Balancing Out Regression Error

Efficient Treatment Effect Estimation without Smooth Propensities

David A. Hirshberg [1] Stefan Wager [2]

January 2018

[1] Department of Statistics, Columbia University

[2] Stanford Graduate School of Business

Average Treament Effects and Notation

- Observe (Y_i, T_i, X_i) iid
  - $Y_i = Y_i(T_i)$ is the observed outcome under treatment T_i
  - T_i is a binary treatment indicator
  - X_i is a covariate
- Nonparametric model for the potential outcomes

$$Y_i(t) = \underbrace{m_t(X_i)}_{\text{outcome model for treatment } t} + \underbrace{\varepsilon_i(t)}_{\substack{\text{unconfounded mean-zero variation} \\ \text{a.k.a. noise}}}$$

- Estimand: Effect of treatment averaged over the whole sample

$$\bar{\tau} = \underbrace{\frac{1}{n}\sum_{i=1}^{n} m_1(X_i)}_{\text{average treatment outcome } \bar{\mu}_1} - \underbrace{\frac{1}{n}\sum_{i=1}^{n} m_0(X_i)}_{\text{average control outcome } \bar{\mu}_0}$$

- For Today: Estimating $\bar{\mu}_1$

Augmented Inverse Probability Weighting: How it works

$$\hat{\mu}_1 = \underbrace{\frac{1}{n}\sum_{i=1}^{n}\hat{m}_1(X_i)}_{\text{averaged predictions}} - \underbrace{\frac{1}{n}\sum_{i:T_i=1}\gamma_i\left(\hat{m}_1(X_i) - Y_i\right)}_{\text{error estimate}}$$

- Start with a regression estimator
  - Fit a nonparametric model for the outcome under treatment
  - Average its predictions over the complete sample
  - Our error is the bias of these predictions, averaged
    $error = \frac{1}{n}\sum_{i=1}^{n}(\hat{m}_1 - m_1)(X_i)$
- Subtract an estimate of our error
  - This estimate is a weighted average of the regression residuals
  - because residuals are noisy measurements of prediction bias.
  - We only have residuals on the treatment subsample
  - so we weight so it's like an average over the complete sample.
    $\widehat{error} = \frac{1}{n}\sum_{i:T_i=1}\gamma_i(\hat{m}_1 - m_1)(X_i) - \frac{1}{n}\sum_{i:T_i=1}\gamma_i\varepsilon_i$
- The error of our corrected estimator is $error - \widehat{error}$

Augmented Inverse Probability Weighting: Weighting

- The error of the AIPW estimator arises from 'imbalance' between
  - Our target population, the whole sample
  - Our weighted treatment group

  in the unobservable regression error function $\xi_n = \hat{m}_1 - m_1$

$$\hat{\mu}_1 - \bar{\mu}_1 = \underbrace{\frac{1}{n}\sum_{i=1}^{n}\xi_n(X_i) - \frac{1}{n}\sum_{i:T_i=1}\gamma_i\xi_n(X_i)}_{\text{'imbalance' } I_{\xi_n}(\gamma)} + \underbrace{\frac{1}{n}\sum_{i:T_i=1}\gamma_i\varepsilon_i}_{\text{'noise'}}$$

- If the weighted treatment sample is just like our target population, our regression error ξ_n gets averaged out to nothing.
- This is too much to hope for.
- We can't define such weights because we don't know much about ξ_n.
- But ensuring that this imbalance is small will be our primary focus.
- And that's fine. However we weight, the noise term is small.
  - it's an average of mean-zero, conditionally independent terms $noise \sim 1/\sqrt{n}$.

The Gold Standard: True Inverse Propensity Weights

- The inverse propensity weights are the unique weights that balance *any* function ξ in mean

$$\mathbb{E}\, I_\xi(1/e) = 0$$

- The imbalance is not just mean zero; it is small with high probability

$$I_\xi(1/e) \sim \frac{\|\xi\|}{\sqrt{n}}$$

- If our regression error $\xi_n \to 0$,
  our imbalance in ξ_n is neglible relative to noise

$$\frac{I_{\xi_n}(1/e)}{noise} \sim \|\xi_n\|$$

- Therefore:
  - This estimator is asymptotically unbiased.
  - Its MSE is asymptotically optimal.
- We can't hope for better.
- We'll imitate its behavior as well as we can.

Imitation by Estimation: Estimated Inverse Propensity Weights

- In observational studies, we don't know the propensity score.
- We can use an estimate: $\hat{\gamma}_i = 1/\hat{e}(X_i)$
- Expand imbalance around the imbalance with the true IPW

$$I_{\xi_n}(1/\hat{e}) - I_{\xi_n}(1/e) = \frac{1}{n} \sum_{i:T_i=1} \left(\frac{1}{\hat{e}(X_i)} - \frac{1}{e(X_i)} \right) \xi_n(X_i)$$

- Our estimator imitates the gold standard if this perturbation is small.
- Well-known sufficient condition via Cauchy-Schwarz

$$\|\xi_n\| \|1/\hat{e} - 1/e\| \ll 1/\sqrt{n}$$

Calibration of Estimated Inverse Propensity Weights

$$\|\xi_n\|\|1/\hat{e} - 1/e\| \ll 1/\sqrt{n}$$

- It's important to think of PS estimation errors on the inverse scale.
- Errors estimating e or $logit(e)$ blow up when mapped to this scale.
- Rule of Thumb: Estimate $e(X_i)$ with error less than $e(X_i)^2$.

$$\frac{1}{\hat{e}(X)} - \frac{1}{e(X)} = \frac{e(X) - \hat{e}(X)}{\hat{e}(X)e(X)}$$

- This can be a lot to ask for.
- It may not be possible to estimate the propensity score this well.
-
- We will approach the problem from a different direction.
- The resulting estimator will be almost completely insensitive to the difficulty of estimating the propensity score.
- We'll need to exploit more of our knowledge about ξ_n.

What do we know about ξ_n?

- Suppose that m is smooth, i.e. bounded partials up to order k
  - Use a smooth estimator $\hat{m}$, e.g. via locally weighted regression
  - Then we know two things about ξ_n.
    - it's smooth
    - it converges at some rate
  - i.e. ξ_n is, up to scale, in a set of smooth functions convergent at that rate

$$\xi_n / \underbrace{\|\xi_n\|_{\mathcal{F}_n}}_{\text{scale}} \in \mathcal{F}_n$$

- Smoothness is just one possible assumption.
- What we need is a condition like this where
  - the scale of ξ_n is bounded whp
  - the set $\mathcal{F}_n$ isn't too complex
- We could assume, if we preferred:
  - m is approximately sparse in some basis
  - m has bounded variation

How to Sidestep PS Estimation

- To balance the regression error ξ_n, balance the set $\mathcal{F}_n$ *uniformly*
  - Define the maximal imbalance over this set

$$I_{\mathcal{F}_n}(\gamma) := \max_{\xi \in \mathcal{F}_n} I_\xi(\gamma).$$

  - Conditional on $\{X_i, T_i\}_{i=1}^n$, the worst case MSE satisfies

$$\frac{1}{2} MSE \leq I_{\mathcal{F}_n}(\gamma)^2 \mathbb{E}\left[\|\xi_n\|_{\mathcal{F}_n}^2 \mid X, T\right] + \frac{1}{n^2} \sum_{i:T_i=1} \gamma_i^2 \mathrm{var}\left[\varepsilon_i(1) \mid X_i\right]$$

  - Minimize assuming the ratio of tuning parameters is the constant σ.

$$\hat{\gamma} := \arg\min_\gamma \ell(\gamma), \ \ell(\gamma) := I_{\mathcal{F}_n}(\gamma)^2 + \frac{\sigma^2}{n^2} \|\gamma\|^2$$

- Remarks
  1. This optimization problem is solvable with fast off-the-shelf software
  2. Our assumption on the tuning parameter ratio is just for motivation. We study its behavior for arbitrary scale and heterogeneous variance.

These Weights are Estimated Inverse Propensity Weights

- Our weights are determined $[\hat{\gamma}_i = \hat{g}(X_i)]$ by a penalized least squares estimate of the *inverse* propensity score

$$\frac{1}{n}\sum_{i:T_i=1}\left[g(X_i) - \frac{1}{e(X_i)}\right]^2 - \frac{1}{n}\sum_{i=1}^{n} U_i\left[g(X_i) - \frac{1}{e(X_i)}\right] + \frac{\|g\|_{\mathcal{F}_n}^2}{n}$$

  with
  - bounded mean-zero noise $U_i = 1 - \frac{T_i}{e(X_i)}$
  - a penalty on the scale $\|g\|_{\mathcal{F}_n}$
- Nice Properties
  - We estimate the PS on the inverse scale – the same way we use it
    - no error-inflating transformations!
  - Our penalty focuses us on balancing the functions we need to, e.g.
    - a smooth estimate of a [nonsmooth] inverse PS will balance a smooth function ξ
  - $\hat{g}$ is *universally consistent*
    - no assumptions on the PS besides overlap

This Estimator Imitates the Uniform Balance of the True IPW

- Our weights $\hat{\gamma}$ minimize the function ℓ

$$\ell(\gamma) := I_{\mathcal{F}_n}(\gamma)^2 + \|\gamma\|^2/n^2.$$

- Compare the value of ℓ at our weights and the true IPW

$$\sqrt{n} I_{\mathcal{F}_n}(\hat{\gamma}) \leq \sqrt{n} I_{\mathcal{F}_n}(1/e) + \sigma \underbrace{\sqrt{\left| \frac{1}{n} \left(\|1/e\|^2 - \|\hat{\gamma}\|^2 \right) \right|}}_{o_p(1)}.$$

- Because our weights consistently estimate the true IPW, they balance $\mathcal{F}_n$ asymptotically as well as the true IPW.
- Nice Consequence
  - the imbalance in the regression error ξ_n is asymptotically negligible
  - therefore our estimator is asymptotically optimal

  if
  - The maximal imbalance on $\mathcal{F}_n$ with the true IPW is $o_p(1/\sqrt{n})$.
  - The scale $\|\xi_n\|_{\mathcal{F}_n}$ is bounded

More Formally

Theorem

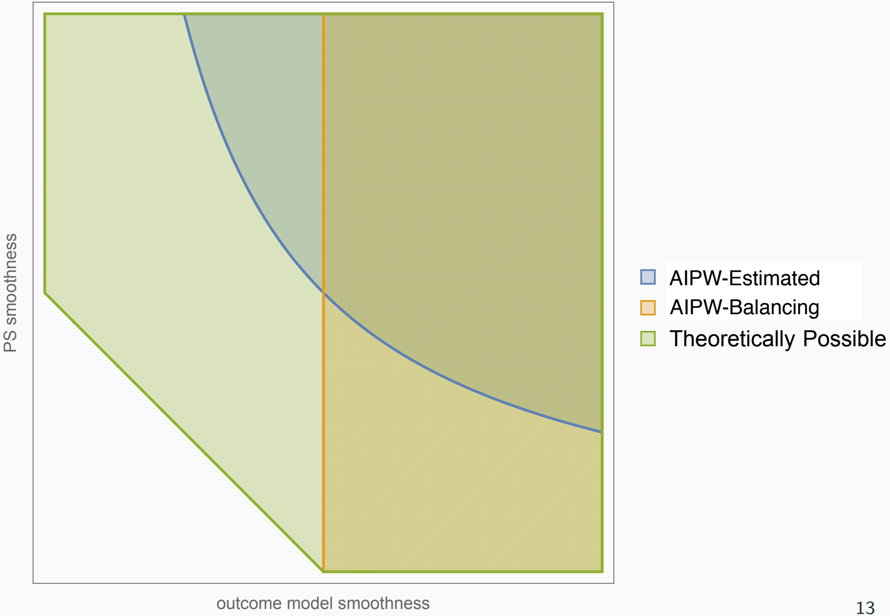
The AIPW estimator $\hat{\mu}_1$ *with*

$$\hat{\gamma} := \arg\min_{\gamma} I_{\mathcal{F}_n}(\gamma)^2 + \|\gamma\|^2/n^2$$

is asymptotically normal with optimal variance if

- *The propensity score is bounded away from zero.*
- *Our noise has a third conditional moment*
- $\mathcal{F}_n$ *is convex and symmetric*
- *The scale of* ξ_n *relative to* $\mathcal{F}_n$ *is* $O_p(1)$
- *The Rademacher complexity of* $\mathcal{F}_n$ *is* $o_p(1/\sqrt{n})$
- *The sequence* $\mathcal{F}_n$ *is dense in the space of square integrable functions*

Range of Asymptotic Efficiency



Simulation Results

	root-mean squared error				coverage			
	BART	AIPW	TMLE	Ours	BART	AIPW	TMLE	Ours
setup 1	0.82	0.18	0.18	**0.17**	0.00	0.88	0.92	0.93
	0.76	0.15	0.15	**0.14**	0.00	0.86	0.88	0.90
	0.99	0.25	0.25	**0.24**	0.00	0.86	0.90	0.90
	0.40	0.12	0.12	**0.09**	0.07	0.92	0.94	0.94
	0.40	0.11	0.11	**0.10**	0.01	0.90	0.93	0.94
	0.65	0.16	0.16	**0.13**	0.01	0.88	0.88	0.94
setup 2	**0.08**	**0.08**	**0.08**	**0.08**	0.92	0.92	0.92	0.93
	0.08	**0.07**	**0.07**	0.08	0.96	0.96	0.98	0.96
	0.07	**0.07**	**0.07**	**0.07**	0.96	0.96	0.97	0.96
	0.07	0.08	0.08	**0.07**	0.96	0.96	0.99	0.98
	0.08	**0.07**	**0.07**	**0.07**	0.94	0.94	0.97	0.96
	0.07	0.08	0.08	0.08	0.98	0.94	0.96	0.96

Simulation Results

	root-mean squared error				coverage			
	BART	AIPW	TMLE	Ours	BART	AIPW	TMLE	Ours
setup 3	0.38	0.72	0.65	**0.31**	0.54	0.82	0.80	0.86
	0.33	0.65	0.61	**0.18**	0.60	0.57	0.56	0.96
	0.40	0.67	0.61	**0.29**	0.48	0.84	0.83	0.85
	0.32	0.61	0.55	**0.18**	0.28	0.64	0.55	0.92
	0.27	0.61	0.57	**0.10**	0.38	0.29	0.22	0.98
	0.41	0.63	0.56	**0.21**	0.12	0.64	0.57	0.87
setup 4	0.30	0.16	0.16	**0.10**	0.08	0.67	0.65	0.90
	0.19	0.15	0.15	**0.07**	0.31	0.64	0.60	0.96
	1.01	0.29	0.30	**0.21**	0.00	0.22	0.16	0.44
	0.37	0.18	0.18	**0.11**	0.02	0.59	0.58	0.85
	0.23	0.17	0.17	**0.08**	0.16	0.56	0.55	0.94
	1.02	0.36	0.35	**0.28**	0.00	0.04	0.06	0.20

Variations: Past and Future

1. Athey et al. [2016] studied this estimator with High-Dimensional Linear Outcome Models. We've refined their argument, which we hope will enable sharper characterization in that that setting.
2. Kallus [2016] studied these weights in the context of linear estimators, i.e. without regression. He established a rate using a simplified version of our argument. Proving efficiency will require some new arguments, which we are working on.
3. Our argument can work with balanced sets $\mathcal{F}_n$ that depend on the complete data $\{X_i, Y_i, T_i\}_{i=1}^{n}$. Wide Open: How should we base our balanced set $\mathcal{F}_n$ on a selected model?

Summary

- The AIPW with Uniform Balancing Weights can compete with top-performing estimators like the TMLE.
- Its insensitivity to the complexity of the PS can be a big advantage in some problems.
- The essential reason for this difference is that our balancing approach imitates the balance of the true IPW in a coarser way.
    - With EIPW, we try to imitate the imbalance of the true IPW for all functions
    - With UBW, we try to imitate the maximum of this imbalance over some set. This is easier.
- Paper on Arxiv, Software coming soon.

References

Susan Athey, Guido W Imbens, and Stefan Wager. Approximate residual balancing: De-biased inference of average treatment effects in high dimensions. *arXiv preprint arXiv:1604.07125*, 2016.

Nathan Kallus. Generalized optimal matching methods for causal inference. *arXiv preprint arXiv:1612.08321*, 2016.