Adaptive normalization for IPW estimation

Johan Ugander
Associate Professor, Management Science & Engineering, Stanford

Joint work with Samir Khan

UChicago Econometrics and Statistics Colloquium, June 2022

Set-up

- Consider the problem of estimating the mean μ of $Y_1, \cdots, Y_n$, where $I_k \sim \text{Ber}(p_k)$ is an indicator of whether or not unit k was observed.

- Horvitz–Thompson and Hájek (self-normalizing) estimators of μ:

$$\hat{\mu}_{\text{HT}} = \hat{S}/n \quad \text{and} \quad \hat{\mu}_{\text{Hájek}} = \hat{S}/\hat{n}$$

  where

$$\hat{S} = \sum_{k=1}^{n} \frac{Y_k I_k}{p_k} \quad \text{and} \quad \hat{n} = \sum_{k=1}^{n} \frac{I_k}{p_k}.$$

- What about the following?

$$\hat{\mu}_\lambda = \frac{\hat{S}}{(1-\lambda)n + \lambda\hat{n}}, \quad \lambda \in \mathbb{R}.$$

IPW: Horvitz–Thompson vs. Hájek

- Fundamental to survey sampling, causal inference, policy learning.
- Only difference between HT and Hájek is how they normalize $\hat{S}$: n vs. $\hat{n}$, an unbiased estimate of n.
- Hájek introduced his ratio estimator in a reply to Basu's "elephants" essay (Basu, 1971).
- Hájek's approach introduces bias, but typically reduces variance (Särndal et al., 2003).
- Connections to self-normalized importance sampling (SNIS) in Monte Carlo, which trace back to Trotter & Tukey (1956) ...

Self-normalization in Monte Carlo

Trotter and Tukey, 1956:

CONDITIONAL MONTE CARLO FOR NORMAL SAMPLES+

Hale F. Trotter and John W. Tukey

Princeton University

The techniques presented here represent what sometimes happens to lazy people who start doing a computation before they quite know what they are going to do. Contrary to experience and high moral principles, this time it worked out all right. We solved our problem (cp. "Monte Carlo techniques in a complex problem about normal samples" pp. 80 ff., below) and we were able to extract some relatively general techniques from what we were driven to do in attempting to solve a particular problem.

As aside, famous for: "the only good Monte Carlos are dead Monte Carlos — the one's we don't have to do."

Self-normalization in Monte Carlo

Trotter & Tukey consider both HT and Hájek estimators . . .

7. If we have N weighted samples (y_1, w_1), (y_2, w_2), ..., (y_N, w_N) from some distribution and are interested in ave $[\phi(z) \mid \text{distribution}]$ we can estimate the average either by

$$\frac{\Sigma w_i \phi(y_i)}{N} \quad \text{or by} \quad \frac{\Sigma w_i \, \phi(y_i)}{\Sigma w_i}$$

Experience shows that the former is almost always better than the latter (as well as being unbiased).

and their "experience" favors HT over Hájek.

Self-normalizing in Monte Carlo

Now consider this uncut gem:

Initially we didn't realize how important it was to distinguish these two estimates -- but experience showed the advantage of dividing by N, the average total weight, rather than by the realized total weight, Σw_i. Indeed, it is possible to use

$$\frac{\Sigma w_i \, \phi(y_i)}{\lambda N + (1 - \lambda)\Sigma w_i}$$

for any real λ, and there is some ground for anticipating that λ's somewhat larger than unity will often be best. (We don't know of any practical experience with such more general estimates.)

Trotter–Tukey proposal

- In our notation, consider the family

$$\hat{\mu}_\lambda = \frac{\hat{S}}{(1-\lambda)n + \lambda\hat{n}}, \quad \lambda \in \mathbb{R}.$$

- At first, $\lambda \in [0, 1]$ seems reasonable. Or?

Trotter–Tukey proposal

- In our notation, consider the family

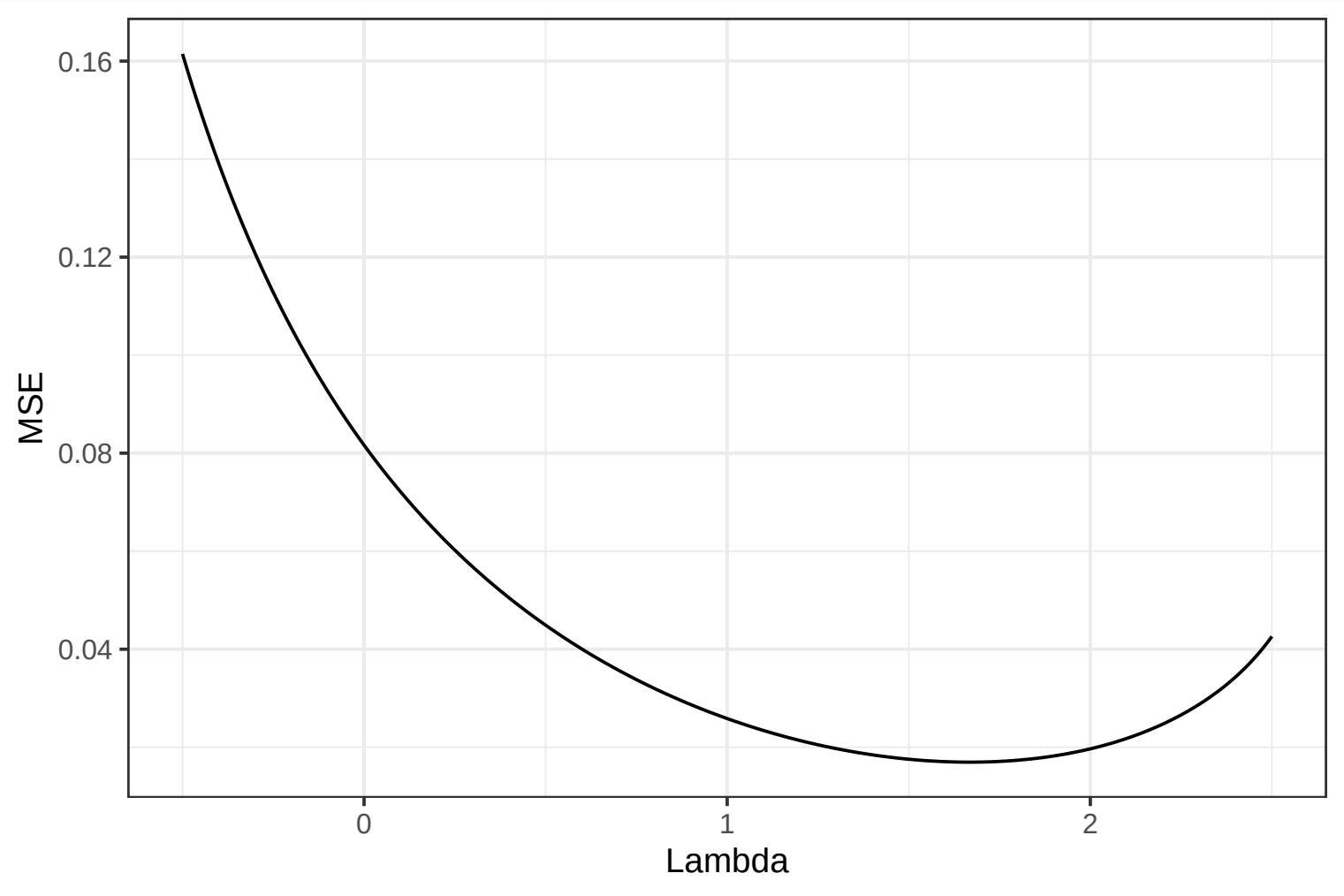
$$\hat{\mu}_\lambda = \frac{\hat{S}}{(1-\lambda)n + \lambda\hat{n}}, \quad \lambda \in \mathbb{R}.$$

- At first, $\lambda \in [0, 1]$ seems reasonable. Or?
- Consider a toy example where

$$Y_1 = Y_2 = \cdots = Y_{10} = 1, \qquad p_1 = 10^{-5}, p_2 = \cdots = p_{10} = 0.5.$$

What happens to HT/Hájek when Y_1 is observed?

Realistic example, MSE of $\hat{\mu}_\lambda$



Based on a realistic example from later in talk.

Adaptive normalization

Model

- Suppose pairs $(Y_1, p_1), \cdots, (Y_n, p_n)$ are drawn i.i.d. from a super-population distribution $\mathcal{D}$ on $\mathbb{R} \times [0, 1]$.
- Our goal is to estimate $\mu = \mathbb{E}[Y_k]$.
- We assume throughout that $|Y_k| \leq M$ and $\delta \leq p_k \leq 1 - \delta$ almost surely.
- Our results continue to hold in a finite population model, with slightly different assumptions.

Choosing a value of λ

Recall:

$$\hat{\mu}_\lambda = \frac{\hat{S}}{(1-\lambda)n + \lambda\hat{n}}, \lambda \in \mathbb{R}.$$

How do we pick values of λ other than $\lambda = 0$ and $\lambda = 1$?

Choosing a value of λ

Recall:

$$\hat{\mu}_\lambda = \frac{\hat{S}}{(1-\lambda)n + \lambda\hat{n}}, \lambda \in \mathbb{R}.$$

How do we pick values of λ other than $\lambda = 0$ and $\lambda = 1$?

Theorem

For any fixed $\lambda \in \mathbb{R}$, we have the CLT

$$\sqrt{n}(\hat{\mu}_\lambda - \mu) \xrightarrow{d} N\left(0, \sigma_\lambda^2\right), \quad \sigma_\lambda^2 = \mathbb{E}\left[\frac{1-p_k}{p_k}(Y_k - \lambda\mu)^2\right].$$

Minimizing asymptotic variance suggests that we should use

$$\lambda^* = \frac{\mathbb{E}\left[\frac{1-p_k}{p_k} Y_k\right]}{\mathbb{E}\left[\frac{1-p_k}{p_k}\right] \mu} := \frac{T}{\pi\mu}.$$

Interpreting λ^*

- What does

$$\lambda^* = \frac{\mathbb{E}\left[\frac{1-p_k}{p_k} Y_k\right]}{\mathbb{E}\left[\frac{1-p_k}{p_k}\right] \mathbb{E}[Y_k]} := \frac{T}{\pi\mu}$$

  look like in different cases?

- If Y_k and p_k are positively correlated, then Y_k and $\frac{1-p_k}{p_k}$ are negatively correlated, so $T < \mu\pi$ and $\lambda^* < 1$.
- If Y_k and p_k are negatively correlated, we have $\lambda^* > 1$.
- Extends the conventional wisdom that Hájek is preferable to HT when Y_k and p_k are negatively correlated (Särndal et al., 2003).
- Trotter–Tukey's Monte Carlo experience was probably in a setting where Y_k and p_k were positively correlated.

Estimating λ^* from the data

- Since we do not know λ^*, we have to estimate it from the data.
- We can estimate μ by $\hat{\mu}_{\mathsf{HT}}$ and T, π by the IPW estimators

$$\hat{T} = \frac{1}{n} \sum_{k=0}^{n} \frac{1 - p_k}{p_k} Y_k \frac{I_k}{p_k}, \quad \hat{\pi} = \frac{1}{n} \sum_{k=0}^{n} \frac{1 - p_k}{p_k} \frac{I_k}{p_k},$$

- This leads to the estimators

$$\hat{\lambda}^* = \frac{\hat{T}}{\hat{\pi} \hat{\mu}_{\mathsf{HT}}}, \quad \hat{\mu}_{\hat{\lambda}^*} = \frac{\hat{S}}{(1 - \hat{\lambda}^*) n + \hat{\lambda}^* \hat{n}}.$$

Estimating λ^* from the data

- Since we do not know λ^*, we have to estimate it from the data.
- We can estimate μ by $\hat{\mu}_{\mathsf{HT}}$ and T, π by the IPW estimators

$$\hat{T} = \frac{1}{n} \sum_{k=0}^{n} \frac{1 - p_k}{p_k} Y_k \frac{I_k}{p_k}, \quad \hat{\pi} = \frac{1}{n} \sum_{k=0}^{n} \frac{1 - p_k}{p_k} \frac{I_k}{p_k},$$

- This leads to the estimators

$$\hat{\lambda}^* = \frac{\hat{T}}{\hat{\pi} \hat{\mu}_{\mathsf{HT}}}, \quad \hat{\mu}_{\hat{\lambda}^*} = \frac{\hat{S}}{(1 - \hat{\lambda}^*) n + \hat{\lambda}^* \hat{n}}.$$

- Wouldn't it be better to estimate λ^* using $\hat{\mu}_{\hat{\lambda}^*}$ instead of $\hat{\mu}_{\mathsf{HT}}$?

Estimating λ^* from the data

- In general, there is an EM-like iteration: a better estimate of λ^* leads to a better estimate of μ, and a better estimate of μ leads to a better estimate of λ^*.
- Suggests an iterative sequence of estimators initialized at $(\hat{\lambda}^{(0)}, \hat{\mu}^{(0)}) = (0, \hat{\mu}_{\mathsf{HT}})$ and defined for $t \geq 1$ by

$$\hat{\lambda}^{(t)} = \frac{\hat{T}}{\hat{\pi}\hat{\mu}^{(t-1)}}, \quad \hat{\mu}^{(t)} = \frac{\hat{S}}{(1-\hat{\lambda}^{(t)})n + \hat{\lambda}^{(t)}\hat{n}}.$$

Fixed points of the iteration

- The iterations

$$\hat{\lambda}^{(t)} = \frac{\hat{T}}{\hat{\pi}\hat{\mu}^{(t-1)}}, \quad \hat{\mu}^{(t)} = \frac{\hat{S}}{(1-\hat{\lambda}^{(t)})n + \hat{\lambda}^{(t)}\hat{n}}.$$

  have two possible limiting behaviors
- One is $\mu^{(t)} \to 0, \hat{\lambda}^{(t)} \to \infty$.
- The other is convergence to the fixed point

$$\hat{\mu}_{\mathsf{AN}} = \frac{\hat{S}}{n} + \frac{\hat{T}}{\hat{\pi}}\left(1 - \frac{\hat{n}}{n}\right).$$

- $\hat{\mu}_{\mathsf{AN}}$ is $\hat{\mu}_{\mathsf{HT}}$ plus a correction factor(!?).
- Can also derive $\hat{\mu}_{\mathsf{AN}}$ as a direct joint minimization, over (λ, μ), of the asymptotic variance.

Convergence of the iterations

Theorem

Suppose $\mu \neq 0$, and consider the sequence of estimators $(\hat{\lambda}^{(t)}, \hat{\mu}^{(t)})$ initialized at $\hat{\lambda}^{(0)} = 0, \hat{\mu}^{(0)} = \hat{\mu}_{HT}$ and defined for $t \geq 1$ by the recursions above. Then

(i) *the sequence $\hat{\mu}^{(t)}$ converges as $t \to \infty$ to an estimator $\hat{\mu}_{lim}$;*

(ii) *the estimator $\hat{\mu}_{lim}$ satisfies*

$$\lim_{n \to \infty} \mathbb{P}\left(\hat{\mu}_{lim} = \hat{\mu}_{AN}\right) = 1,$$

so that $\hat{\mu}_{lim} - \hat{\mu}_{AN}$ converges in probability to 0.

So the process of repeatedly learning better and better estimates of λ^* culminates in $\hat{\mu}_{\mathrm{AN}}$.

Proof sketch

- The first step is to show that the $\hat{\mu}^{(t)}$ converges to $\hat{\mu}_{\mathsf{AN}}$ on the event

$$\left|\frac{\hat{S}}{n}\right| > \left|\frac{\hat{T}}{\hat{\pi}}\left(1 - \frac{\hat{n}}{n}\right)\right|$$

- The second step is to establish that the above event occurs with high probability using the fact that $\hat{S}/n$ concentrates around μ and $\frac{\hat{T}}{\hat{\pi}}\left(1 - \frac{\hat{n}}{n}\right)$ concentrates around 0.

Asymptotic variance of $\hat{\mu}_{\mathsf{AN}}$

Recall:

$$\hat{\mu}_{\mathsf{AN}} = \frac{\hat{S}}{n} + \frac{\hat{T}}{\hat{\pi}}\left(1 - \frac{\hat{n}}{n}\right).$$

Theorem

The estimator $\hat{\mu}_{AN}$ satisfies the CLT

$$\sqrt{n}(\hat{\mu}_{AN} - \mu) \xrightarrow{d} N\left(0, \mathbb{E}\left[\frac{1 - p_k}{p_k}\left(Y_k - \frac{T}{\pi}\right)^2\right]\right).$$

Furthermore, the asymptotic variance above is always smaller than the asymptotic variances of $\hat{\mu}_{HT}$ and $\hat{\mu}_{Hájek}$, and is strictly smaller except if equivalent.

Connections to regression controls

- Consider the regression control family

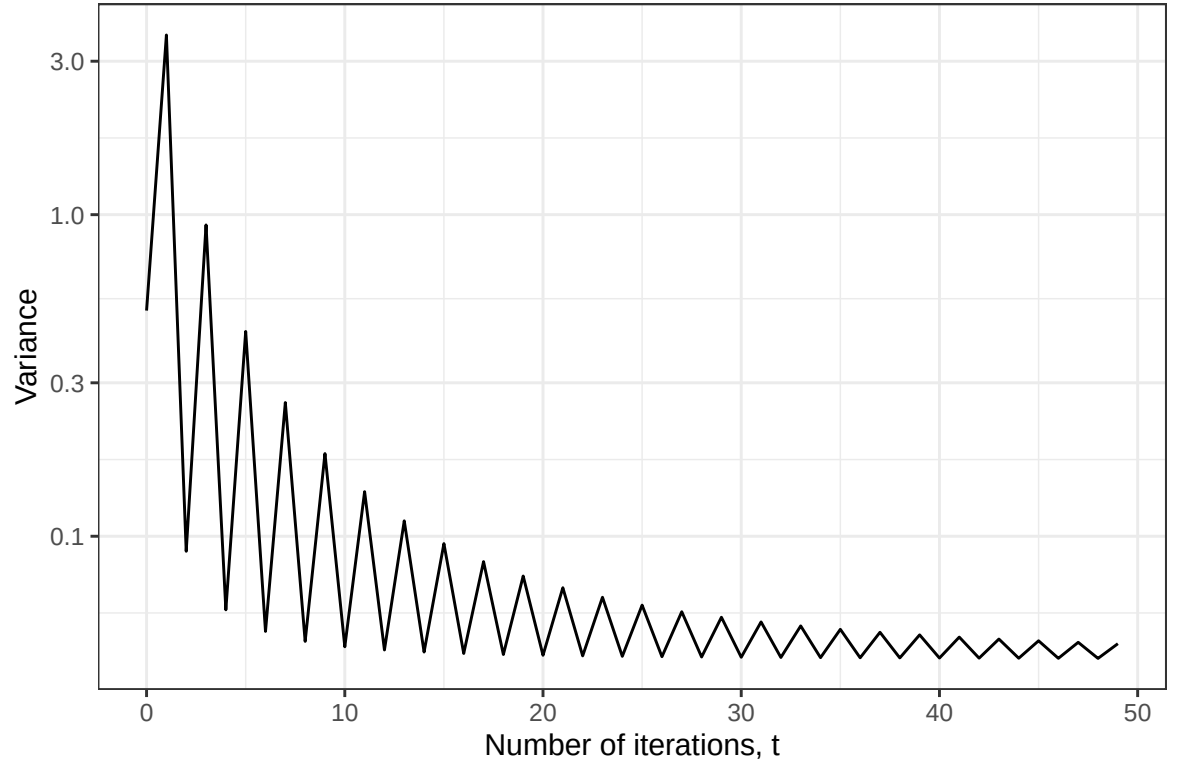
$$\hat{\mu}_\beta = \frac{1}{n}\sum_{k=1}^{n}\frac{Y_k I_k}{p_k} - \beta\left(\frac{1}{n}\sum_{k=1}^{n}\frac{I_k}{p_k} - 1\right)$$

  and selecting β^* to minimize the variance.
- The choice of I_k/p_k as a regression control in Monte Carlo problems is considered in Hesterberg (1995). Owen (2013) recommends $\hat{\mu}_{\beta^*}$ over HT/Hájek for Monte Carlo!
- The regression control estimator $\hat{\mu}_{\beta^*}$ is, surprisingly, algebraically equivalent to $\hat{\mu}_{\text{AN}}$, i.e. adaptive normalization.

A finite-sample variance conjecture

- Starting from $\hat{\mu}^{(0)} = \hat{\mu}_{\mathsf{HT}}$, move to better estimates $\hat{\mu}^{(1)}, \hat{\mu}^{(2)}, \ldots$ based on better information about the correlation structure.
- See paper for (incomplete!) contraction mapping argument for finite-sample variance reduction.
- See also Hansen & Lee (2021), studying variance reduction from iterated GMM proceedure. Same incomplete argument.

A finite-sample variance conjecture



- The variance of the iterative estimator $\hat{\mu}^{(t)}$ as a function of t.
- Although a single iteration may increase the variance, we observe that, always in simulation, every two iterations reduce variance.

Applications beyond survey sampling

AIPW estimation

- Consider the more general model where we have pairs $(Y_1, X_1), \cdots, (Y_n, X_n)$ and $p_k = p(X_k)$ is a function of the covariates.
- In this context, the AIPW estimator (Robins, Rotnitzky, & Zhao 1994) of μ first estimates the response surface $\mu(X_k) = \mathbb{E}[Y_k \mid X_k]$ and the propensity map $p(X_k)$ non-parametrically, and then estimates μ by

$$\hat{\mu}_{\text{AIPW}} = \frac{1}{n}\sum_{k=1}^{n} \hat{\mu}(X_k) + \frac{1}{n}\sum_{k=1}^{n} \frac{(Y_k - \hat{\mu}(X_k))I_k}{\hat{p}(X_k)}$$

  .
- The second term here is a Horvitz–Thompson estimator—can we replace it with an adaptively normalized estimator?

Adaptively normalized AIPW estimation

- This suggests the estimator

$$\hat{\mu}_{\text{AIPW,AN}} = \frac{1}{n}\sum_{k=1}^{n}\hat{\mu}(X_k) + \frac{1}{n}\sum_{k=1}^{n}\frac{(Y_k - \hat{\mu}(X_k))I_k}{\hat{p}(X_k)}$$

$$+\frac{1}{\hat{\pi}}\left(\sum_{k=1}^{n}(Y_k - \hat{\mu}(X_k))\frac{1-\hat{p}(X_k)}{\hat{p}(X_k)}\frac{I_k}{\hat{p}(X_k)}\right)\left(1 - \frac{\hat{n}}{n}\right)$$

Theorem

Assume that $\hat{\mu}(\cdot)$ and $\hat{p}(\cdot)$ are uniformly consistent and that they also satisfy the risk decay condition

$$\mathbb{E}\left[(\hat{\mu}(X_k) - \mu(X_k))^2 \mid \mathcal{T}_n\right] \times \mathbb{E}\left[(\hat{p}(X_k) - p(X_k))^2 \mid \mathcal{T}_n\right] = o_P(n^{-1}).$$

Then

$$\sqrt{n}(\hat{\mu}_{AIPW,AN} - \hat{\mu}_{AIPW}) \xrightarrow{\mathbb{P}} 0.$$

Policy learning

- Suppose individual k has potential outcomes $Y_k(1)$ and $Y_k(0)$ depending on whether or not they receive a treatment, and we wish to learn a policy π that maps known covariates X_k to a treatment assignment in $\{0, 1\}$.
- The value of a policy π is $V(\pi) = \mathbb{E}[Y_k(\pi(X_k))]$. We would like to maximize V, but we cannot compute it, so Kitagawa & Tetenov (2018) propose estimating V from historical data $(Y_1(I_1), X_1), \cdots, (Y_n(I_n), X_n)$ by the surrogate

$$\hat{V}_{\text{IPW}}(\pi) = \frac{1}{n} \sum_{k=1}^{n} \frac{1\{I_k = \pi(X_k)\}\, Y_k}{\mathbb{P}(I_k = \pi(X_k) \mid X_k)}.$$

Adaptively normalized policy learning

- Continuing our theme, we propose minimizing $\hat{V}_{\mathrm{AN}}(\pi) = \hat{V}_{\mathrm{IPW}}(\pi) +$

$$\frac{\sum_{k=1}^{n} Y_k \frac{1-\mathbb{P}(I_k=\pi(X_k)|X_k)}{\mathbb{P}(I_k=\pi(X_k)|X_k)} \frac{1\{I_k=\pi(X_k)\}}{\mathbb{P}(I_k=\pi(X_k)|X_k)}}{\sum_{k=1}^{n} \frac{1-\mathbb{P}(I_k=\pi(X_k)|X_k)}{\mathbb{P}(I_k=\pi(X_k)|X_k)} \frac{1\{I_k=\pi(X_k)\}}{\mathbb{P}(I_k=\pi(X_k)|X_k)}} \left(1 - \frac{1}{n}\sum_{k=1}^{n} \frac{1\{I_k = \pi(X_k)\}}{\mathbb{P}(I_k = \pi(X_k) \mid X_k)}\right)$$

  instead.

Theorem

Fix a class of policies Π *with finite VC-dimension and assume the potential outcomes* $Y_k(1), Y_k(0)$ *are bounded. Let* $\hat{\pi}_{AN} = \arg\max_{\pi\in\Pi} \hat{V}_{AN}(\pi)$ *and* $\pi^* = \arg\max_{\pi\in\Pi} V(\pi)$*. Then*

$$\mathbb{E}[V(\pi^*) - V(\hat{\pi}_{AN})] \leq O\left(\frac{M}{\delta}\sqrt{\frac{VC(\Pi)}{n}}\right).$$

Experiments

Survey sampling of Swiss municipalities

- Data set of 2896 municipalities in Switzerland.
- Two responses: Y_1, the wooded area, and Y_2, the industrial area.
- Assume sampling scheme in which p_k is proportional to total area of municipality and $\sum_k p_k$ is either 50 or 250.
- Want to estimate sum from sample with non-uniform probabilities.
- Test problem from R package `sampling`.

	Problem specification			
	$\sum = 50, Y_1$	$\sum = 250, Y_1$	$\sum = 50, Y_2$	$\sum = 250, Y_2$
$\hat{\mu}_{\text{HT}}$	68.4	27.8	2.51	1.07
$\hat{\mu}_{\text{Hájek}}$	95.3	39.3	2.52	1.06
$\hat{\mu}_{\text{AN}}$	61.5	23.1	2.45	1.01

Table 1: RMSE of estimators on Swiss municipality data; Y_1 is wood area and Y_2 is industrial area, while Σ is the sum of the p_k; probabilities are chosen proportional to total municipality area, which is strongly positively correlated with Y_1 and weakly positively correlated with Y_2. RMSEs are averaged over 100,000 trials.

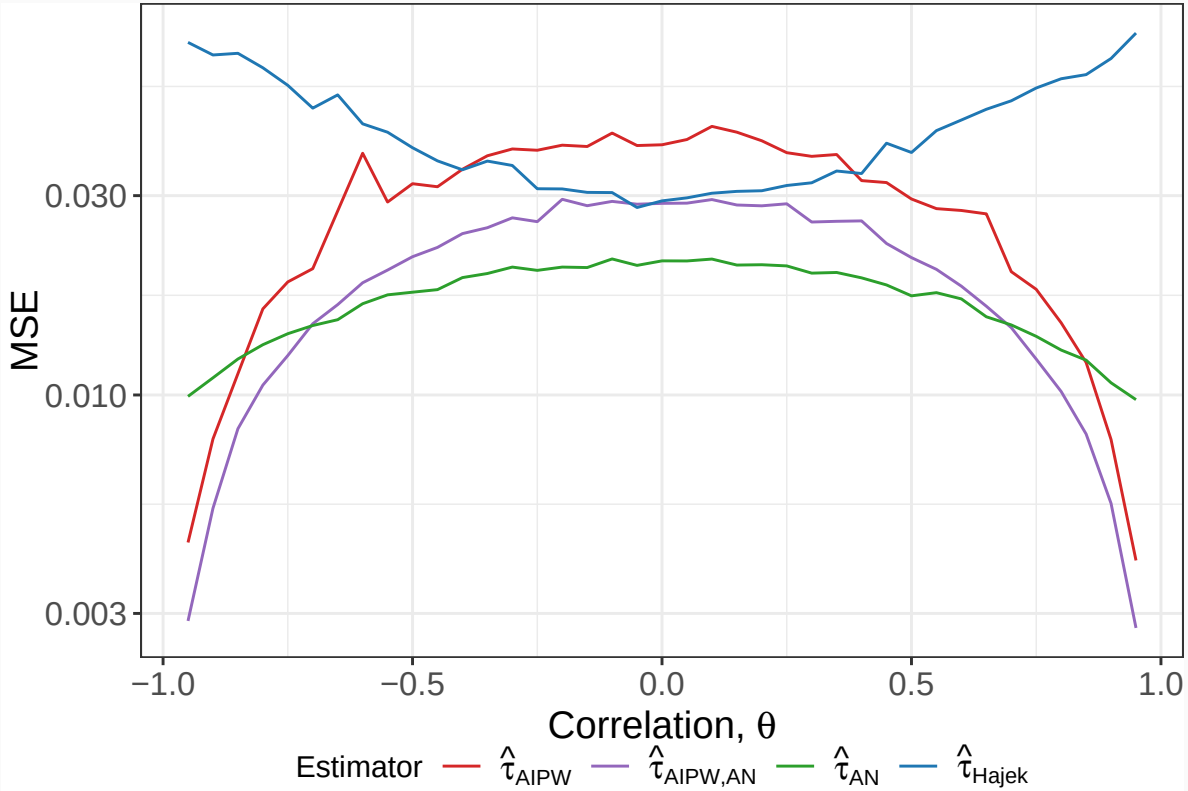
ATE estimation in a normal model

- Consider data generated according to the normal model

$$(Y_k(0), X_k) \sim N\left(\begin{bmatrix} \mu \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \theta \\ \theta & 1 \end{bmatrix}\right), \quad Y_k(1) = Y_k(0) + \tau,$$

  and $p_k = \frac{1}{1+\exp(-2X_k)}$.
- This represents a setting where Y_k and p_k have an approximately linear relationship, the strength of which is controlled by θ.
- For the AIPW estimators, we estimate p_k from logistic regression on X_k and Y_k from a GAM fit on X_k.

Normal model simulations



The estimated MSE of all discussed estimators on data generated from the normal model with $n = 500$ and $\mu = 1$ for different values of θ.

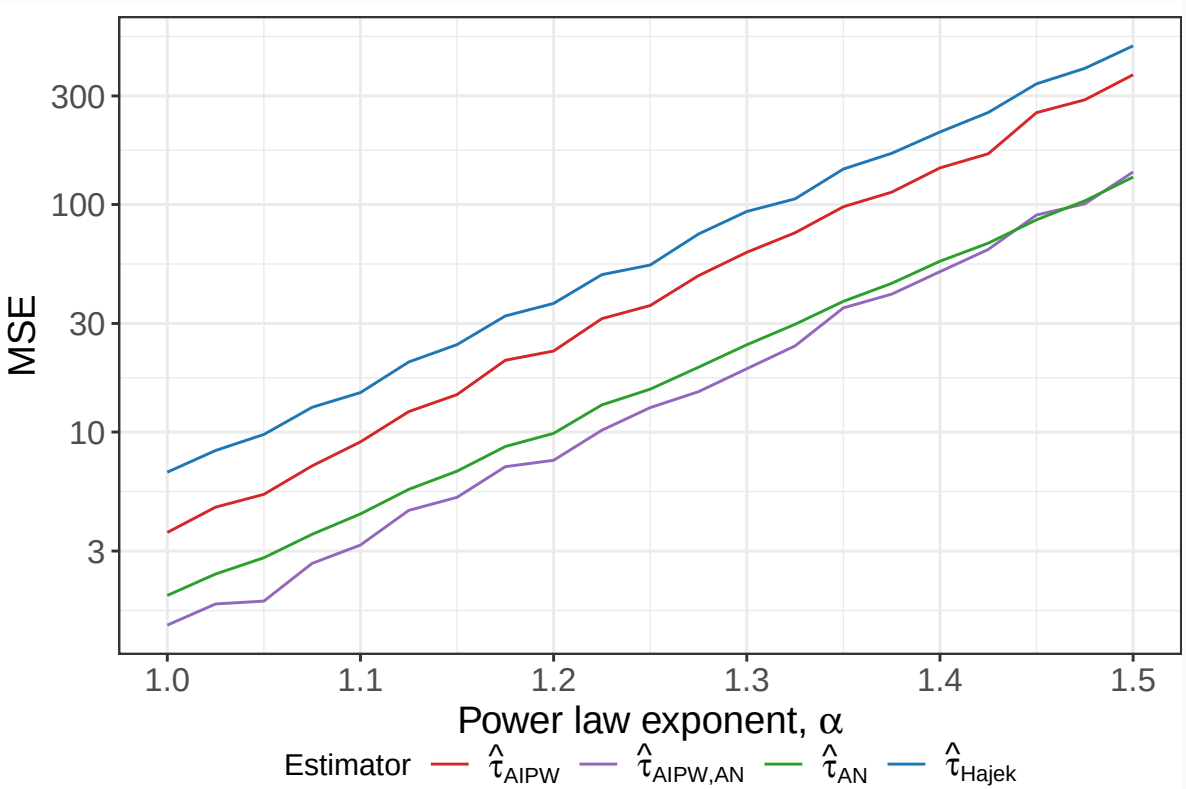
Survey sampling in a power law model

- For a more challenging setting, we generate

$$p_k \sim \text{Uni}(\epsilon, 1-\epsilon), \quad Y_k(0) = p_k^{-\alpha} + N(0, \sigma^2),$$

  and $Y_k(1) = Y_k(0) + \tau, X_k = \log\left(\frac{1-p_k}{p_k}\right)$.
- This corresponds to Y_k and p_k with a strong negative relationship, whose strength is controlled by α.

Power law model simulations



The estimated MSE of all discussed estimators on data generated from the power law model with $n = 500$ and α varying.

Policy learning experiments

- We generate data inspired by Athey & Wager (2021):

$$X_k \sim N(0, I_{3\times 3}), \quad p(X_k) = \frac{1}{1+\exp(-X_{k,1})},$$
$$Y_k(0) = X_{k,1},\ Y_k(1) = Y_k(0) + \text{sgn}(X_{k,2} + X_{k,3})$$

  where $X_{k,i}$ is the i^{th} entry of X_k.
- We learn a policy of the form $1\{X_{k,2} > T\}$ for $T \in [-1, 1]$ by grid search on $\hat{V}_{\text{IPW}}$ and $\hat{V}_{\text{AN}}$.

	Sample size			
Objective	$n = 250$	$n = 500$	$n = 750$	$n = 1000$
$\hat{V}_{\text{IPW}}$	-0.057	-0.035	-0.026	-0.020
$\hat{V}_{\text{AN}}$	-0.039	-0.015	-0.010	-0.004

Table 2: Thresholds learned by optimizing $\hat{V}_{\text{IPW}}$ and $\hat{V}_{\text{AN}}$ on samples of different sizes of data generated. Each entry is the average threshold chosen over 100,000 trials. The optimal policy is to threshold at 0, so we see that minimizing $\hat{V}_{\text{AN}}$ consistently learns better thresholds.

Summary

- Trotter & Tukey (1956) had a simple, powerful, overlooked idea.
- IPW with adaptive normalization, minimizing asymptotic variance, is a good idea.
- Magic upgrade for IPW in AIPW, ATE estimation, policy learning, your problem?
- Open problem: finite sample variance reduction?

- Khan & Ugander (2021): arXiv:2106.07695
- Trotter & Tukey (1956): `stanford.edu/~jugander/rare/`

- Thank you! Questions?