Causal Segmentation Analysis with Machine Learning in Large-Scale Digital Experiments

Nima Hejazi

Friday, 05 November 2021

Payments and Partnerships Team
Data Science and Engineering
Netflix

nshejazi
nhejazi
nimahejazi.org
with Wenjing Zheng and Sathya Anand

MIT Conference on Digital Experimentation

Why Segment A Population?

- Netflix has 200M+ (and growing) users, constituting a diversity of potential population segments.
- Better understanding this eclectic user-base helps us to
    - improve treatment schedule allocation in A/B experiments,
    - understand and prioritize fairness of treatment impacts,
    - assess differential impacts of proposed product alterations.
- Causal inference provides a *formal* language for discovering and evaluating segments through their treatment effects.

Defining Treatment Effect Heterogeneity

- For a given A/B test or quasi-experiment, we assume data on each of the n units may be expressed $O = (W, A, Y)$, where
  - W: baseline (e.g., region, device type, viewing history),
  - $A \in \{0, 1\}$: treatment assignment (i.e., A vs. B arm),
  - Y: the outcome of interest (e.g., viewing hours).
- Among W, we choose a set of segmentation variables $V \subset W$, whose realizations correspond to user segments of interest.
- The conditional average treatment effect (CATE) evaluates the treatment effect *within a stratum* $v \in V$. The CATE is

$$\text{CATE}(v) = \mathbb{E}[\, \underbrace{\mathbb{E}(Y \mid A = 1, W) - \mathbb{E}(Y \mid A = 0, W)}_{\text{counterfactual mean difference of A=1 vs. A=0}} \mid V = v]$$

Doubly Robust Estimation of Treatment Effect Heterogeneity

- Utilize **doubly robust** estimators of the CATE (Luedtke and van der Laan 2017, VanderWeele et al. 2019).
    - Accurate (i.e., consistent) estimate even when one of the two nuisance quantities is modeled poorly.
    - Efficient (minimal variance) estimate when both well-modeled.
- Ensemble **machine learning** (e.g., van der Laan et al. 2007) for flexible estimation of nuisance quantities.
- **Cross-fitting** (Bickel et al. 1993, Zheng and van der Laan 2011) to identify "should-treat" segments while *preserving inference* for effect measures estimated with machine learning.

Detecting Segments Benefiting from Treatment

- Goal: identify segments T benefiting from treatment, where T are a subset of strata $v \in V$.
- Absolute benefit, defined as $T = \{v : \text{CATE}(v) > \theta\}$ for a given user-specified cutoff $\theta \in \mathbb{R}^+$.
- Relative benefit (subject to cost or side-effects), in which only strata benefiting from treatment ($T \subseteq \{v : \text{CATE}(v) > 0\}$) are subjected to a constraint like $\sum_{v \in T} \text{cost}(v) p(V) \leq \text{budget}$.
- Assign treatment based on CATE point estimates or through hypothesis testing $H_0 : \text{CATE}(v) \leq \theta, H_1 : \text{CATE}(v) > \theta$.

Population Effects of Dynamically Treating Segments

- Dynamic rule: assign treatment only to segments T benefiting from treatment in terms of CATE, i.e., $A = d(V) = \mathbb{I}(v \in T)$.
- Use doubly robust estimators of heterogeneous treatment or optimal treatment effects (HTE, OTE).
- OTE: $\psi_{\text{OTE}} = \mathbb{E}[\mathbb{E}(Y \mid A = d(V), W) - \mathbb{E}(Y \mid A = 1, W)]$, compare dynamic treatment to “treat-all” strategy.
- HTE: ψ_{HTE} compares treatment effects of “should-treat” ($V \in T$) and “should-not-treat” ($V \notin T$) segments.
- Both OTE and HTE characterize the efficacy of the *learned* dynamic rule, informing how interventions should be deployed.

The `sherlock` R Package

- Free and open source data science tool implementing our causal segment discovery framework.
- Supports both causal segment detection and population effect estimation of segment-specific dynamic treatment rules.
- Out-of-the-box machine learning via `sl3` (Coyle et al. 2021) and cross-validation via `origami` (Coyle and Hejazi 2018).
- Available at `https://github.com/Netflix/sherlock`, with plans in place for a release on the CRAN repository.

Illustration in a Quasi-Experimental Study

num_devices	is_p2plus	is_newmarket	baseline_ltv	baseline_viewing	treatment	outcome_viewing
3	0	1	0.539	0.000	1	0.406
2	1	1	1.328	1.637	0	2.328
3	1	0	0.000	0.000	1	3.400
2	1	0	1.027	0.000	0	1.934
2	1	1	0.000	0.000	0	1.376
3	0	1	0.000	1.401	0	2.683

Measurements on six random units from a *synthetic* dataset.

- Baseline covariates (W): account's number of devices, whether a newly enrolled member, being in a new market region, lifetime value of account, account's baseline viewing hours.

Illustration in a Quasi-Experimental Study

num_devices	**is_p2plus**	is_newmarket	baseline_ltv	baseline_viewing	treatment	outcome_viewing
3	0	1	0.539	0.000	1	0.406
2	1	1	1.328	1.637	0	2.328
3	1	0	0.000	0.000	1	3.400
2	1	0	1.027	0.000	0	1.934
2	1	1	0.000	0.000	0	1.376
3	0	1	0.000	1.401	0	2.683

Measurements on six random units from a *synthetic* dataset.

- Segmentation variables ($V \subset W$): account's number of devices (num_devices), whether a newly enrolled member (is_p2plus).

Illustration in a Quasi-Experimental Study

num_devices	is_p2plus	is_newmarket	baseline_ltv	baseline_viewing	**treatment**	outcome_viewing
3	0	1	0.539	0.000	1	0.406
2	1	1	1.328	1.637	0	2.328
3	1	0	0.000	0.000	1	3.400
2	1	0	1.027	0.000	0	1.934
2	1	1	0.000	0.000	0	1.376
3	0	1	0.000	1.401	0	2.683

Measurements on six random units from a *synthetic* dataset.

- Treatment (A, non-randomized): a new user interface.

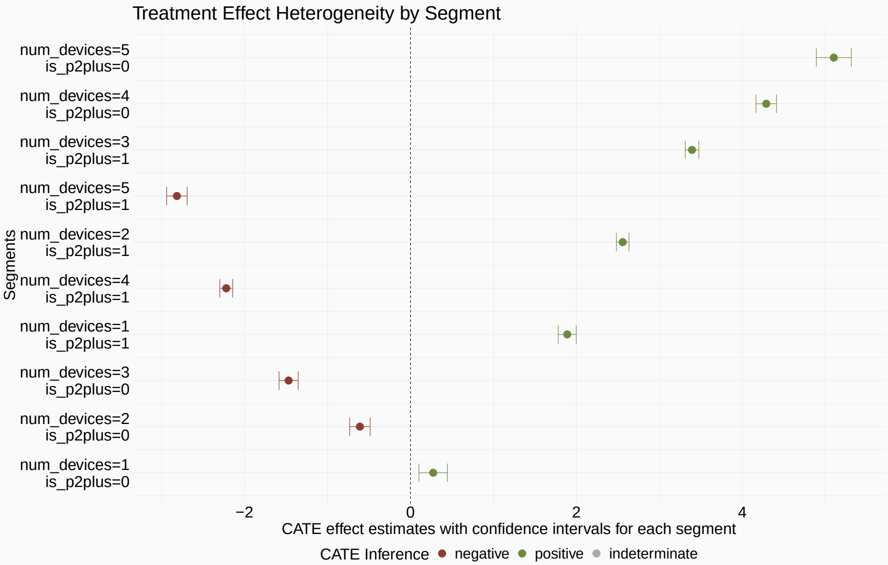
Illustration in a Quasi-Experimental Study

num_devices	is_p2plus	is_newmarket	baseline_ltv	baseline_viewing	treatment	**outcome_viewing**
3	0	1	0.539	0.000	1	0.406
2	1	1	1.328	1.637	0	2.328
3	1	0	0.000	0.000	1	3.400
2	1	0	1.027	0.000	0	1.934
2	1	1	0.000	0.000	0	1.376
3	0	1	0.000	1.401	0	2.683

Measurements on six random units from a *synthetic* dataset.

- Outcome (Y): metric of account's viewing hours.

Treatment Heterogeneity Across Segments



https://github.com/Netflix/sherlock

Thank you!

https://nimahejazi.org

https://twitter.com/nshejazi

https://github.com/nhejazi

https://arxiv.org/abs/2111.01223

References

Bickel, P. J., Klaassen, C. A., Ritov, Y., and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press Baltimore.

Coyle, J. R. and Hejazi, N. S. (2018). origami: A generalized framework for cross-validation in R. *Journal of Open Source Software*, 3(21):512.

Coyle, J. R., Hejazi, N. S., Malenica, I., Phillips, R. V., and Sofrygin, O. (2021). *sl3: Modern Pipelines for Machine Learning and Super Learning*. R package version 1.4.3.

Luedtke, A. R. and van der Laan, M. J. (2017). Evaluating the impact of treating the optimal subgroup. *Statistical Methods in Medical Research*, 26(4):1630–1640.

van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super Learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1).

VanderWeele, T. J., Luedtke, A. R., van der Laan, M. J., and Kessler, R. C. (2019). Selecting optimal subgroups for treatment using many covariates. *Epidemiology*, 30(3):334.

Zheng, W. and van der Laan, M. J. (2011). Cross-validated targeted minimum-loss-based estimation. In *Targeted Learning: Causal Inference for Observational and Experimental Data*, pages 459–474. Springer.