Thinking Causally with High-Dimensional Databases

Debashis Ghosh

Department of Biostatistics and Informatics, Colorado School of Public Health

Statistical and Computational Challenges in Precision Medicine

Joint work with Wei Luo (CUNY Baruch College), Yeying Zhu (University of Waterloo), Efrén Cruz-Cortés, Kevin Josey and Elizabeth Juaréz-Colunga (Colorado School of Public Health)

Introduction

- Motivation: Electronic Data Warehouses: `https://pub.healthdatacompass.org`
- Interest in performing *in silico* observational studies/comparative effectiveness research
- High-dimensional confounder setting (large n and large p, $n > p$)
- Statistical interest in causal inference/optimal treatment methods

Potential Outcomes: Notation and Assumptions

- Let $T \in \{0, 1\}$ denote the treatment
- Let $\{Y(0), Y(1)\}$ denote the potential outcomes for Y under each of the treatments
- Targets of estimation:

$$\text{ACE} = E[Y_i(1) - Y_i(0)]$$

- Strong Ignorability of Treatment Assignment Assumption (SITA):

$$T \perp \{Y(0), Y(1)\}|\mathbf{X}$$

where $\mathbf{X}$ are covariates

Potential Outcomes (cont'd.)

- Propensity score: $e(\mathbf{X}) = P(T = 1|\mathbf{X})$
- SITA implies:

$$T \perp \{Y(0), Y(1)\}|e(\mathbf{X})$$

- Treatment Positivity: $0 < e(\mathbf{X}) < 1$ for all $\mathbf{X}$
- Estimate propensity score using
  - logistic regression
  - machine learning methods

Causal Inference as Missing Data Problem

- Data visualization

$Y(0)$	$Y(1)$	T	X_1	$\cdots$	X_p
?	y_1	1	x_{11}	$\cdots$	x_{1p}
y_2	?	0	x_{21}	$\cdots$	x_{2p}
?	y_3	1	x_{31}	$\cdots$	x_{3p}
?	y_4	1	x_{41}	$\cdots$	x_{4p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_n	?	0	x_{n1}	$\cdots$	x_{np}

Optimal Treatment as Missing Data Problem

- Let $\mathcal{T}$ denote the set of potential of treatments
- "Impute" $\{Y_i(t) : t \in \mathcal{T}\}$, $i = 1, \ldots, n$
- Select "optimal" treatments based on a utility/value function
- Much of this talk focuses on causal inference, although ideas should transfer to optimal treatment problem.

Today's talk

- The forgotten assumption for causal inference in high dimensions
- Boosting and covariate balance: two sides of the same coin

Treatment Positivity

1. Treatment Positivity:

$$0 < P(T = 1|\mathbf{X}) < 1$$

for all $\mathbf{X}$

2. Note that for regular semiparametric estimators of causal effects, we strengthen this to

$$\eta < P(T = 1|\mathbf{X}) < 1 - \eta$$

uniformly in $\mathbf{X}$ where $\eta \in (0, 0.5)$ (strict positivity)

Robins and Ritov, SIM, 1997

1. Strict positivity is a necessary assumption for causal effect estimators to behave with usual asymptotic properties.
2. Without strict positivity, Robins and Ritov argue that this allows for 'pathological data generating distributions' that yield causal effect estimators with irregular asymptotic properties.

Example of Irregular Causal Effect Estimator

- Simulations from Luo et al., Biometrika, 2017

Table 2. *Compare estimators of average causal effect*

Models	GCL	GCQ	GAM	RF	IPW	Matching	DR	DR-oracle	TMLE	a-SORCI
I	.096	−.139	−.035	−.140	−.083	−.080	−.132	−.004	−.035	.004
	(.105)	(.158)	(.086)	(.069)	(.218)	(.085)	(.278)	(.089)	(.208)	(.049)
II	−.007	.008	.000	.009	.004	−.043	.000	−.007	−.016	−.008
	(.113)	(.120)	(.126)	(.099)	(.163)	(.093)	(.129)	(.090)	(.139)	(.068)
III	−.005	.008	.005	.020	.043	−.016	−.013	−.008	−.020	.005
	(.123)	(.073)	(.094)	(.096)	(.181)	(.132)	(.152)	(.094)	(.169)	(.064)
IV	−.088	.010	−.007	−.019	−.004	−.090	.000	.001	−.052	.001
	(.346)	(.068)	(.073)	(.150)	(.318)	(.248)	(.070)	(.068)	(.217)	(.072)
V	−.001	.000	.001	.003	.021	−.001	.000	−.001	−.001	.000
	(.022)	(.022)	(.025)	(.057)	(.088)	(.081)	(.022)	(.021)	(.023)	(.023)

Why?

- Under assumptions of Theorem 3 of Luo et al. (2017), violation of strict positivity and dimension reduction functions as a *sparsity*-inducing assumption
- Thus, we can beat doubly-robust estimators for certain classes of models
- Tradeoff: our estimator can behave erratically under other model configurations

Strict Positivity

- Bayes rule says that strict positivity is equivalent to

$$\eta < \frac{\alpha f(\mathbf{X}|T=1)}{(1-\alpha) f(\mathbf{X}|T=0)} < 1-\eta$$

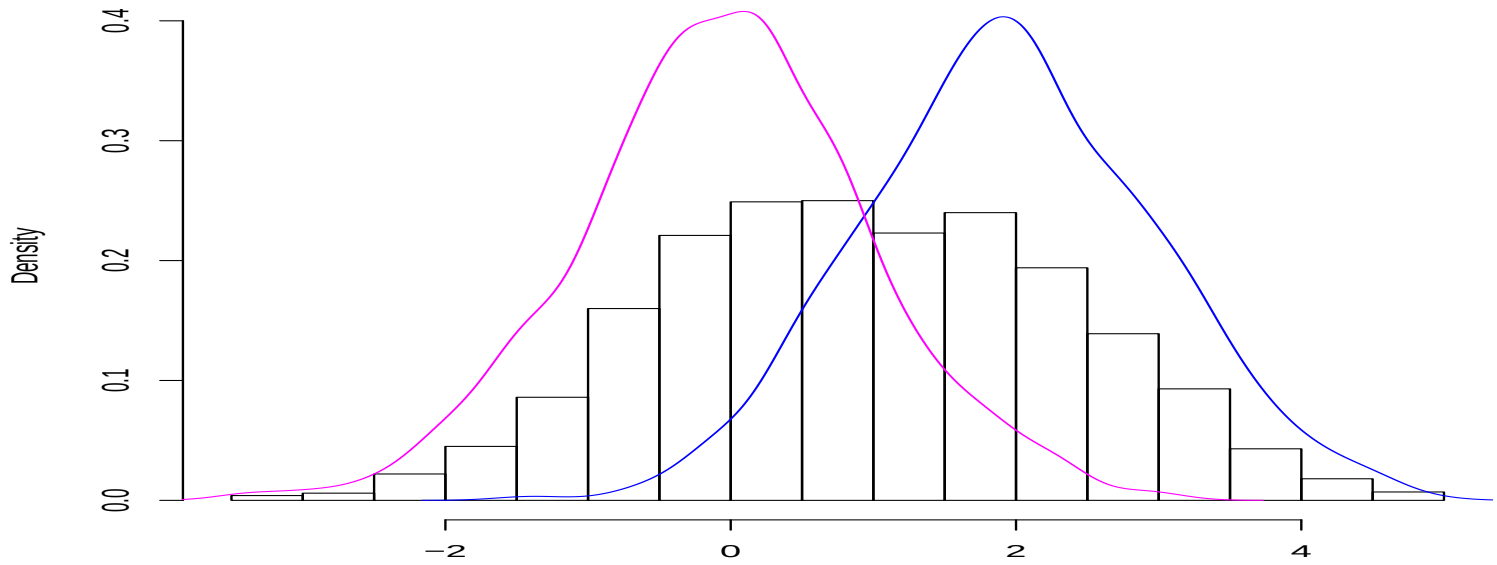
  where $\alpha = P(T=1)$, and $f(\mathbf{X}|T=t)$ is distribution of confounders given treatment.
- This is a bounded likelihood ratio assumption.
- Implication: classifiers that can predict treatment will eventually violate this assumption for sufficiently large sample size.

Strict Positivity in High Dimensions

- Assume we can model $\mathbf{X}(t)|T = c,\ c = 0, 1$ using Gaussian Processes to generalize densities from previous slide
- Then we have that
  (a). strict positivity $\Leftrightarrow$ Gaussian measures given treatment are equivalent
  (b). There is a dichotomy: strict positivity violated $\Leftrightarrow$ Gaussian measures given treatment are orthogonal

Strict Positivity (cont'd.)

- In high dimensions, for regular causal effect estimators, we assume that the support for confounders have full overlap.



- One possible solution to relax this: use the margin to define the points for estimating causal effects (Ghosh 2018, in press).

Covariate Balance

- A lot of recent research in causal inference
- Idea: reduce sensitivity of causal effect estimators to the misspecification of propensity score by treating covariate balance as an "optimization criterion"

Covariate Balance proposals

- Generalized methods of moments (CBPS, Imai and Ratkovic, 2014)
- Calibration estimators (ATE, Chan et al., 2016)
- Boosting propensity scores (TWANG, McCaffrey et al., 2004)
- Entropy balancing (Ebal, Hainmueller, 2012)
- Practical question: which one to use??
- Extensive comparison of TWANG with CBPS in Setodji et al. (2017, Epidemiology)

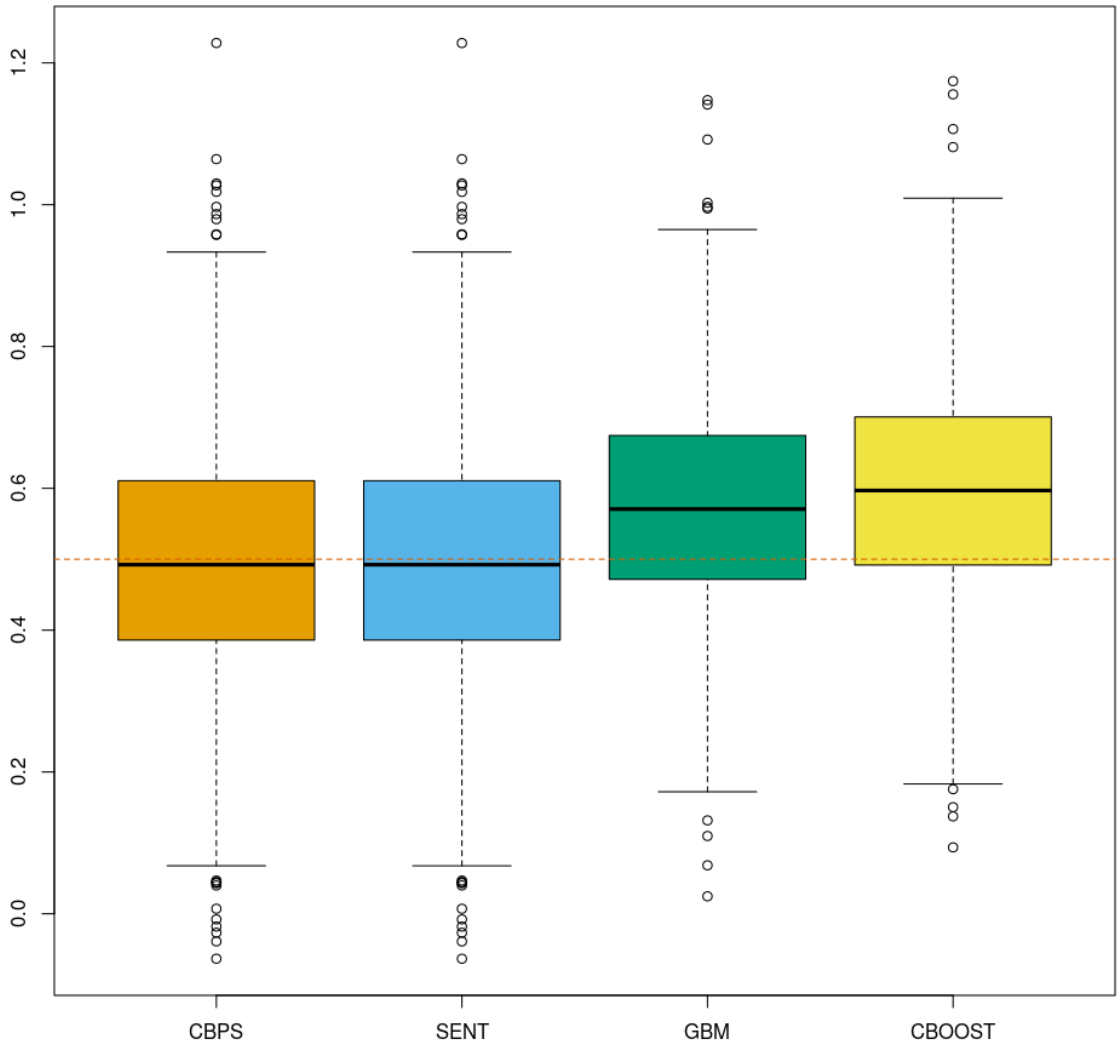
Covariate Balance proposals: a unified approach

- Idea: estimate person-specific weights p_i for causal effects based on minimizing the **Bregman divergence**, defined as

$$D_F(\mathbf{p} \parallel \mathbf{q}) = F(\mathbf{p}) - F(\mathbf{q}) - (\mathbf{p} - \mathbf{q})^\intercal \nabla F(\mathbf{q}), \tag{1}$$

  where F is a function and $\mathbf{q}$ are the base weights.
- For specific functional form of F, can get back CBPS and entropy balancing
- Alternatively, can perform boosting directly on the Bregman divergence to estimate weights directly in contrast to TWANG

Example simulation setting from Lee et al. (2010)



Conclusion

- In high dimensions, there is an interplay between covariate overlap, treatment positivity and regular estimation for causal effects that is not appreciated and needs to be better understood. This will have implications for propensity score modelling.
- Many of the methods for covariate balance are roughly doing the same thing, i.e., constructing weights that have properties of calibration estimators.

Funding and References

- Grohne-Stepp Endowment from Univ. Colorado Cancer Center
- Data Science Patient Value (D2V) Pilot Award, Univ. Colorado

- Luo, W., Zhu, Y. and Ghosh, D. (2017). On estimating regression causal effects using sufficient dimension reduction, *Biometrika* **104**, 51 – 65.
- Ghosh, D. (2018). Relaxed covariate overlap and margin-based causal effect estimation, *Statistics in Medicine*, in press. Available at `https://arxiv.org/abs/1801.00816`.
- Ghosh, D. and Cortés, E. C. (2018). A note on strict functional overlap in causal inference problems with high-dimensional covariates, submitted. Available at `http://arxiv.org/abs/1801.03185`
- Josey, K. P., Juarez-Colunga, E. and Ghosh, D. (2018). A unified approach to covariate balance using Bregman divergences, in preparation.