Nonparametric Bayes
and
Causal Sensitivity Analysis

A.W. van der Vaart

TU Delft

Lund Statistics Day 2024

joint with


Bart Eggen, TU Delft


Stephanie van der Pas, AMC


Chris van Vliet, TU Delft

Causal Inference

$A \in \{0,1\}$ treatment indicator

Y^1, Y^0 outcomes if $A=1,0$. "counterfactual"

Causal effect $EY^1 - EY^0$

Causal Inference

$A \in \{0,1\}$ treatment indicator

y^1, y^0 outcomes if $A=1,0$. "counterfactual"

Causal effect $Ey^1 - Ey^0$ only estimable under conditions

Z covariates

Observed (A, y, Z), $y = y^A$.

LEM if $y^1, y^0 \perp\!\!\!\perp A \mid Z$, then

$$Ey^1 - Ey^0 = E_Z\left(E(y \mid A=1, Z) - E(y \mid A=0, Z)\right)$$

LEM if $y^1, y^0 \perp\!\!\!\perp A \mid Z$, then

$$E y^1 - E y^0 = E_Z \left(E(y \mid A=1, Z) - E(y \mid A=0, Z) \right)$$

proof

$$\begin{aligned} E\, y^a &= E_Z\, E(y^a \mid Z) \\ &= E_Z\, E(y^a \mid A=a, Z) && \text{(assumption)} \\ &= E_Z\, E(y \mid A=a, Z) && \text{(consistency)} \end{aligned}$$

(need also that conditioning on $\{A=a, Z\}$ makes sense:

positivity: $P(A=a \mid Z) > 0$)

$\square$

Sensitivity Analysis

$Y^1, Y^0 \perp\!\!\!\perp A \mid Z$

"Conditional Exchangeability"
or
"No unmeasured confounding"

Z has to be rich enough to make this true
(and not too rich)

Sensitivity Analysis: consider deviations

Sensitivity Analysis

$y^1, y^0 \perp\!\!\!\perp A \mid Z$

"Conditional Exchangeability"
or
"No unmeasured confounding"

Z has to be rich enough to make this true
(and not too rich)

Sensitivity Analysis: consider deviations

Two approaches

- Assume $\exists\, U$ with $y^1, y^0 \perp\!\!\!\perp A \mid Z, U$
- Model $A \mid Z, y^a$, $a \in \{0,1\}$

Bayesian Sensitivity Analysis

Both approaches add an <u>unidenfiable</u>

"sensitivity parameter"

> Causal effect is identifiable only given the sensitivity parameter.
>
> Bayesian approach puts a prior on this parameter and obtains a posterior of the causal effect

As usual, except

- posterior does not contract to point and
- priors matter

Two approaches to model failure of $Y^1, Y^0 \perp\!\!\!\perp A \mid Z$

- Assume $\exists$ U with $Y^1, Y^0 \perp\!\!\!\perp A \mid Z, U$
- Model $A \mid Z, Y^a$, $a \in \{0,1\}$

McCandless, Gustafson, Levy, 2007
McCandless, Gustafson, Levy, Richardson, 2010, 2011

$$\text{logit}\ P(Y=1 \mid A, Z, U) = \beta_0 + \beta_1 A + \alpha U + \eta^T Z$$

$$\text{logit}\ P(U=1 \mid A, Z) = \gamma_0 + \gamma_1 A + \xi^T Z$$

A	beta blockers
Y	1-year mortality
U	unmeasured
Z	covariates

$$Y^1, Y^0 \perp\!\!\!\perp A \mid Z, U \implies \text{logit}\ P(Y^a=1 \mid A=a, Z, U) = \beta_0 + \beta_1 a + \alpha U + \eta^T Z$$

"causal effect" ↔ β_1

get posterior of β_1 by Gibbs sampling with U missing:

repeat $\beta_0, \beta_1, \alpha, \eta, \gamma_0, \gamma_1 \mid \text{DATA}, U$

$U \mid \text{DATA}, \beta_0, \beta_1, \alpha, \eta, \gamma_0, \gamma_1$

Dorie, Harada, Bohme Carnegie, Hill, 2016

$$y \mid A, Z, U \sim N(f(A,Z) + \alpha_1 U, \sigma^2)$$

$$P(A=1 \mid Z, U) = \Phi(Z\beta + \alpha_2 U)$$

$$U \perp\!\!\!\perp Z, \quad U \sim \text{Bernoulli}(\pi)$$

prior on f : BART

$$y^1, y^0 \perp\!\!\!\perp A \mid Z, U \Rightarrow E y^a = E\,E(y \mid A=a, Z, U) = E f(a,Z) + \alpha_1 E U$$

$$E y^1 - E y^0 = E f(1,Z) - E f(0,Z).$$

Get posterior of $Ey^1 - Ey^0$ from posterior of f by Gibbs sampling with U missing

Nonparametric Bayes (1): BART

Chipman, George, McCulloch, 1998.

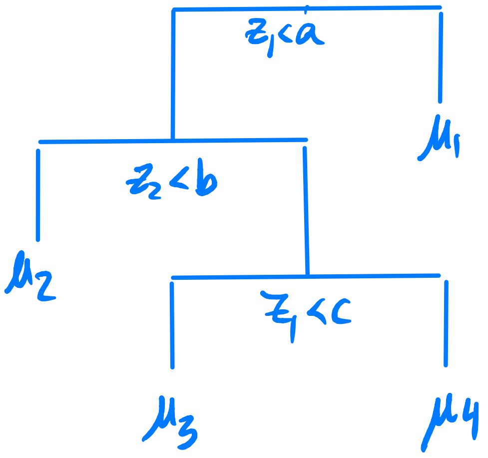
Bayesian Additive Regression Tree

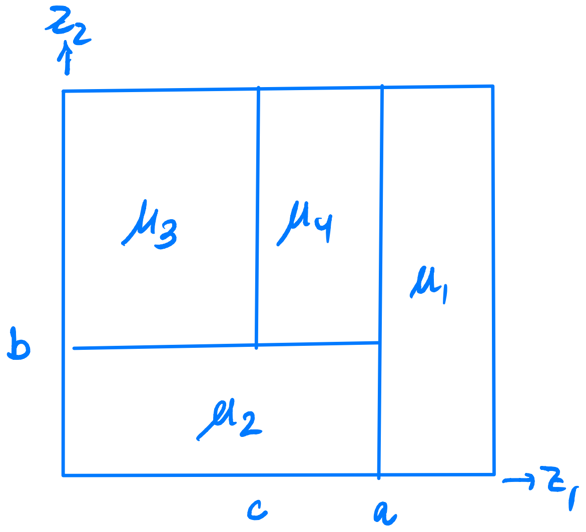
$$f(z) = g(z; T_1, M) + g(z; T_2, M) + \cdots + g(z; T_m, M_m)$$

$z \mapsto g(z; T, M)$

T = tree

$M = (a, b, c, \cdots, \mu_1, \mu_2 \cdots)$





PRIOR

- Node at level h is split with probability $0.95/(1+h)^2$
- Splitting variable z_i and value $a, b, c, \cdots$ chosen uniformly.
- Leaf values $\mu_1, \mu_2, \cdots \overset{\text{iid}}{\sim} N(0, \sigma^2/m)$.

+ empirical Bayes

Two approaches

to model failure of $Y^1, Y^0 \perp\!\!\!\perp A \mid Z$

- Assume $\exists$ U with $Y^1, Y^0 \perp\!\!\!\perp A \mid Z, U$
- Model $A \mid Z, Y^a$, $a \in \{0,1\}$

Robins, Rotnitzky, Scharfstein, 2000
Scharfstein, Daniels, Robins, 2003

FOCUS ON y^1
y^0 ANALOGOUS

$$\text{logit}\, P(A=1 \mid y^1, Z) = \eta(Z) + q(y^1 \mid Z)$$

sensitivity model

Robins, Rotnitzky, Scharfstein, 2000
Scharfstein, Daniels, Robins, 2003

FOCUS ON y^1
y^0 ANALOGOUS

$$\operatorname{logit} P(A=1 \mid y^1, Z) = \eta(Z) + q(y^1 \mid Z)$$

sensitivity model

$$P(y^1 \in \cdot \mid Z) = P(\cdot \mid Z)$$

outcome

Robins, Rotnitzky, Scharfstein, 2000
Scharfstein, Daniels, Robins, 2003

FOCUS ON y^1
y^0 ANALOGOUS

$\text{logit}\, P(A=1 \mid y^1, z) = \eta(z) + g(y^1 \mid z)$ sensitivity model

$P(y^1 \in \cdot \mid Z) = P(\cdot \mid Z)$ outcome

- -

$P(A=1 \mid Z) = \pi(Z)$ propensity score

$P(y^1 \in \cdot \mid A=1, Z) = P_1(\cdot \mid Z)$ observed outcome

- -

Robins, Rotnitzky, Scharfstein, 2000
Scharfstein, Daniels, Robins, 2003

FOCUS ON y'
y^0 ANALOGOUS

$\operatorname{logit} P(A=1 \mid y', Z) = \eta(Z) + q(y' \mid Z)$ sensitivity model

$P(y' \in \cdot \mid Z) = P(\cdot \mid Z)$ outcome

- -

$P(A=1 \mid Z) = \pi(Z)$ propensity score

$P(y' \in \cdot \mid A=1, Z) = P_1(\cdot \mid Z)$ observed outcome

- -

$P(y' \in \cdot \mid A=0, Z) = P_0(\cdot \mid Z)$ <u>un</u>observed outcome

LEM

- $\forall q \; \exists$ bijections $(\eta, P) \leftrightarrow (\pi, P_1) \leftrightarrow \mathcal{L}(A, Ay' \mid Z)$
- $dP_0(y \mid z) \propto e^{-q(y \mid z)} \, dP_1(y \mid z)$

Prior Modelling

Full Data (A, y', Z) Observed Data (A, Ay', Z)

Sensitivity Model $\text{logit}\, P(A=1 \mid Z, y') = \eta(Z) + q(y' \mid Z)$

$\pi(Z) = P(A=1 \mid Z)$ $P_1(\cdot \mid Z) = \mathcal{L}(y' \mid A=1, Z)$ $P(\cdot \mid Z) = \mathcal{L}(y' \mid Z)$

Strategy 1

prior on P^Z, π, P_1, q, independent

$$E y' = E_Z E(y' \mid Z) = E_Z \int y \left[(1-\pi(Z))\, dP_0(y \mid Z) + \pi(Z)\, dP_1(y \mid Z) \right]$$

$$\uparrow \propto e^{q(y \mid Z)} dP_1(y \mid Z)$$

Strategy 2

Prior on P^Z, η, P, q, independent

$$E y' = E_Z E(y' \mid Z) = E_Z \int y \, dP(y \mid z)$$

Findings

Both strategies: posterior of $Ey' \not\to \delta_\theta$,
even if $|DATA| \to \infty$

Strategy 1 : posterior of q = prior of q

Strategy 2 : posterior of q depends on DATA

Nonparametric Bayes (2): Dirichlet process

P ~ DP(a) $\iff (P(A_1), \dots, P(A_k)) \sim \text{Dirichlet}(a(A_1), \dots, a(A_k)) \quad \forall\, \mathcal{X} = \sqcup_i A_i$

$\iff P \sim \sum_{i=1}^{\infty} W_i \delta_{\theta_i}, \quad \theta_i \overset{iid}{\sim} \frac{a}{|a|} \perp\!\!\!\perp W_i = V_i \prod_{j<i}(1-V_j), \quad V_i \overset{iid}{\sim} Be(1, |a|)$

PROP If $P \sim DP(a)$, $X_1 \cdots X_n \mid P \overset{iid}{\sim} P$, then $P \mid X_1 \cdots X_n \sim DP(a + n\mathbb{P}_n)$

Ferguson, 1973
Lo, 1983

$\rightarrow$ • $E(P \mid X_1 \cdots X_n) = \mathbb{P}_n \frac{n}{n+|a|} + a \frac{|a|}{n+|a|} \approx \mathbb{P}_n$.

• $\sqrt{n}(P - \mathbb{P}_n) \mid X_1 \cdots X_n \rightsquigarrow G_{P_0}$

↑ empirical measure

Using DP(α) prior is Bayesian equivalent of estimating a distribution by $\mathbb{P}_n$ (= MLE)

Special Case: No Covariates , given q

Full Data (A, Y') Observed Data (A, AY')

$\pi = P(A=1)$, $P_1 = \mathcal{L}(Y' \mid A=1)$, $P = \mathcal{L}(Y)$, $\text{logit}\, P(A=1 \mid Y) = \eta + q(Y)$

Strategy 1 $\pi \sim \text{Beta}(\alpha, \beta) \perp\!\!\!\perp P_1 \sim DP(a) \perp\!\!\!\perp q$

Strategy 2 $\eta \sim \mathcal{U}[\alpha, \beta] \perp\!\!\!\perp P \sim DP(a) \perp\!\!\!\perp q$

THM For both strategies, Bernstein-von Mises:

$$\sqrt{n}\,(\underset{\overset{\uparrow}{EY'}}{\theta} - \hat{\theta}_{n,q}) \mid \text{DATA}_n, q \rightsquigarrow N(0, \sigma_q^2)$$

$\uparrow$

efficient ("double robust) estimator of EY'

Proofs

Strategy 1

$\pi \sim \mathrm{Beta}(\alpha, \beta) \perp\!\!\!\perp P_1 \sim DP(a) \perp\!\!\!\perp q$

P_1 is law of data → conjugacy of DP

Strategy 2

$\eta \sim \mathcal{U}[\alpha, \beta] \perp\!\!\!\perp P \sim DP(a) \perp\!\!\!\perp q$

$$P \sim DP(a) \implies dP_1(y) \propto \frac{1}{1+e^{\eta+q(y)}} \, dP(y)$$

$$\implies P_1 \sim \text{"NCRM"}$$

Nonparametric Bayes (3): NCRM

Kingman, 1975

Ψ Completely Random Measure if $\Psi(A_1) \perp\!\!\!\perp \cdots \perp\!\!\!\perp \Psi(A_k)$ $\forall\ \mathcal{X} = \coprod_{i=1}^{n} A_i$

- intensity measure $\nu = \nu^d + \nu^c$
- $\Psi = \sum_{i=1}^{\infty} V_i \delta_{a_i} + \sum_{i=1}^{\infty} W_i \delta_{\theta_i}$, $(a_i) \subset \mathcal{X}$, $(V_i) \sim \nu^d$, $(W_i, \theta_i) \sim \text{Poisson}(\nu^c)$

$P = \dfrac{\Psi}{\Psi(\mathcal{X})}$ Normalised C R M

THM If $\nu^c(dy, ds) = s^{-1} e^{-s b(y)} ds\, d\alpha(y)$, then for Donsker class $\mathcal{F}$

$$\sqrt{n}\,(P - \underset{\uparrow}{\mathbb{P}_n}) \mid Y_1 \cdots Y_n \rightsquigarrow \underset{\uparrow}{\mathbb{B}_{P_0}} \quad \text{in } \ell^{\infty}(\mathcal{F})$$

empirical measure

Brownian bridge

b bounded below

$P_0 b^q + P_0 F^r + \alpha F < \infty$

$\frac{1}{q} + \frac{1}{r} < \frac{1}{2}$

Nonparametric Bayes (3): NCRM

Kingman, 1975

Ψ Completely Random Measure if $\Psi(A_1) \perp\!\!\!\perp \cdots \perp\!\!\!\perp \Psi(A_k)$ $\forall\, \mathfrak{X} = \bigsqcup_{i=1}^{n} A_i$

- intensity measure $\nu = \nu^d + \nu^c$
- $\Psi = \sum_{i=1}^{\infty} V_i \delta_{a_i} + \sum_{i=1}^{\infty} W_i \delta_{\theta_i}$, $(a_i) \subset \mathfrak{X}$, $(V_i) \sim \nu^d$, $(W_i, \theta_i) \sim \text{Poisson}(\nu^c)$

$P = \dfrac{\Psi}{\Psi(\mathfrak{X})}$ Normalised C R M

THM If $\nu^c(dy, ds) = s^{-1-\sigma} e^{-s b(y)} \, ds \, d\alpha(y)$, then for Donsker class $\mathcal{F}$

$$\sqrt{n}\left(P - \mathbb{P}_n - \sigma \frac{K_n}{n}\left(G - \tilde{\mathbb{P}}_n\right)\right) \mid Y_1 \cdots Y_n \rightsquigarrow \mathbb{B}_{P_0} \quad \text{in } \ell^\infty(\mathcal{F})$$

empirical measure

#distinct values

empirical measure distinct values

Gaussian process

b bounded below

$P_0 b^q + P_0 F^r + \alpha F < \infty$

$\frac{1}{q} + \frac{1}{r} < \frac{1}{2}$

proof

PROP If $P \sim NCRM$, $y^1 \cdots y^n | P \overset{iid}{\sim} P$, then $P | y_1 \cdots y_n \sim$ mixture of NCRMs

James, Lijoi, Prunster, 2009

$P | y_1, \cdots, y_n, \lambda$ has $\nu = \nu^d_{n,\lambda} + \nu^c_{n,\lambda}$

$$\nu^d_{n,\lambda}(\tilde{y}_j, s) \propto s^{N_{jn}-1} e^{-\lambda(\lambda+b/\tilde{y}_j))} ds$$

↑ distinct values

multiplicity $\tilde{y}_j$

- Show $\Pi_n(\lambda \gtrsim \frac{n}{\log n} | y_1 \cdots y_n) \xrightarrow{P} 1$
- Show discrete part dominates

(if $\sigma > 0$ latter is not true)

Special Case: No Covariates

, $q \sim$ prior

Full Data (A, y') Observed Data (A, Ay')

$\pi = P(A=1)$, $P_1 = \mathcal{L}(y' \mid A=1)$, $P = \mathcal{L}(y)$, $\operatorname{logit} P(A=1 \mid y) = \eta + q(y)$

Strategy 1 $\pi \sim \text{Beta}(\alpha, \beta) \perp\!\!\!\perp P_1 \sim DP(a) \perp\!\!\!\perp q$

Strategy 2 $\eta \sim \mathcal{U}[\alpha, \beta] \perp\!\!\!\perp P \sim DP(a) \perp\!\!\!\perp q$

THM Strategy 1 : $q \mid \text{DATA} \sim \underline{q}$

$\mathcal{L}(A, Ay')$

Strategy 2 : $q \mid \text{DATA} \sim q \mid H = \mathbb{H}_n$

This works through in posterior of Ey'.

Special Case: No Covariates

computation

Full Data (A, Y') Observed Data (A, AY')

$\pi = P(A=1)$, $P_1 = \mathcal{L}(Y' | A=1)$, $P = \mathcal{L}(Y)$, $\text{logit}\, P(A=1 | Y) = \eta + q(Y)$

Strategy 1

$\pi \sim \text{Beta}(\alpha, \beta) \perp\!\!\!\perp P_1 \sim DP(a) \perp\!\!\!\perp q$

Explicit Dirichlet posterior

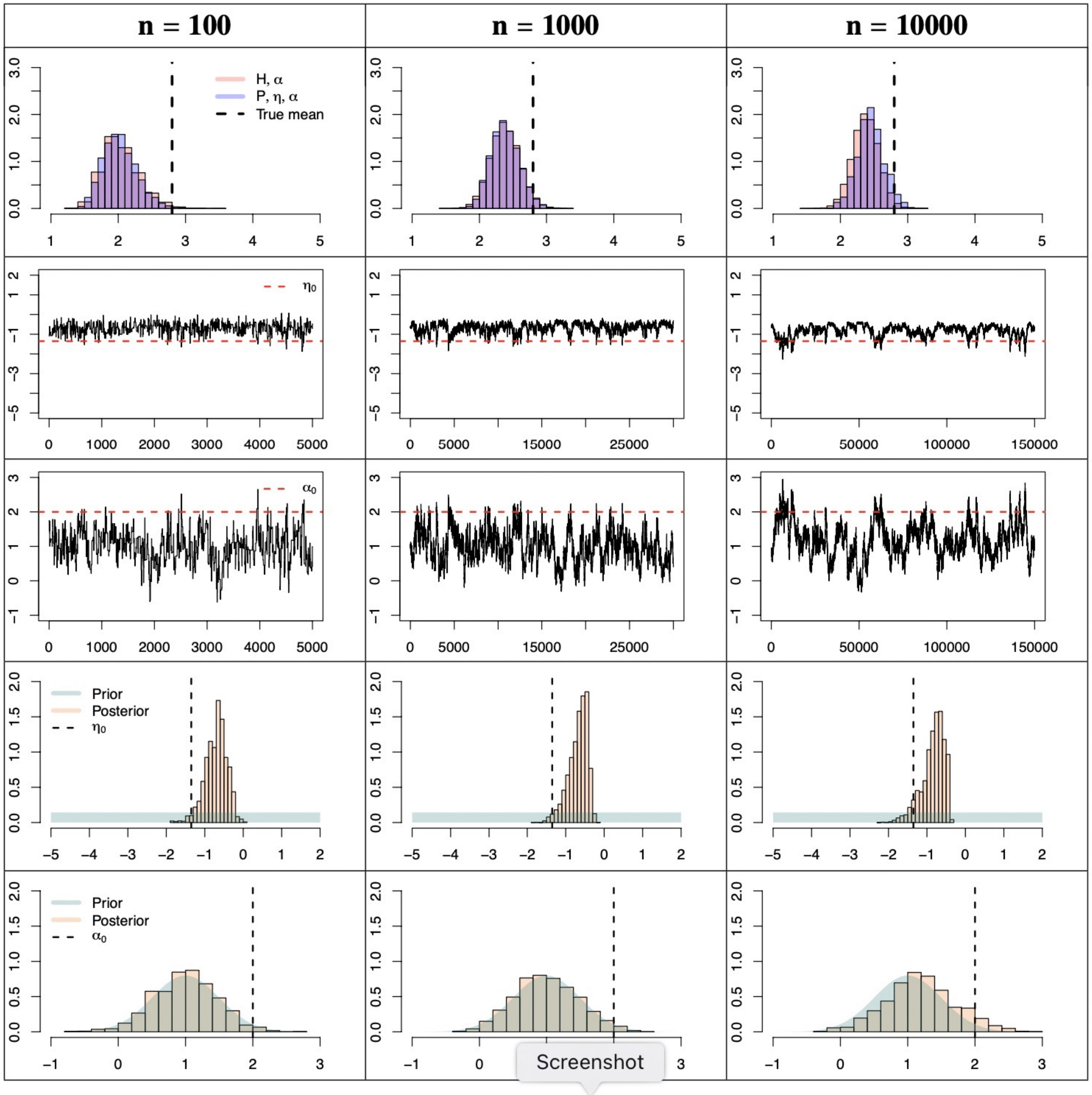
Strategy 2

$\eta \sim \mathcal{U}[\alpha, \beta] \perp\!\!\!\perp P \sim DP(a) \perp\!\!\!\perp q$

Posterior by Gibbs sampling missing outcomes

Setting where sensitivity model is <u>not</u> centered at truth

Sensitivity function $g(y) = \alpha \log y \quad , y \in (0, \infty)$



Ey'

η

α

η

α

Special Case: Survival Outcome

Full Data (A, Y', Z) Observed Data (A, AY', Z)

$Y' \in (0, \infty)$

$P(\cdot|Z) = \mathcal{L}(Y'|Z)$ $\text{logit}\, P(A=1|Z, Y') = \eta(Z) + q(Y'|Z)$

Strategy 2

$P^Z \sim DP(a) \perp\!\!\!\perp \eta(Z) = \gamma^T Z, \gamma \sim \pi \perp\!\!\!\perp P(\cdot|Z) \sim$ Bayesian Cox

Nonparametric Bayes (4): Cox Model

Cox model $\Lambda(t|z) = e^{\beta^T z}\Lambda(t)$

$$\iff P(T > t \,|\, z, \beta, \Lambda) = e^{-e^{\beta^T z}\Lambda(t)}$$

Prior model $\beta \sim \Pi$, $\Lambda \sim$ Beta process

Hjort, 1990

Data model $T_1 \cdots T_n \,|\, \beta, \Lambda, Z_1 \cdots Z_n \overset{ind}{\sim} Cox(\beta, \Lambda)$

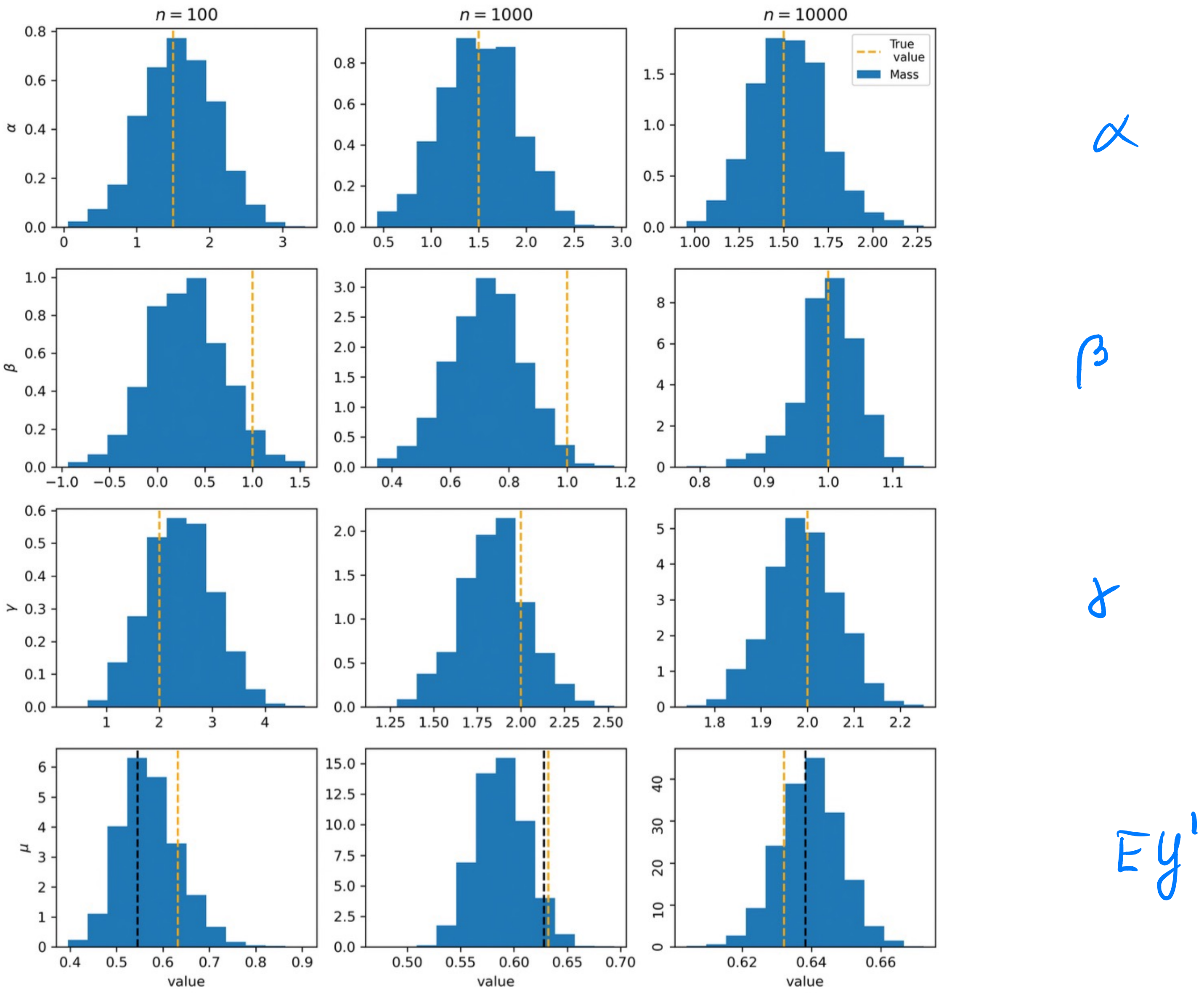
Using Beta process prior is Bayesian equivalent of using the Cox estimator (= MLE)

THM $\sqrt{n}(\beta - \hat{\beta}_{Cox}) \,|\, T_1 \cdots T_n Z_1 \cdots Z_n \rightsquigarrow N(0, \tilde{i}_{Cox}^{-1})$

$\sqrt{n}(\Lambda - \hat{\Lambda}_{Cox}) \,|\, T_1 \cdots T_n Z_1 \cdots Z_n \rightsquigarrow B_{Cox}$

Kim, Lee, 2003

$$\operatorname{logit} P(A=1 \mid y', z) = \gamma z + \overbrace{\alpha y' 1_{(0.5,1]}(z)}^{q(y'|z)} \qquad \Lambda(y'|z) = e^{\beta z} \Lambda(y)$$



Special Case: Binary Outcome

Full Data (A, Y', Z) Observed Data (A, AY', Z)

$Y' \in \{0,1\}$

$P(z) = P(Y'=1|Z)$, $\text{logit}\, P(A=1|Z,Y') = \eta(Z) + q(Y'|Z)$

Strategy 2 $P^Z \sim DP(a) \perp\!\!\!\perp \text{logit}\, P(Y'=1|Z) \sim GP \perp\!\!\!\perp \eta \sim GP$

or Double Robust

Gaussian process

Ray, vdV, 2017

CONJECTURE

- $\sqrt{n}(\theta - \hat{\theta}_{n,q}) \mid \text{DATA}_n, q \rightsquigarrow N(0, \sigma_q^2)$
- $q \mid \text{DATA}_n$ depends on DATA_n

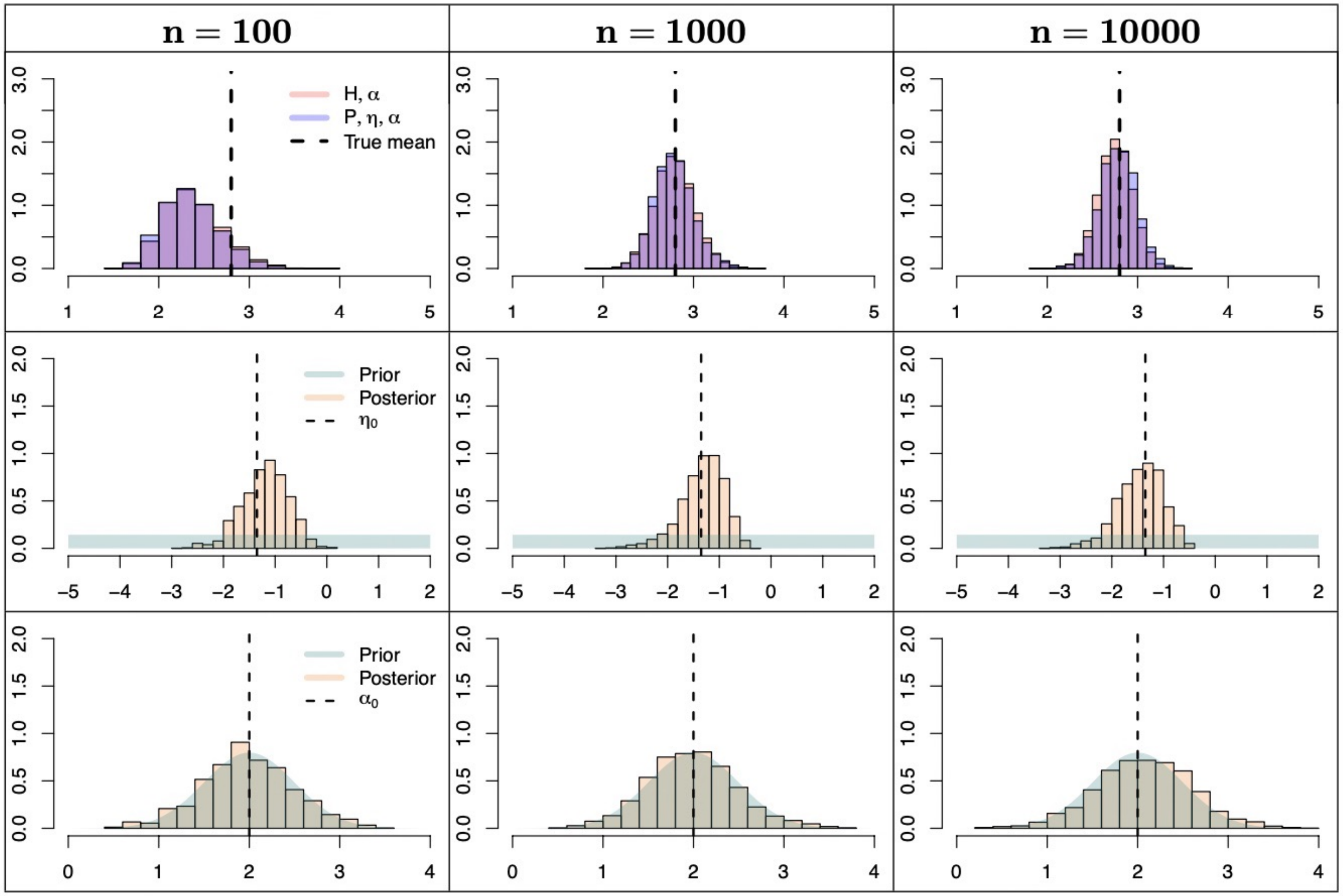
Conclusion

Even though sensitivity parameter is not identified, in a Bayesian setting part of it can be learned from data, even with nonparametric modelling, depending on prior parameterization.

extra slides ⟶

Setting where sensitivity model is centered at truth

Sensitivity function $g(y) = \alpha \log y$, $y \in (0, \infty)$



Ey'

η

α

Gibbs Sampling

With Strategy 2 the $DP(a)$ prior is placed on $P = \mathcal{L}(y')$.
Posterior given $(y_i' : i=1,\cdots,n)$ is $DP(a + \sum_{i=1}^{n} \partial_{y_i'})$.

We observe $(y_i' : A_i = 1)$ but not $(y_i' : A_i = 0)$.

→ Generate $(y_i' : A_i = 0)$ using

$$dP_0(y) \propto e^{-q(y)} dP_1(y)$$

and current posterior draws of parameters (q, P_1).

→ Repeat

Future Research: General Outcome

Challenge: nonparametric prior on $f(y'|z)$

e.g.
- Dependent Dirichlet
- BART + Gaussian error

Challenge: double robustness (or "debiased machine learning")