i)\, d\Pi(\eta)} \to^{P_0} 1$$

BvM (3) — with propensity score-dependent prior and Dirichlet prior

Prior on $w \perp\!\!\!\perp F \sim \mathrm{DP} \perp\!\!\!\perp \lambda \sim N(0, \sigma_n^2)$, estimate $\hat{a}$ with rate ρ_n

$$b(z) = \Psi\Big(w(z) + \frac{\lambda}{\hat{a}(z)}\Big)$$

Thm Assume $\exists \mathcal{H}_n^b$ and numbers $u_n, \varepsilon_n^b \to 0$,

$$\Pi\big((w,\lambda)\colon w + (\lambda + tn^{-1/2})/\hat{a} \in \mathcal{H}_n^b \,|\, X_1, \ldots, X_n\big) \to^{P_0} 1$$

$$\Pi\big(\lambda\colon |\lambda| \le u_n \sigma_n^2 \sqrt{n} \,|\, X_1, \ldots, X_n\big) \to^{P_0} 1$$

$$\sup_{b: \eta^b \in \mathcal{H}_n^b} \|b - b_0\|_{L^2(F_0)} \le \varepsilon_n^b$$

$$\sup_{b: \eta^b \in \mathcal{H}_n^b} \big|\mathbb{G}_n[b - b_0]\big| \to^{P_0} 0$$

Then BvM theorem holds if $n\sigma_n^2 \to \infty$ and $\sqrt{n}\rho_n\varepsilon_n^b \to 0$

Random λ makes prior "vague" in least favorable direction

*** Exponential BVM for Dirichlet

$$F_n \mid Z_1, \ldots, Z_n \sim \mathrm{DP}(\nu + n\mathbb{F}_n), \qquad \mathbb{F}_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{Z_i}$$

Thm If $\sup_{g\in\mathcal{G}_n} |\mathbb{F}_n g - F_0 g| \to^P 0$, $\nu G_n = O(1)$, $F_0 G_n^{2+\delta} = O(1)$, then, for small $|t|$,

$$\sup_{g\in\mathcal{G}_n} \left| \mathrm{E}\left[e^{t\sqrt{n}(F_n g - \mathbb{F}_n g)} \mid Z_1, \ldots, Z_n\right] - e^{t^2 F_0(g - F_0 g)^2/2} \right| \xrightarrow{P} 0$$

Independent Gaussian process priors

$W^b \perp\!\!\!\perp W^f$ Riemann-Liouville ($\bar\beta$) and ($\bar\gamma$)

$$b(z) = \Psi(W^b_z), \qquad f(z) \propto e^{W^f_z}$$

Cor If $a_0 \in C^\alpha$, $b_0 \in C^\beta$, $f_0 \in C^\gamma$ with

$$\alpha, \beta > \frac{d}{2}, \quad \frac{d}{2} < \bar\beta < \alpha + \beta - \frac{d}{2}, \quad \bar\gamma < \gamma + \beta - \frac{d}{2}$$

and

$$\frac{\beta \wedge \bar\beta}{2\bar\beta + d} + \frac{\gamma \wedge \bar\gamma}{2\bar\gamma + d} > \frac{1}{2},$$

then BvM theorem holds

Example: if $\beta = \bar\beta$, $\gamma = \bar\gamma$, then need $\alpha, \beta > d/2$ and γ large enough

No double robustness, f important

Independent Gaussian process prior with Dirichlet

$W^b \sim$ Riemann-Liouville$(\bar\beta) \perp\!\!\!\perp F \sim \mathrm{DP}$

$$b(z) = \Psi(W_z^b)$$

Cor If $a_0 \in C^\alpha$, $b_0 \in C^\beta$, and

$$\alpha, \beta > \frac{d}{2}, \qquad \frac{d}{2} < \bar\beta < \alpha + \beta - \frac{d}{2},$$

then BvM theorem holds

Example: if $\beta = \bar\beta$, then need $\alpha, \beta > d/2$

No condition on f, no double robustness, undersmooth b

Gaussian process prior with propensity score and Dirichlet

$W^b \perp\!\!\!\perp \lambda \sim N(0, \sigma_n^2)$, estimate a with rate ρ_n

$$b(z) = \Psi\Big(W_z^b + \frac{\lambda}{\hat{a}(z)}\Big)$$

Cor If $a_0 \in C^\alpha$ and $b_0 \in C^\beta$, and

$$\Big(\frac{n}{\log n}\Big)^{-\frac{\beta\wedge\bar\beta}{2\bar\beta+d}} \ll \sigma_n \lesssim 1, \quad \beta \wedge \bar\beta > \frac{d}{2}, \quad \sqrt{n}\rho_n \Big(\frac{n}{\log n}\Big)^{-\frac{\beta\wedge\bar\beta}{2\bar\beta+d}} \to 0,$$

then BvM theorem holds

Example: if $\beta = \bar\beta$ and $\hat{a}$ optimal, then need $\beta > d/2$ and

$$\frac{\alpha}{2\alpha + d} + \frac{\beta}{2\beta + d} > \frac{1}{2}$$

No condition on f, half of double robustness:
smoothness of b can compensate low smoothness of a

Regularity conditions?

- $b \in C^\beta$ for $\beta > d/2$ needed for "regularity condition"

$$\sup_{b:\eta^b \in \mathcal{H}_n^b} \left|\mathbb{G}_n[b - b_0]\right| \to^{P_0} 0$$

Regularity conditions?

- $b \in C^\beta$ for $\beta > d/2$ needed for "regularity condition"

$$\sup_{b:\eta^b \in \mathcal{H}_n^b} \left| \mathbb{G}_n[b - b_0] \right| \to^{P_0} 0$$

- $a \in C^\alpha$ and $b \in C^\beta$ for $\alpha, \beta > d/2$ needed for prior:

$$b(z) = \Psi\left(W_z^b + \frac{\lambda}{W_z^a}\right), \qquad W^b \perp\!\!\!\perp W^a \perp\!\!\!\perp \lambda \sim N(0, \sigma_n^2).$$

Regularity conditions?

- $b \in C^\beta$ for $\beta > d/2$ needed for "regularity condition"

$$\sup_{b:\eta^b \in \mathcal{H}_n^b} \left|\mathbb{G}_n[b - b_0]\right| \to^{P_0} 0$$

- $a \in C^\alpha$ and $b \in C^\beta$ for $\alpha, \beta > d/2$ needed for prior:

$$b(z) = \Psi\Big(W_z^b + \frac{\lambda}{W_z^a}\Big), \qquad W^b \perp\!\!\!\perp W^a \perp\!\!\!\perp \lambda \sim N(0, \sigma_n^2).$$

This prevents double robustness
Is this an artefact of the proof?

Semiparametric Bayesian estimation, with or without bias

Kolyan Ray and Aad van der Vaart

King’s College London and Universiteit Leiden

BNP12, Oxford, June 2019