Debiased Bayesian inference for average treatment effects

Kolyan Ray & Botond Szabó

King's College London & Leiden University

Motivation

Estimating a causal effect from observational data is a common problem, e.g. in healthcare.

If treatment assignment is independent of outcome (e.g. a randomized controlled trial), one can use standard methods. However, often not the case.

Goal: estimate population average treatment effect from observational data, e.g. to decide whether to recommend a new treatment or policy.

Two major difficulties:

- missing counterfactual outcomes,
- selection bias in treatment assignment.

Bayesian (and other) methods can be **badly biased**, especially in complex models or with high-dimensional features (see picture).

Causal inference model

In the potential outcomes model with binary treatments, individual i has two 'potential outcomes':

$$\begin{aligned} Y_i^{(1)} &\quad \text{with treatment} \\ Y_i^{(0)} &\quad \text{without treatment.} \end{aligned}$$

Study the (unobserved) treatment effect $Y_i^{(1)} - Y_i^{(0)}$.

Use a nonparametric causal regression model:

$$Y_i = m(X_i, R_i) + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2),$$

$i = 1, \dots, n$, where

- $X_i \sim^{iid} F$ are features in $\mathbb{R}^d$,
- $R_i = 1$ (treatment) or $R_i = 0$ (no treatment),
- $Y_i = R_i Y^{(1)} + (1 - R_i) Y^{(0)} \in \mathbb{R}$ is the observed outcome.

We assume unconfoundedness:

$$R \perp\!\!\!\perp Y^{(1)}, Y^{(0)} | X,$$

i.e. R (treatment assignment) and $Y^{(0)}, Y^{(1)}$ (outcomes) are conditionally independent given measured features X.

Main idea

Study the marginal posterior for the population average treatment effect

$$\psi = E[Y^{(1)} - Y^{(0)}] = \int_{\mathbb{R}^d} m(x,1) - m(x,0) dF(x).$$

Carefully incorporate an estimate $\hat{\pi}$ of the propensity score $\pi(x) = P(R = 1 | X = x)$ (e.g using logistic regression) into the prior:

$$m(x,r) = W(x,r) + \lambda \left(\frac{r}{\hat{\pi}(x)} + \frac{1-r}{1-\hat{\pi}(x)} \right), \qquad F \sim \text{Dirichlet Process}, \qquad \lambda \sim N(0, \nu_n^2) \quad \text{independent},$$

with W a standard nonparametric prior (e.g. Gaussian process) and $\nu_n^2 \sim 1/n$.

Idea: increase/decrease prior correlation within/across treatment groups in a heterogeneous way compared to a standard prior ($\lambda = 0$).

Reduces bias (correct centering) at the cost of moderately inflating the posterior variance.

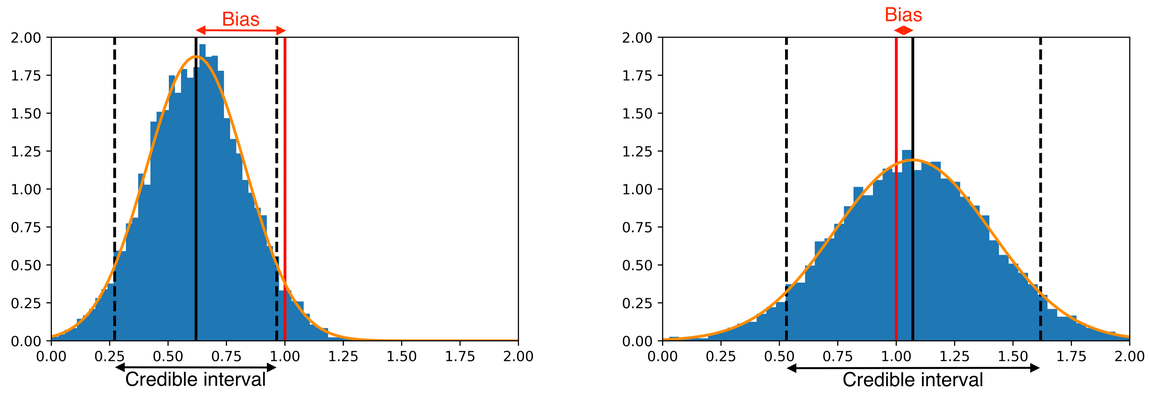

Figure 1: Posterior for average treatment effect ψ with true $\psi = 1.0$ **Left**: unmodified GP prior ($\lambda = 0$) with bias; **Right**: propensity score corrected GP prior with bias reduction

Sample simulation

With $n = 1000$ observations and $d = 100$ features: $X_1, \dots, X_{1000} \sim^{iid} N(0, I_{100})$, non-linear response surface m, non-linear propensity score π and heterogeneous treatment effect $m(x,1) - m(x,0)$.

Method	Abs. error±sd	Width 95% CI±sd	Coverage
GP (unmodified)	0.321 ± 0.027	0.613 ± 0.027	0.38
GP with debiasing (our method)	$\mathbf{0.063 \pm 0.042}$	0.883 ± 0.040	**1.00**
BART	0.228 ± 0.186	1.723 ± 0.490	**1.00**
BART with propensity score	0.134 ± 0.092	0.741 ± 0.079	0.99
Bayesian Causal Forests	0.144 ± 0.109	0.535 ± 0.066	0.87
Causal Forests (AIPW)	0.138 ± 0.097	$\mathbf{0.695 \pm 0.103}$	0.96
Causal Forests (TMLE)	0.136 ± 0.100	0.891 ± 0.152	0.99
Propensity Score Matching	0.234 ± 0.178	1.282 ± 0.158	0.97

Results

- The unmodified GP performs badly, with biased estimation and poor coverage.
- Our method substantially improves both the estimation accuracy and coverage of the GP.
- Makes GPs competitive with state-of-the-art.
- Provides a general route to improving priors (e.g. BART).
- Can put a prior on π, but much slower than using estimator $\hat{\pi}$ ('empirical Bayes').

Intuition & theory

- Should help most when m or π are difficult to estimate, especially with high-dimensional features.
- In such cases, bias $\gg$ variance, so need bias correction (like we do here).
- Idea theoretically investigated in an idealized setting in Ray & van der Vaart (2018).
- Show that for large sample sizes, the posterior for ψ is asymptotically Gaussian (see picture):

$$\psi | (X_i, R_i, Y_i)_{i=1}^n \approx^d N(\hat{\psi}_n, n^{-1} I_0^{-1}),$$

centred at a good (efficient) estimator $\hat{\psi}_n$ with best possible variance (the 'Bernstein-von Mises theorem').

Future directions

- Improve scalability: sparse GP approximations, variational Bayes, distributed computing methods.
- Higher order bias corrections for Bayes (our method corrects 'first order' bias)
- Other causal models.

References

Ray & van der Vaart (2018). Semiparametric Bayesian causal inference. *Annals of Statistics, to appear.*