Targeted Bayesian Learning for Causal Inference: The Best of Both Worlds?

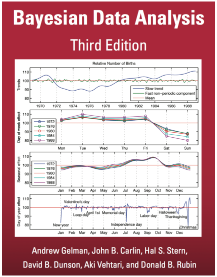
Herbert Susmann
NYU Grossman School of Medicine

ISBA 2024, Venice, Italy

Acknowledgements: We would like to thank Iván Díaz and Antoine Chambaz for helpful comments. The computational requirements for this work were supported in part by the NYU Langone High Performance Computing (HPC) Core's resources and personnel.

My background

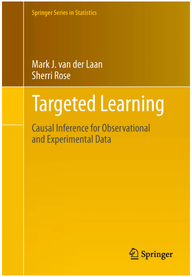
- Framing: science as a social process (Kuhn, 1962)
  - Statistical cultures (Breiman, 2001) and causal inference cultures (Bonvini et al., 2021).
- PhD at UMass Amherst with Leontine Alkema – Bayesian modeling
- Doctoral and postdoc work with Antoine Chambaz (Université Paris Cité) – mathematical statistics, causal inference, non-parametric theory
- Current postdoc work with Iván Díaz (NYU Grossman School of Medicine)



Bayesian Data Analysis

Third Edition

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin



Springer Series in Statistics

Mark J. van der Laan
Sherri Rose

Targeted Learning

Causal Inference for Observational and Experimental Data

Springer

NYU Grossman School of Medicine

Takeaways

- Targeted Maximum Likelihood Estimation (TMLE): frequentist method for estimating smooth finite-dimensional parameters in non-parametric models (van der Laan and Rose, 2011)
- Bayesian TMLE: (pseudo) Bayesian analogue of TMLE. (Díaz Muñoz et al., 2011).
- *This talk:* Bayesian TMLE + hierarchical modeling for estimating group-specific treatment effects.

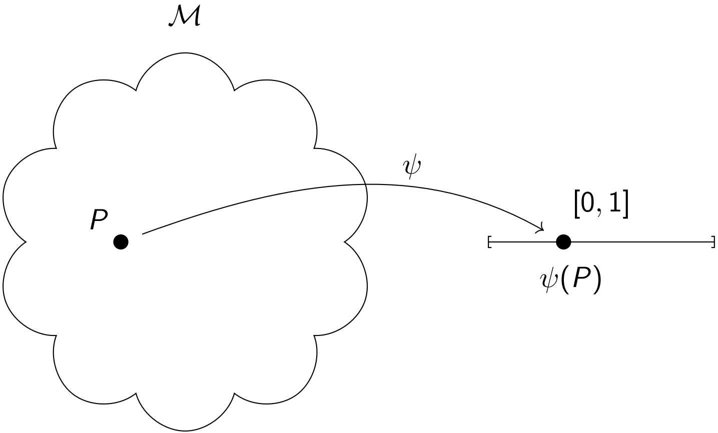
Average Treatment Effect

- Suppose we observe n i.i.d. draws $O_1, \ldots, O_n$ of a variable $O = (X, A, Y) \sim P_0$.
  - X: vector of covariates.
  - A: binary treatment indicator.
  - Y: binary outcome.
- We assume that P_0 falls in the non-parametric model $\mathcal{M}$.
- For any $P \in \mathcal{M}$, define the *Average Treatment Effect* functional as

$$\psi(P) = \mathrm{E}_P[\mathrm{E}_P[Y|A = 1, X] - \mathrm{E}_P[Y|A = 0, X]].$$

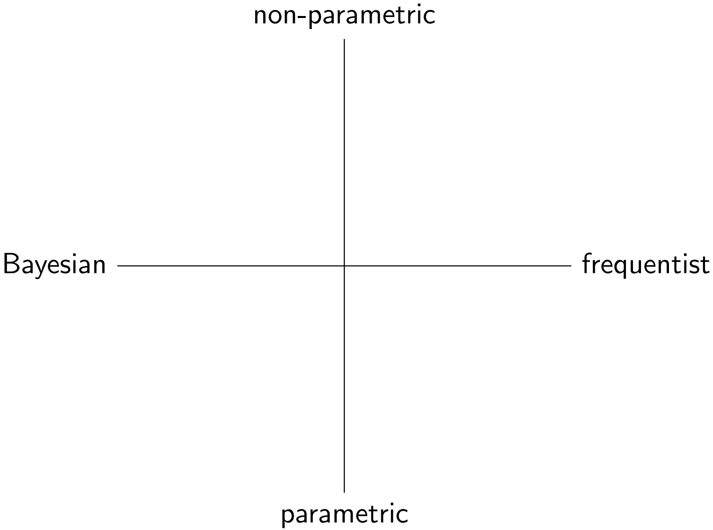
- Notation:
  - Marginal law of X: $Q_P(X) = dP(X)$.
  - Propensity score: $g_P(A, X) = P(A|X)$.
  - Outcome model: $\mu_P(A, X) = \mathrm{E}_P[Y|A, X]$.

Target parameter

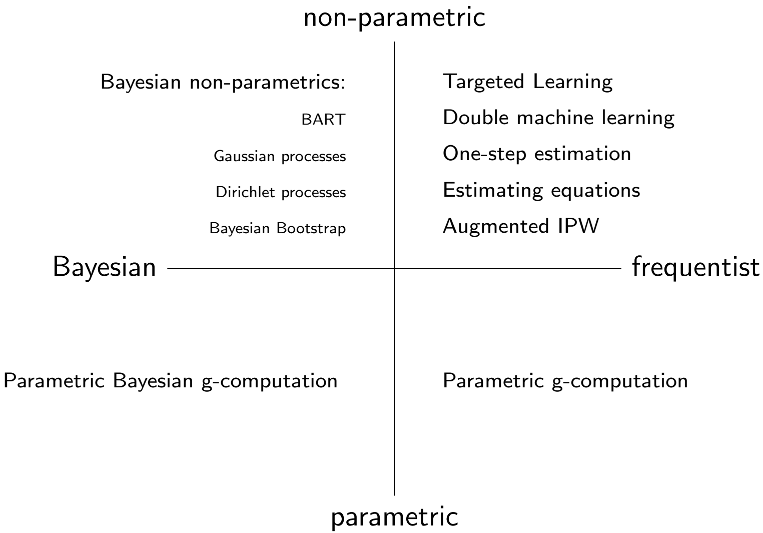


(Warning: infinite dimensional space visualized in two dimensions!)

A crude taxonomy



A crude taxonomy



non-parametric

Bayesian non-parametrics:

BART

Gaussian processes

Dirichlet processes

Bayesian Bootstrap

Targeted Learning

Double machine learning

One-step estimation

Estimating equations

Augmented IPW

Bayesian

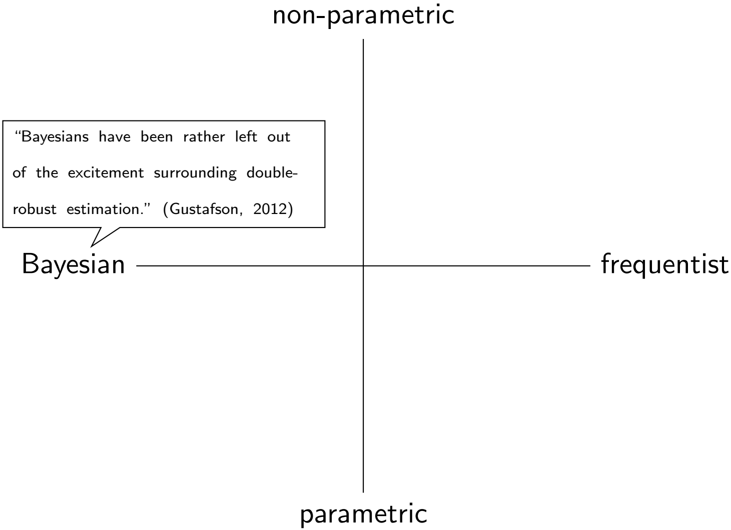
frequentist

Parametric Bayesian g-computation

Parametric g-computation

parametric

A crude taxonomy



non-parametric

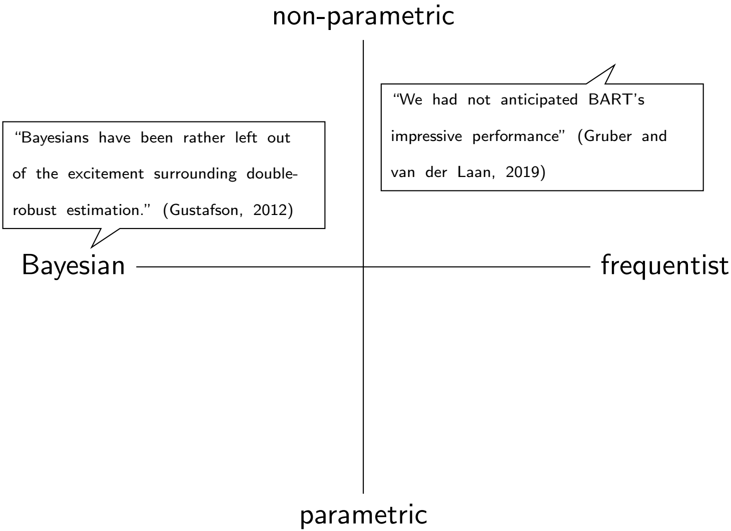
“Bayesians have been rather left out of the excitement surrounding double-robust estimation.” (Gustafson, 2012)

Bayesian

frequentist

parametric

A crude taxonomy



non-parametric

“We had not anticipated BART’s impressive performance” (Gruber and van der Laan, 2019)

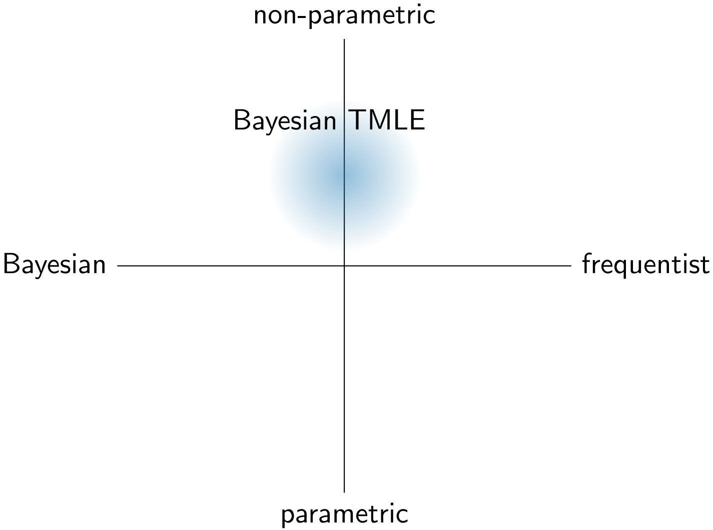
“Bayesians have been rather left out of the excitement surrounding double-robust estimation.” (Gustafson, 2012)

Bayesian

frequentist

parametric

A crude taxonomy



non-parametric

Bayesian TMLE

Bayesian

frequentist

parametric

Frequentist estimation of the ATE

- Any regular estimator $\hat{\psi}_n$ of $\psi(P_0)$ satisfies

$$\hat{\psi}_n = \psi(P_0) + \frac{1}{n}\sum_{i=1}^{n} \mathsf{IF}(P_0)(O_i) + o_p(n^{-1/2}),$$

  where $\mathsf{IF}(P_0) : \mathcal{O} \to \mathbb{R}$ is called an *influence function* of ψ at P_0.

- The influence function with the smallest variance is called the *efficient influence function* (EIF).
- The non-parametric efficiency bound for estimating $\psi(P_0)$ is given by the variance of the efficient influence function: $\mathrm{Var}_{P_0}(\mathsf{EIF}(P_0)(O))$.
- A typical goal for frequentists is to propose an estimator that achieves the above efficiency bound.
- Classical references for semi-parametric efficiency theory: Bickel et al. (1993); van der Vaart and Wellner (1996). Excellent and accessible review: Kennedy (2023)

Efficient influence function of the ATE

- Efficient influence function for the ATE, EIF(P):

$$O \mapsto \frac{2A-1}{g_P(A,X)}\left(Y-\mu_P(A,X)\right)+\mu_P(1,X)-\mu_P(0,X)-\psi(P).$$

- The non-parametric efficiency bound for estimating $\psi(P_0)$:

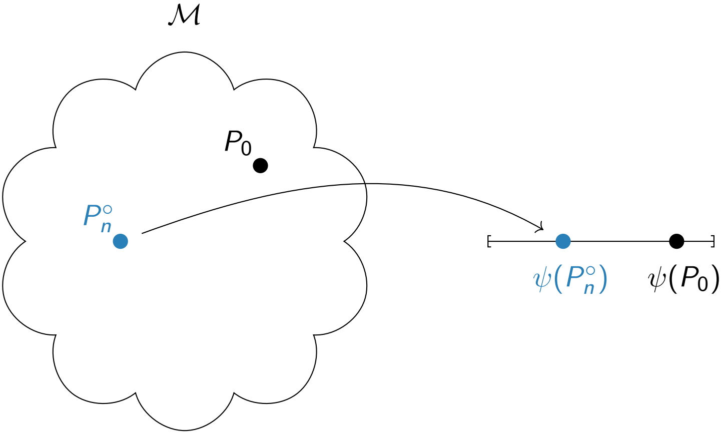
$$\mathsf{E}_{P_0}\left[\frac{\mathrm{Var}_{P_0}(Y|A,X)}{g_{P_0}(A,X)^2}+\left(\mu_{P_0}(1,X)-\mu_{P_0}(0,X)\right)^2\right].$$

Plug-in estimation

- Suppose we have an initial estimate $P_n^\circ = \{\mu_n^\circ, Q_n^\circ\}$ of the parts of P_0 relevant to ψ.
    - Covariate distribution Q_n°: empirical distribution.
    - Outcome model μ_n°: logistic regression, machine learning algorithms, ...
- Form a plug-in estimator of ψ:

$$
\begin{aligned}
\hat{\psi}^{\mathrm{plug-in}} &= \psi(P_n^\circ) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mu_n^\circ(1, X_i) - \mu_n^\circ(0, X_i).
\end{aligned}
$$

Plug-in estimation



Analysis of plug-in estimator

- Apply expansion of the ATE (von Mises, 1947):

$$\psi(P_n^\circ) = \psi(P_0) + \frac{1}{n}\sum_{i=1}^{n} \mathrm{IF}(P_n^\circ)(O_i) + \mathrm{R}(P_n^\circ, P_0).$$

- Problems with plug-in estimator:
  - First-order bias term $\frac{1}{n}\sum_{i=1}^{n} \mathrm{IF}(P_n^\circ)(O_i)$ not necessarily zero.
  - Second-order remainder term $\mathrm{R}(P_n^\circ, P_0)$ doesn't necessarily converge to zero, or does so too slowly.

Targeted Learning

- Many different ways to form unbiased and asymptotically efficient estimators of parameters like the ATE
  - e.g. double machine learning, one-step estimation, estimating equations, augmented IPW
- Frequentist analysis typically proceeds by showing how they make the first-order bias term disappear, and under which conditions the second-order remainder term is $o_p(n^{-1/2})$.
- We focus on *Targeted Maximum Likelihood Estimation* (TMLE) (van der Laan and Rubin, 2006; van der Laan and Rose, 2011).

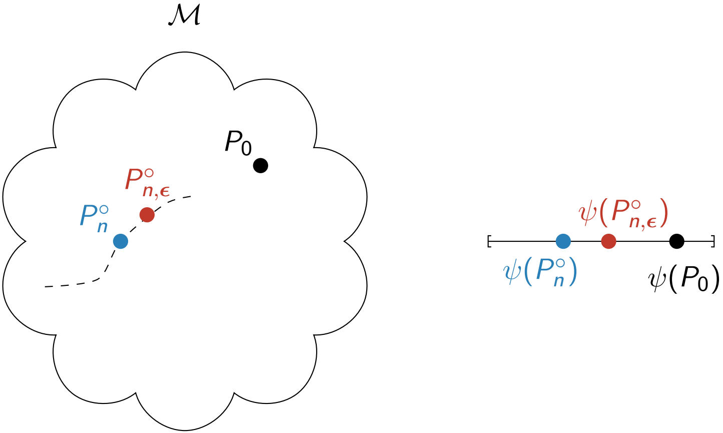
Targeted Maximum Likelihood Estimation

- Key idea: carefully *fluctuate* the initial estimate P_n° in order that the first-order bias term becomes zero.
- More specifically, plug-in estimator of the form

$$\hat{\psi}^{\mathrm{TMLE}} = \psi(P^\circ_{n,\epsilon_n^*})$$

with $\{P^\circ_{n,\epsilon} : \epsilon \in \mathbb{R}\} \subset \mathcal{M}$ a well-chosen *fluctuation* of P_n°, and ϵ_n^* chosen by maximum-likelihood with respect to a carefully chosen likelihood $\mathcal{L}$.

Targeted Maximum Likelihood Estimation



TMLE: details for the ATE

- Fluctuation submodel:

$$\begin{aligned}\operatorname{logit}\left\{\mu^{\circ}_{n,\epsilon}(A,X)\right\} &= \operatorname{logit}\left\{\mu^{\circ}_{n}(A,X) + \frac{2A-1}{g_P(A,X)}\epsilon\right\}, \\ Q^{\circ}_{n,\epsilon}(X) &\propto Q^{\circ}_{n}(X)\exp\left\{\mu^{\circ}_{n}(1,X) - \mu^{\circ}_{n}(0,X) - \psi(P^{\circ}_{n})\right\}.\end{aligned}$$

- Log-likelihood:

$$\begin{aligned}\mathcal{L}(O|\epsilon) =& Y\log(\mu^{\circ}_{n,\epsilon}(A,X)) \\ &+ (1-Y)\log(1-\mu^{\circ}_{n,\epsilon}(A,X)) + \log Q^{\circ}_{n,\epsilon}(X).\end{aligned}$$

TMLE: details for the ATE

- Key property of log-likelihood and fluctuation submodel:

$$\mathsf{EIF}(P) \in \mathrm{span}\left\{ \frac{d}{d\epsilon}\mathcal{L}(\cdot|\epsilon)\bigg|_{\epsilon=0} \right\}$$

- Estimate ϵ_n^* be maximizing $\epsilon \mapsto \mathcal{L}(O_1, \ldots, O_n|\epsilon)$.
- Form a new plug-in estimator as

$$\hat{\psi}^{\mathrm{TMLE}} = \psi(P^{\circ}_{n,\epsilon_n^*})$$

- It can then be shown that

$$\frac{1}{n}\sum_{i=1}^{n} \mathsf{EIF}(P^{\circ}_{n,\epsilon_n^*})(O_i) \approx 0,$$

i.e. the bias term of the expansion is zero.

NYU Grossman School of Medicine

Frequentist properties of TMLE

- Double-robust consistency: remarkably, $\hat{\psi}^{\mathrm{TMLE}}$ is consistent as long as *either* μ_n° *or* g_n° consistently estimate μ_{P_0} or g_{P_0}.
- Asymptotically normal and efficient so long as
  - Nuisance estimation rates: $\|\mu_n^\circ - \mu_{P_0}\| \times \|g_n^\circ - g_{P_0}\| = o_p(n^{-1/2})$.
  - Nuisance estimators are not too complex (e.g. fall in Donsker class). Note such assumptions can be obviated through the use of cross-fitting (Zheng and van der Laan, 2011).

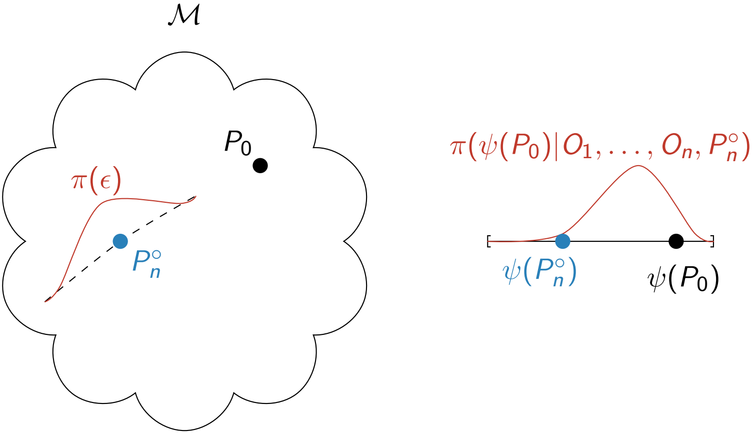
Bayesian TMLE

- Idea: in Bayesian TMLE, apply Bayesian inference to derive a (pseudo) posterior distribution for ϵ:

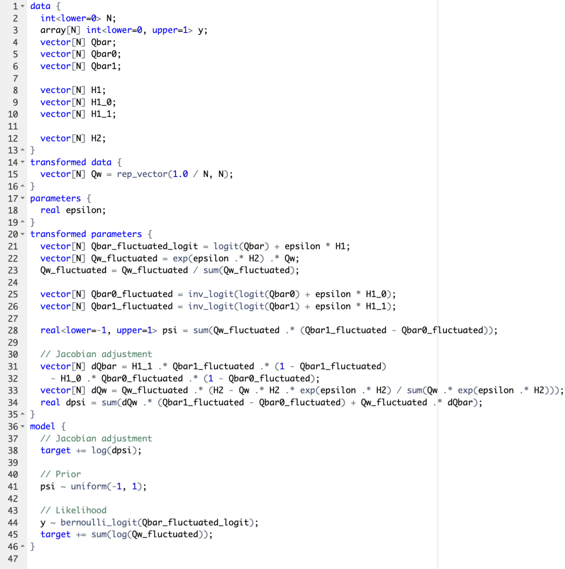
$$\pi(\epsilon | O_1, \ldots, O_n, P_n^\circ) \propto \prod_{i=1}^{n} \mathcal{L}(O_i | \epsilon)\pi(\epsilon),$$

  where $\pi(\epsilon)$ is a prior on ϵ.
- A posterior distribution for ψ can then be formed via the mapping $\psi(P_{n,\epsilon}^\circ)$.
- Bernstein von-Mises type results suggest posterior converges to normal distribution centered on efficient estimator with variance given by the non-parametric efficiency bound, and is double robust.
- Bayesian TMLE has been developed for a few parameters: ATE, class proportions in unlabeled data, parameter of marginal structural models (Díaz Muñoz et al., 2011; Díaz et al., 2020; Susmann and Chambaz, 2023).

Bayesian TMLE



Bayesian TMLE in Stan



```
data {
  int<lower=0> N;
  array[N] int<lower=0, upper=1> y;
  vector[N] Qbar;
  vector[N] Qbar0;
  vector[N] Qbar1;

  vector[N] H1;
  vector[N] H1_0;
  vector[N] H1_1;

  vector[N] H2;
}
transformed data {
  vector[N] Qw = rep_vector(1.0 / N, N);
}
parameters {
  real epsilon;
}
transformed parameters {
  vector[N] Qbar_fluctuated_logit = logit(Qbar) + epsilon * H1;
  vector[N] Qw_fluctuated = exp(epsilon .* H2) .* Qw;
  Qw_fluctuated = Qw_fluctuated / sum(Qw_fluctuated);

  vector[N] Qbar0_fluctuated = inv_logit(logit(Qbar0) + epsilon * H1_0);
  vector[N] Qbar1_fluctuated = inv_logit(logit(Qbar1) + epsilon * H1_1);

  real<lower=-1, upper=1> psi = sum(Qw_fluctuated .* (Qbar1_fluctuated - Qbar0_fluctuated));

  // Jacobian adjustment
  vector[N] dQbar = H1_1 .* Qbar1_fluctuated .* (1 - Qbar1_fluctuated)
    - H1_0 .* Qbar0_fluctuated .* (1 - Qbar0_fluctuated);
  vector[N] dQw = Qw_fluctuated .* (H2 - Qw .* H2 .* exp(epsilon .* H2) / sum(Qw .* exp(epsilon .* H2)));
  real dpsi = sum(dQw .* (Qbar1_fluctuated - Qbar0_fluctuated) + Qw_fluctuated .* dQbar);
}
model {
  // Jacobian adjustment
  target += log(dpsi);

  // Prior
  psi ~ uniform(-1, 1);

  // Likelihood
  y ~ bernoulli_logit(Qbar_fluctuated_logit);
  target += sum(log(Qw_fluctuated));
}

```

Group Treatment Effects

- Suppose that instead of estimating a single Average Treatment Effect, we would like to estimate a set of *group-specific* treatment effects.
- Data structure is now $O = (X, A, D, Y)$, where
  - X: vector of covariates.
  - $D \in \{1, \dots, G\}$: indicator of group membership.
  - A: binary treatment indicator.
  - Y: binary outcome.
- For $g \in \{1, \dots, G\}$, define *group-specific treatment effect* as

$$\psi_g(P) = \mathsf{E}_P[\mathsf{E}_P[Y|A = 1, D, X] - \mathsf{E}_P[Y|A = 0, D, X]|D = g].$$

- Note that the functional ψ_g has a causal interpretation under similar "standard causal assumptions" as the ATE.

Partial pooling

- How should we estimate $\psi_g(P_0)$, especially in instances where there are many groups (G is large) or there are groups with few observations?
- For Bayesians, natural to think about *partial pooling* of group-specific treatment effects (Feller and Gelman, 2015).
- Place a hierarchical model on $\psi_g(P_0)$, for example:

$$\psi_g(P_0) \sim N(m, \sigma^2),$$

with appropriate hyperpriors on m, σ.

Hierarchical models for Bayesian TMLE

- Bayesian TMLE for group-specific treatment effects: set up fluctuation model separately for each group, with fluctuation parameters ϵ_g, $g = 1, \ldots, G$.
- Each ϵ_g maps to a group-specific treatment effect via $\psi_g(P^{\circ}_{n,\epsilon_g})$.
- Place a hierarchical model on the group-specific treatment effects in order to share information between groups:

$$\psi_g(P^{\circ}_{n,\epsilon_g}) \sim N(m, \sigma^2).$$

Illustrative simulation study

- Data generating process:
  - For each group $g = 1, \ldots, G$, draw a group-specific effect $\lambda_g \sim \mathrm{Uniform}(-0.5, 0.5)$.
  - Observations are drawn from

$$
\begin{aligned}
W &\sim \mathrm{N}_5(0_5, I_5),\\
D|W &\sim \mathrm{Categorical}(\{1, \ldots, G\}, \{1, 2, 1, \ldots, \}),\\
A|D, W &\sim \mathrm{Bernoulli}(\mathrm{logit}^{-1}(0.25 + 0.25W^{(1)} + 0.25W^{(2)})),\\
Y|W, D, A &\sim \mathrm{Bernoulli}(\mathrm{logit}^{-1}(-0.25 + 0.5W^{(1)} + 0.25W^{(2)} + A + \lambda_D)).
\end{aligned}
$$

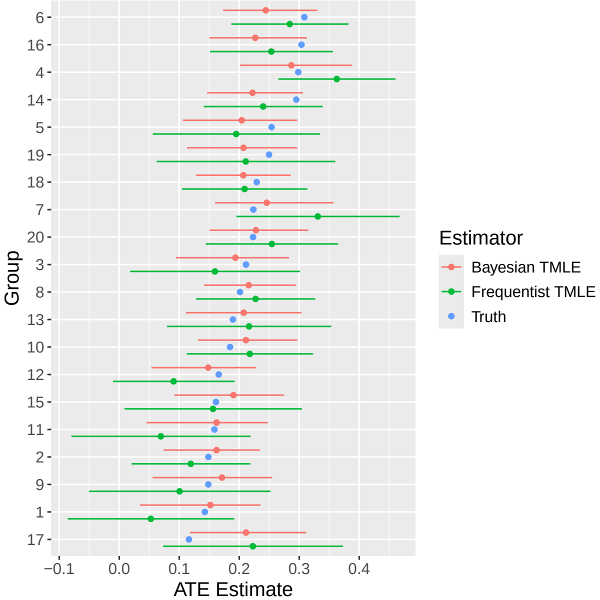
- Form simulation datasets by drawing $N = 2500, 5000, 10000$ observations 100 times from the above data generating process, with $G = 20$ groups.

NYU Grossman School of Medicine

Illustrative simulation study

- Estimate propensity score and outcome models using ensemble estimators.
    - SuperLearner with candidate algorithms `SL.mean`, `SL.knn`, `SL.glm`, `SL.glm.interaction`, `SL.glmnet`.
- Compare against frequentist TMLE run separately for each group.
- Code available on GitHub: `github.com/herbps10/bayes_tmle`.

Simulation example



NYU Grossman School of Medicine

Simulation results: mean error

N	Frequentist TMLE	Bayes hierarchical TMLE
Both consistent		
2500	-0.008	-0.011
5000	-0.003	-0.005
10000	-0.002	-0.003
Propensity consistent, outcome inconsistent		
2500	-0.025	-0.033
5000	-0.018	-0.025
10000	-0.013	-0.017
Propensity inconsistent, outcome consistent		
2500	-0.013	-0.013
5000	-0.007	-0.008
10000	-0.002	-0.002
Both inconsistent		
2500	-0.041	-0.041
5000	-0.040	-0.041
10000	-0.038	-0.038

Simulation results: mean absolute error

N	Frequentist TMLE	Bayes hierarchical TMLE
Both consistent		
2500	0.073	0.047
5000	0.052	0.040
10000	0.037	0.030
Propensity consistent, outcome inconsistent		
2500	0.075	0.053
5000	0.054	0.043
10000	0.038	0.033
Propensity inconsistent, outcome consistent		
2500	0.072	0.046
5000	0.052	0.040
10000	0.036	0.031
Both inconsistent		
2500	0.080	0.056
5000	0.063	0.051
10000	0.048	0.044

Simulation results: empirical coverage

N	Frequentist TMLE	Bayes hierarchical TMLE
Both consistent		
2500	92.37%	93.25%
5000	94.12%	92.32%
10000	93.81%	95.38%
Propensity consistent, outcome inconsistent		
2500	92.63%	90.88%
5000	93.97%	90.41%
10000	94.90%	93.85%
Propensity inconsistent, outcome consistent		
2500	93.45%	93.09%
5000	93.85%	91.77%
10000	95.00%	93.25%
Both inconsistent		
2500	91.70%	87.47%
5000	89.36%	82.45%
10000	83.57%	81.79%

Bayesian TMLE: Best (or worst) of both worlds?

Pros

- Good asymptotic frequentist properties.
- In practice, prior only needs to be placed on the parameter of interest (the group-specific treatment effects).
- Nuisance parameters can be estimated using any method, including flexible data-adaptive algorithms.
- Simulations suggest hierarchical modeling improves finite-sample performance for group-specific treatment effects.

Cons

- Not strictly Bayesian; interpretation of posterior not clear in finite samples.
- May be example of “frequentist pursuit” (Robins et al., 2015).

Discussion

- Bayesian TMLE: (pseudo) Bayesian method with applications in causal inference.
- To Bayesians, our proposal should be pretty obvious: use hierarchical modeling!
- Moral: bringing together ideas from separate “statistical cultures” combines advantages (and disadvantages) of both.
  - Exciting recent work combining Bayesian inference with concepts from non-parametric efficiency theory (Ray and Szabo, 2019; Ray and van der Vaart, 2020; Yiu et al., 2023; Breunig et al., 2024).

Bibliography I

Bickel, P. J., Klaassen, C. A., Ritov, Y., Klaassen, J., and Wellner, J. A. (1993). *Efficient and adaptive estimation for semiparametric models*, volume 4. Springer.

Bonvini, M., Mishler, A., and Kennedy, E. H. (2021). Comment on "statistical modeling: The two cultures" by leo breiman.

Breiman, L. (2001). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199 – 231.

Breunig, C., Liu, R., and Yu, Z. (2024). Double robust Bayesian inference on average treatment effects.

Díaz Muñoz, I., Hubbard, A. E., and van der Laan, M. J. (2011). *Targeted Bayesian Learning*, pages 475–493. Springer New York, New York, NY.

Díaz, I. (2019). Machine learning in the estimation of causal effects: targeted minimum loss-based estimation and double/debiased machine learning. *Biostatistics*, 21(2):353–358.

Díaz, I., Savenkov, O., and Kamel, H. (2020). Nonparametric targeted Bayesian estimation of class proportions in unlabeled data. *Biostatistics*, 23(1):274–293.

Feller, A. and Gelman, A. (2015). *Hierarchical Models for Causal Effects*, pages 1–16. John Wiley & Sons, Ltd.

Gruber, S. and van der Laan, M. J. (2019). Comment on "Automated Versus Do-It-Yourself Methods for Causal Inference: Lessons Learned from a Data Analysis Competition". *Statistical Science*, 34(1):82 – 85.

Bibliography II

Gustafson, P. (2012). Double-robust estimators: Slightly more Bayesian than meets the eye? *The International Journal of Biostatistics*, 8(2):1–15.

Kennedy, E. H. (2023). Semiparametric doubly robust targeted double machine learning: a review.

Kuhn, T. S. (1962). *The structure of scientific revolutions.* The structure of scientific revolutions. Chicago, University of Chicago Press.

Ray, K. and Szabo, B. (2019). Debiased bayesian inference for average treatment effects. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Ray, K. and van der Vaart, A. (2020). Semiparametric Bayesian causal inference. *The Annals of Statistics*, 48(5):2999 – 3020.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9):1393–1512.

Robins, J. M., Hernán, M. A., and Wasserman, L. (2015). Discussion of "On Bayesian Estimation of Marginal Structural Models". *Biometrics*, 71(2):296–299.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational psychology.*, 66(5).

Bibliography III

Susmann, H. and Chambaz, A. (2023). Inference in marginal structural models by automatic targeted bayesian and minimum loss-based estimation.

van der Laan, M. J. and Rose, S. (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, New York.

van der Laan, M. J. and Rubin, D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1).

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer New York, New York, NY.

von Mises, R. (1947). On the Asymptotic Distribution of Differentiable Statistical Functions. *The Annals of Mathematical Statistics*, 18(3):309 – 348.

Yiu, A., Fong, E., Holmes, C., and Rousseau, J. (2023). Semiparametric posterior corrections.

Zheng, W. and van der Laan, M. J. (2011). *Cross-Validated Targeted Minimum-Loss-Based Estimation*, pages 459–474. Springer New York, New York, NY.

Appendix: Causal inference for the average treatment effect

- Separate task into two components: (1) definition and (2) estimation of causal effects (Díaz, 2019).
- Observe n i.i.d. draws $O_1, \ldots, O_n$ of a variable $O = (X, A, Y) \sim P_0$
  - X: vector of covariates
  - A: binary treatment indicator
  - Y: binary outcome
- Let $Y(0)$ and $Y(1)$ be the *potential outcomes* under treatment assignment $A = 0$ and $A = 1$, respectively (Rubin, 1974).
- Define *Average Treatment Effect (ATE)*:

$$\Psi = \mathrm{E}\left[Y(1) - Y(0)\right]$$

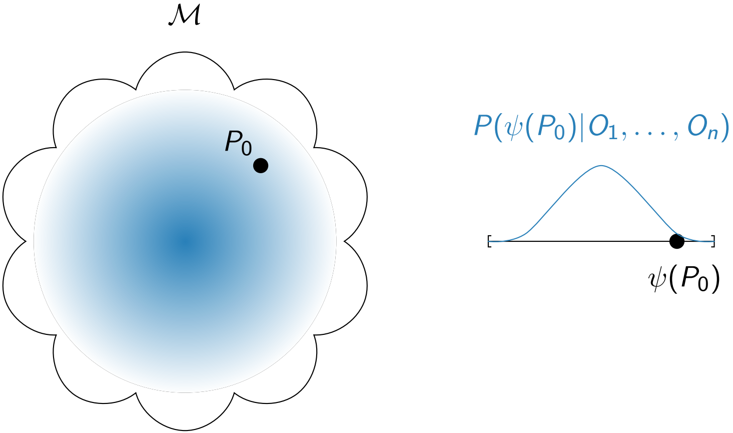
Appendix: Identification

- Assume “Standard causal assumptions”:
  - Consistency: $Y = Y(A)$,
  - Ignorability: $Y(0), Y(1) \perp\!\!\!\perp A | X$,
  - Positivity (overlap): $0 < P(A = 1|X) < 1$ for all X.
- g-computation identification result: (Robins, 1986):

$$\psi(P) = \mathsf{E}_P \left[\mathsf{E}_P[Y|A = 1, X] - \mathsf{E}_P[Y \mid A = 0, X] \right].$$

- We are now firmly in the realm of statistics. How shall we go about estimating $\psi(P)$?

Appendix: Bayesian non-parametrics


$\mathcal{M}$
P_0
$P(\psi(P_0)|O_1, \ldots, O_n)$
$\psi(P_0)$