Balance in Causal Inference

from Poststratification to Regularized Riesz Representers

David A. Hirshberg, Stanford GSB

Online Causal Inference Seminar.   September 1, 2020

based on Augmented Minimax Linear Estimation with Stefan Wager

Estimating the Treatment-Specific Mean

We observe iid units described by

- X_i, a vector of covariates
- W_i, a level of treatment
- $Y_i = Y_i(W_i)$, an outcome under that level of treatment

When treatment assignment is unconfounded, the treatment-specific mean is an average involving the regression function of Y on X, W.

$$\mathbb{E}[Y_i(1)] = \mathbb{E}[m(X_i, 1)] \text{ where } m(X_i, W_i) = \mathbb{E}[Y_i \mid X_i, W_i].$$

It's the average of its **treated arm value** $m(X_i, 1)$.

This arm isn't the one we actually observe for all units.
What we observe is the **natural arm value**, $m(X_i, W_i)$, plus 'noise'.

$$Y_i = m(X_i, W_i) + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i \mid X_i, W_i] = 0.$$

Basic Estimation Approaches: Imputation

To estimate $\psi(m) = \mathbb{E}[m(X, 1)]$:

1. Find an estimator for the regression function m.
2. Average its treated arm value $\hat{m}(X_i, 1)$ over the sample.

$$\hat{\psi}_{imp} = n^{-1} \sum_{i \leq n} \hat{m}(X_i, 1)$$

Basic Estimation Approaches: Weighting

To estimate $\psi(m) = \mathbb{E}[m(X,1)]$:

1. Find weights $\hat{\gamma}_\psi$ that approximately equate
    - averages of the treated arm value of m
    - weighted averages of the natural arm value of m

$$n^{-1}\sum m(X_i,1) \approx n^{-1}\sum \hat{\gamma}_\psi(X_i,W_i)m(X_i,W_i).$$

2. Take a weighted average of the observed outcomes Y_i, which are the natural arm value (plus noise).

$$\hat{\psi}_{ipw} = n^{-1}\sum \hat{\gamma}_\psi(X_i,W_i)Y_i$$

Inverse propensity weights will do. For each square integrable function m,

$$\mathbb{E}[m(X,1)] = \mathbb{E}[\gamma_\psi(X,W)m(X,W)] \text{ for } \gamma_\psi(X,W) = \frac{1(W=1)}{P(W=1 \mid X)}.$$

Augmented Inverse Propensity Weighting (AIPW)

- Imputation works if we estimate the regression function well.
- Weighting works if we estimate the inverse propensity weights well.
- Combining both yields a more robust estimator: AIPW (Robins, Rotnitzky, and Zhao, 1995)

We estimate the error of the imputation estimator
by weighting our regression residuals — and subtract it off.

$$\hat{\psi}_{aipw} = \underset{\text{imputation estimator}}{n^{-1}\sum \hat{m}(X_i, 1)} - \underset{\text{estimated imputation error}}{n^{-1}\sum \hat{\gamma}_{\psi}(X_i, W_i)(\hat{m}(X_i, W_i) - Y_i)}.$$

The weighting works just like before.

- The error of the imputation estimator is the average of the treated arm value of the **regression error function** $\delta_m = \hat{m} - m$.
- The regression residuals are its natural arm value (plus noise).
- Our weights are chosen to approximately equate them.

A Common Worry: Instability of (A)IPW

Inverse propensity weighting has a reputation for **extreme variability** when treatment is rare within some strata (Kang and Schafer, 2007).

In part, this is based on a bad example (Robins et al., 2007). In K & S,

- Treatment is so rare for some that estimation is nearly impossible.
- An imputation estimator miraculously works.

But we should also blame the usual IPW workflow.

1. Estimate the prop. score by regressing a treatment indicator on X.
2. Use its reciprocal to form inverse propensity weights, $\hat{\gamma}_{\psi}(w, x) = 1(w = 1) \;/\; \hat{P}(W = 1 \mid X = x)$.

Inversion dramatically **inflates errors** where the propensity score is small.

$$\frac{1}{\hat{p}} - \frac{1}{p} = \frac{p - \hat{p}}{\hat{p}p}.$$

The Balancing Workflow

$$
\begin{aligned}
&\hat{\psi}_{aipw} - \psi(m) \\
&= n^{-1}\sum \delta_m(X_i,1) - n^{-1}\sum \hat{\gamma}_\psi(X_i,W_i)\delta_m(X_i,W_i) \quad \text{[imbalance]} \\
&+ n^{-1}\sum[\hat{\gamma}_\psi(X_i,W_i)\varepsilon_i + m(X_i,1) - \psi(m)] \quad \text{[mean zero]}
\end{aligned}
$$

Error arises primarily from a difference between

- the average of the treated arm value of δ_m
- the weighted average of the natural arm value of δ_m

We call this the **imbalance** in the function $\delta_m = \hat{m} - m$.

If we think δ_m will be in some set $\mathcal{M}$ — a **model** for δ_m — we can choose weights controlling the **maximal imbalance** in that model.

$$I_{\mathcal{M}}(\hat{\gamma}) = \sup_{f\in\mathcal{M}} |n^{-1}\sum f(X_i,1) - n^{-1}\sum \hat{\gamma}(X_i,W_i)f(X_i,W_i)|.$$

The Minimax Approach to Balancing

Accounting for the mean-zero term as well, we minimize the **maximal mean squared error** of our estimator in the model.

$$\hat{\gamma} = \arg\min_{\gamma} CMSE_{\mathcal{M}}(\gamma) := I^2_{\mathcal{M}}(\gamma) + \frac{\sigma^2}{n} \|\gamma\|^2_{L_2(\mathrm{P_n})}$$

Specifically, maximal **conditional-on-design** mean squared error for:

- regression error functions δ_m in the model $\mathcal{M}$
- variances $\mathrm{Var}\left[Y_i \mid X_i, W_i\right]$ are bounded by σ^2

The Minimax Weights

These weights $\hat{\gamma}$ are a **method of moments** estimate of the inverse propensity weights γ_ψ.

They approximately solve a set of sample moment conditions: **approximate balance conditions** for functions in our model $\mathcal{M}$.

$$\begin{aligned} n^{-1}\sum \hat{\gamma}(X_i, W_i) f(X_i, W_i) &\approx n^{-1}\sum f(X_i, 1) \quad && \forall f \in \mathcal{M}, \\ \mathbb{E}\,\gamma_\psi(X, W) f(X, W) &= \mathbb{E}\, f(X, 1) && \forall f \ : \ \mathbb{E}\, f^2(X, W) < \infty. \end{aligned}$$

The inverse propensity weights uniquely solve analogous **population balance conditions** for all square integrable functions.

This is a direct estimate of the inverse propensity weights; there's **no error-inflating inversion step** in this workflow. We get stable estimates, even in the problematic K & S example. (H, Maleki, and Zubizarreta, 2019)

The Approach with Discrete Covariates

These are **poststratification weights** if our model is completely general.

Maximal imbalance $I_{\mathcal{M}}(\gamma) = \infty$ unless we weight treated units so their empirical covariate distribution **exactly matches** the whole sample's.

$$\begin{aligned} I_{\mathcal{M}}(\gamma) = \sup_{f_{x,w}} n^{-1} \sum_{i,x,w} f_{x,w} \Big[& 1\{(X_i, 1) = (x, w)\} \\ - \quad & 1\{(X_i, W_i) = (x, w)\}\gamma(x, w) \Big] \\ = \infty \text{ unless } \ \gamma(x,w) = 1\{w = 1\} & \frac{\sum_i 1\{X_i = x\}}{\sum_i 1\{(X_i, W_i) = (x, w)\}}. \end{aligned}$$

The Approach with Continuous Covariates

If covariates are **continuous**, we can't do this.
No two units will have exactly the same covariates, so we need there to be a sense in which two empirical distributions are *close enough*.

We can control imbalance only in **models that limit discontinuity**, e.g.,

1. Functions with several $(> \dim(X)/2)$ bounded derivatives.
2. Functions of bounded (Hardy-Krause) variation.

This is just a requirement that we can learn about a function from its samples. We need the same to do any kind of estimation.

Asymptotic Efficiency (H and Wager, 2017)

Choose weights minimizing $CMSE_{\mathcal{M}}$ for a model $\mathcal{M}$ that is convex, uniformly bounded, and admits a uniform CLT.[1]

Our estimator is **asymptotically efficient** if

1. the inverse propensity weighting function γ_{ψ} is square integrable.
2. our regression is mean-square consistent: $\|\delta_m\|_{L_2(\mathrm{P})} = o_p(1)$.
3. $\mathcal{M}$ is a valid model for our regression error: $\delta_m \in \alpha\mathcal{M},\ \alpha = O_p(1)$.

Asymptotic efficiency means that the usual confidence intervals are asymptotically valid with **minimal length**. For example,

- Wald type intervals: $\hat{\psi} \pm 1.96\,\mathrm{se}(\hat{\psi})$.
- Bootstrap intervals.

[1] Our example models, as specified in parentheticals, do satisfy these assumptions.

Practical Considerations: Choosing Parameters

We have three ‘parameters’ to choose.

1. Noise level σ^2.
2. Outcome regression $\hat{m}$.
3. Regression error model $\mathcal{M}$

The good news is that the estimator is **adaptive**: it is sensitive to properties of the regression error δ_m that aren't baked into our model $\mathcal{M}$.

It behaves a bit **like we had perfectly modeled** the regression error's

1. consistency, measured by mean squared error $\left\|\delta_m\right\|_{L_2(P_n)}$.
   (H and Wager, 2017)
2. smoothness, measured by polynomial decay of its fourier coefficients.
   (H, Maleki, and Zubizarreta, 2019, in RKHS models)

And it's not sensitive to σ: $\sigma = 1$ is usually fine in theory and practice.

This means we do pretty well with reasonable default parameters.
This is good; we still have a lot to learn about automated tuning.

Other Estimands

Estimating many other things is a trivial change — often one line of code.

$\psi(m) = \mathbb{E}[m(X,1) - m(X,0)]$ average treatment effect

$\psi(m) = \mathbb{E}[m(X,\pi(X))]$ average policy value

$\psi(m) = \mathbb{E}\left[\frac{\partial}{\partial w} m(X,w) \mid_{w=W}\right]$ average slope

$\psi(m) = \mathbb{E}[h(X,W,m)]$ generally, where $h(x,w,m)$ is linear in m.

Substitute h for 'treated arm value' earlier in the talk and it all works.
(H and Wager, 2017)

The inverse prop. weights γ_ψ generalize to the **Riesz representer** of ψ.
(Chernozhukov, Escanciano, Ichimura, and Newey, 2016)

γ_ψ solves $\mathbb{E}\, h(X,W,f) = \mathbb{E}\, \gamma_\psi(X,W) f(X,W), \ \forall f : \mathbb{E}\, f^2(X,W) < \infty.$

When ψ is **mean-square-continuous**, the Riesz representation theorem guarantees that this solution exists and is square integrable.

Discontinuous Estimands

To estimate discontinuous functionals ψ, like the CATE at a point, we can use this approach to estimate a smoothed approximation.

$$\begin{aligned}\psi(m) &= m(x,1) - m(x,0) \;\; \text{at some point } x \\ &\approx \mathbb{E}[K(x,X)\{m(X,1) - m(X,0)\}] \;/\; \omega(K,x)\end{aligned}$$

Typically we smooth by convolution with a kernel K.
(Chernozhukov, Newey, Robins, and Singh, 2018)

What makes balancing work?

Balancing folks emphasize three principles:

(i) directly balancing the sample, not the population.
(ii) balancing the right model $\mathcal{M}$, not just a convenient one.
(iii) estimating the inverse propensity score, not *its* inverse.

Do we need all three? We could **cross-fit** when estimating the weights.

1. Minimize $CMSE_{\mathcal{M}}$ for half the sample to get an estimate $\hat{\gamma}_{\psi}$ of the inverse propensity weighting function.
2. Evaluate it for units on the other half to get weights.

Is this balancing? This violates (i) but maybe (ii) and (iii) are enough.
Intuition about **own-observation bias** suggests it might work better.
Similar methods do work well in theory and in simulations.
(Chernozhukov, Newey, Robins, and Singh, 2018)

Cross-fitting and directly balancing the sample are at odds,
and we don't know when to prefer one over the other.

Where we're going

Historically, there's been tension between folks who emphasize simple balance properties — like from poststratification — and those who emphasize semiparametric efficiency. But in the last decade, with more general notions of balance, it's become harder to see the boundary between these approaches.

Still, there's a stereotype that *balancing estimators* and *semiparametric estimators* have fundamentally different robustness properties.

balancing	near-minimaxity for the observed design
semiparametric	double robustness

The truth is, these stereotypical properties may just be the ones that are easiest to prove.

There's work to be done in understanding exactly how these properties arise, and whether they are compatible or we must make a choice. This isn't work for one faction or another. It's for all of us.

References

V. Chernozhukov, J. C. Escanciano, H. Ichimura, and W. K. Newey. Locally robust semiparametric estimation. *arXiv preprint arXiv:1608.00033*, 2016.

V. Chernozhukov, W. Newey, J. Robins, and R. Singh. Double/de-biased machine learning of global and local parameters using regularized riesz representers. *arXiv preprint arXiv:1802.08667*, 2018.

D. A. H and S. Wager. Augmented minimax linear estimation. *arXiv preprint arXiv:1712.00038*, 2017.

D. A. H, A. Maleki, and J. Zubizarreta. Minimax linear estimation of the retargeted mean. *arXiv preprint arXiv:1901.10296*, 2019.

J. D. Kang and J. L. Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, pages 523–539, 2007.

J. Robins, A. Rotnitzky, and L. Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(1):106–121, 1995.

J. Robins, M. Sued, Q. Lei-Gomez, and A. Rotnitzky. Comment: Performance of double-robust estimators when “inverse probability” weights are highly variable. *Statistical Science*, 22(4):544–559, 2007.