Discussion of Kitagawa and Tetenov

Interactions Workshop
Kei Hirano

September 2016

Statistical Treatment Assignment

Let data be $Z = \{Y_i, X_i, i = 1, \ldots, n\} \sim P \in \mathcal{P}$.

A statistical decision rule maps data into actions:

$$\delta : \text{supp}(Z) \to \mathcal{A}.$$

Here

$$\mathcal{A} = \mathcal{G} = \text{a collection of subsets of supp}(X),$$

where $G \in \mathcal{G}$ interpreted as the set of X values to treat.

Welfare:

$$\begin{aligned} W(G) &= E_P\left[Y_1 1(X \in G) + Y_0 1(X \notin G)\right] \\ &= E_P\left[\frac{YD}{e(X)} 1(X \in G) + \frac{Y(1-D)}{1-e(X)} 1(X \notin G)\right]. \end{aligned}$$

(assuming unconfoundedness)

Best rule in $\mathcal{G}$:

$$G^* \in \arg\max_{G \in \mathcal{G}} W(G).$$

Proposed rule:

$$\hat{G} \in \arg\max_{G \in \mathcal{G}} \hat{W}(G)$$

where $\hat{W}(\cdot)$ is a suitable empirical analog.

Sensible and has the potential to be useful in practice.

A key step is to restrict $\mathcal{G}$ so that it is small in the sense of finite VC dimension.

- Develop concentration inequalities to get finite-sample bounds on loss in welfare relative to $G*$
- ...which yield minmax convergence rates.

In applications we often want to, or are required to, restrict $\mathcal{G}$, so this serves a clever dual purpose.

Can we say more beyond rates?

Comparison with Stoye (2009)

- Finite sample minmax regret with $\mathcal{G}$ unrestricted.
- Achieved by stratifying completely on X, even when X takes on many values (“no-data rules”)
- Intuition: conditional means can be arbitrarily wiggly, so no gain from using nearby x values.
- Here we also allow wigglyness, *but* the VC condition on $\mathcal{G}$ restricts us to “smooth” rules.
- A type of regularization.

Choice of $\mathcal{G}$

- Suppose there are no (or limited) external constraints on $\mathcal{G}$.
- In some applications this might be the case. (E-commerce, perhaps.)
- How to choose $\mathcal{G}$ (or sequence $\mathcal{G}_n$)?
- Data-driven choice possible?
- A related problem: suppose we have a constraint on the VC-dimension of $\mathcal{G}$, but not on the specific class $\mathcal{G}$.

Hybrid Rules

Involve estimation of propensity score $e(x)$ or conditional treatment effect $\tau(x)$.

Nonparametric estimation of $e(x)$ or $\tau(x)$ leads to slower rate for *upper bound* of welfare regret.
(Not clear for lower bound.)

Recall for estimation of ATE under unconfoundedness:

- Not knowing $e(x)$ does not change the efficiency bound. (See Hahn.)
- IPW estimator with $\hat{e}(x)$ is efficient but IPW with true $e(x)$ is inefficient. (HIR)

What's different here? Could this be explored in simple cases and/or simulations?

Margin Assumption

Borrowed from the classification literature, essentially limits the fraction of population with close to zero treatment effect.

Allows faster rates of convergence.

But seems artificial, and against the spirit of the minmax arguments.

I think the results without the margin assumption are more important.