Stat 236 Final Project: High-Dimensional Distributed Learning for Causal Inference (Group 3)

Larry Han, Christina Howe, Lara Maleyeff, Avi Swartz

Dec 12, 2020

Introduction

- The ability to integrate information across multiple sites can improve generalizability and accelerate decision-making
- COVID-19 → need for novel approaches to efficiently and safely analyze data across different healthcare systems.
- Motivation: International Consortium of EHR for COVID-19 (4CE), a joint effort of hundreds of hospitals across 50 sites and 17 countries to inform physicians, epidemiologists, and the policymakers (Brat et al., 2020)

Why is this an interesting problem?

- Integrated analysis using EHRs is challenging because of
  1. privacy provisions that make it impossible to share patient-level data across sites
  2. communication costs of time and human resources associated with transferring even summary-level data between sites
  3. the dimension of observed covariates may be very large and heterogeneous across sites

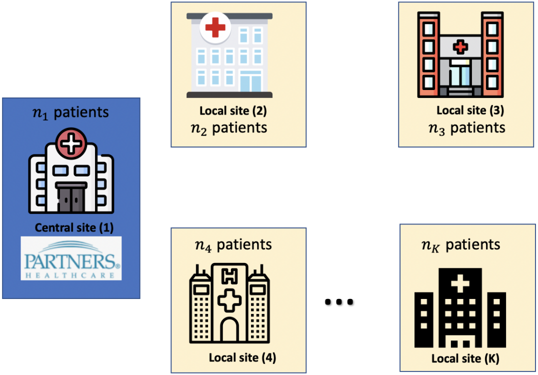
What have others already done?

- Previously, distributed algorithms were developed in the low-dimensional regression setting by decomposing tasks to be completed within each site:
    - Linear regression (Chen et al., 2006)
    - Logistic regression (Duan et al., 2020a, Wu et al., 2012)
    - Cox regression (Duan et al., 2020b, Lu et al., 2015)
- Recently, much research has focused on the high-dimensional regression by imposing sparsity assumptions (Battey et al., 2018, Lee et al., 2017) and constructing a surrogate likelihood function as approximation of the global likelihood function (Jordan et al., 2018)
- Little development of distributed learning algorithms for causal models, either in the low-dimensional or high-dimensional settings.

High-Level Goal

- We aim to address this knowledge gap by proposing two novel approaches.
  1. A two-step procedure that first fits a penalized regression at the central site to obtain the covariates to be used in the second step, where we robustly estimate target model parameters at local sites using only site-specific patient data and summary level-information from other sites
  2. A gradient based optimization approach that leverages the software infrastructure developed for neural network inference to move the estimators to the data.

Visualizing the Problem Setting and Notation



- K sites
- n_r patients at site $r; r = 1, \dots, K$
- $n = \sum_{r=1}^{K} n_r$ total patients
- $D_i^r = (Y_i^r, X_i^{r^T}, A_i), i = 1, \dots, n_r$
- Patient-level data is accessible at each site only
- Summary-level data can be shared between sites
- The central site aggregates summary data from local sites

Figure: One central site (blue) and $K - 1$ local sites (yellow)

Approach 1 (Overview of Two Steps)

1. **[Variable Selection]:** For the central site only, we fit a penalized linear regression with Y as the continuous outcome and X as the set of covariates available in the EHR. Penalized linear regression minimizes a penalized sum of squares of the form

$$\frac{1}{n}||Y - \mathrm{X}\beta||_2^2 + \sum_{j=1}^{p} \rho_\lambda(\beta_k),$$

   with respect to β. We consider three common penalty functions for ρ: LASSO, SCAD, and MCP. With cross-validated optimal λ, we extract $\hat{\mathrm{p}}_r^* \subseteq \{1, \ldots, p\}$, the indices of the covariates with a non-null relationship with Y.

2. **[Modeling]:** Estimation outcome regression (OR) model, propensity score (PS) model, and density ratio model (DR). Construct estimators of TATE.

Goal

- Our goal is to estimate an ATE for a specified target population, $\mathcal{T}$ (TATE)

$$\Delta_{\mathcal{T}} = \mu_{1,\mathcal{T}} - \mu_{0,\mathcal{T}}, \quad \text{where } \mu_{a,\mathcal{T}} = E_{\mathrm{X}_{\mathcal{T}}}\{Y^{(a)} \mid R \in \mathcal{T}\}.$$

- Note that the TATE across all sites is simply the ATE, which is the special case when $\mathcal{T} = \{1, \ldots, K\}$.
- To robustly estimate TATE, we propose the estimation of three models:
  - $m_{a,r}(\mathrm{X}) = E(Y \mid A = a, \mathrm{X} = \mathrm{X}, R = r)$ (Outcome regression)
  - $\pi_{a,r}(\mathrm{X}) = P(A = a \mid \mathrm{X} = \mathrm{X}, R = r)$ (Propensity score)
  - $w_r(\mathrm{X}) = \frac{f(\mathrm{X} \mid R = r)}{f(\mathrm{X} \mid R \in \mathcal{T})}$ (Density ratio)

Why is passing $\bar{X}^1$ from central site to local sites sufficient?

- The density ratio weight can be estimated by imposing a semiparametric model, e.g. exponential tilt model (Qin, 1998):

$$f(\mathrm{X} \mid R = r) = f(\mathrm{X} \mid R = 1) \exp(\gamma_r^\top \mathrm{X}),$$

- Then

$$\bar{\mathrm{X}}^1 = \int \mathrm{X} f(\mathrm{X} \mid R = 1) d\mathrm{X} = \int \mathrm{X} f(\mathrm{X} \mid R = r) e^{\gamma_r^\top \mathrm{X}} d\mathrm{X}.$$

- Estimation details for each of the models is given in the paper.

Proposed OR and IPW Estimators

$$\boxed{\widehat{\mathrm{OR}}_r = \frac{1}{n_{1,r}} \sum_{i=1}^{n_{1,r}} \hat{m}_{1,r}(\mathrm{X}_i^r) I(A_i^r = 1) - \frac{1}{n_{0,r}} \sum_{i=1}^{n_{0,r}} \hat{m}_{0,r}(\mathrm{X}_i^r) I(A_i^r = 0)}$$

$$\boxed{\widehat{\mathrm{IPW}}_r = \frac{1}{n_r} \sum_{i=1}^{n_r} \left(\frac{Y_i^r I(A_i^r = 1)}{\hat{\pi}_{1,r}(\mathrm{X}_i^r)} - \frac{Y_i^r I(A_i^r = 0)}{\hat{\pi}_{0,r}(\mathrm{X}_i^r)} \right)}$$

Proposed DR and Weighted DR Estimators

$$\boxed{\widehat{\mathrm{DR}}_r = \hat{\mu}_{r,1} - \hat{\mu}_{r,0}, \quad \text{where}}$$

$$\hat{\mu}_{r,a} = n_r^{-1} \sum_{i=1}^{n_r} \left\{ \frac{Y_i^r I(A_i^r = a)}{\hat{\pi}_{a,r}(\mathrm{X}_i^r)} - \frac{I(A_i^r = a) - \hat{\pi}_{a,r}(\mathrm{X}_i^r)}{\hat{\pi}_{a,r}(\mathrm{X}_i^r)} \hat{m}_{a,r}(\mathrm{X}_i^r) \right\}$$

$$\boxed{\widehat{\mathrm{WDR}}_r = \tilde{\mu}_{r,1} - \tilde{\mu}_{r,0}, \quad \text{where}}$$

$$\tilde{\mu}_{r,a} = n_r^{-1} \sum_{i=1}^{n_r} w_r(\mathrm{X}_i^r) \left\{ \frac{Y_i^r I(A_i^r = a)}{\hat{\pi}_{a,r}(\mathrm{X}_i^r)} - \frac{I(A_i^r = a) - \hat{\pi}_{a,r}(\mathrm{X}_i^r)}{\hat{\pi}_{a,r}(\mathrm{X}_i^r)} \hat{m}_{a,r}(\mathrm{X}_i^r) \right\}$$

Simulation Setup

- $K = 50$ sites with $n_r = (100, 200, 500)$ patients per site
- $p = (10, 100, 500)$ variables
  - Normally distributed with site-specific mean: $X_{i,r}^{(p)} \sim N(\mu_r^{(p)}, \sigma^2)$
  - No correlation between variables
- Outcome depends on $p^* = 5$ variables:
  $E(Y^r \mid A^r = a, \mathrm{X}^r = \mathrm{x}) = \beta_A A^r + .5X_1^r + X_2^r + .5X_3^r - .5X_4^r - X_5^r$
- Treatment assignment depends on same $p^* = 5$ variables:
  $P(A^r = a \mid \mathrm{X}^r = \mathrm{x}) = \mathrm{expit}(.05X_1^r + .1X_2^r + .1X_3^r - .05X_4^r - .1X_5^r)$
- $\beta_A = (0, 10)$ is the true ATE
- $\mathcal{T}$ is either 5 sites or all $K = 50$ sites

Regularization Results

Effect	n_r	Penalty	Sensitivity	Specificity
0	100	LASSO	0.84	0.85
		MCP	0.80	0.98
		SCAD	0.81	0.94
	200	LASSO	0.84	0.86
		MCP	0.80	0.99
		SCAD	0.80	0.97
	500	LASSO	0.83	0.87
		MCP	0.80	0.99
		SCAD	0.80	0.98
10	100	LASSO	1.00	0.84
		MCP	1.00	0.99
		SCAD	1.00	0.98
	200	LASSO	1.00	0.85
		MCP	1.00	0.99
		SCAD	1.00	0.97
	500	LASSO	1.00	0.86
		MCP	1.00	0.99
		SCAD	1.00	0.98

$K = 50$ sites of $n_r \in \{100, 200, 500\}$ people, each with $p = 100$ covariates $\sim \mathcal{N}(0, 1)$. β_A was either 0 or 10. Sensitivity refers to the percentage of correctly identified non-zero β's; specificity refers to the percentage of correctly identified zero β's.

Low-D Estimator Results

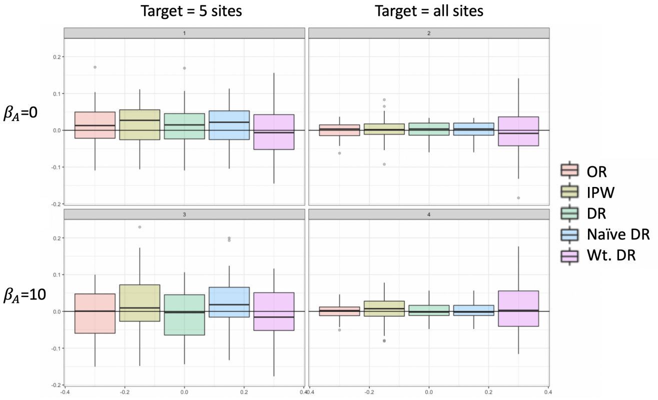
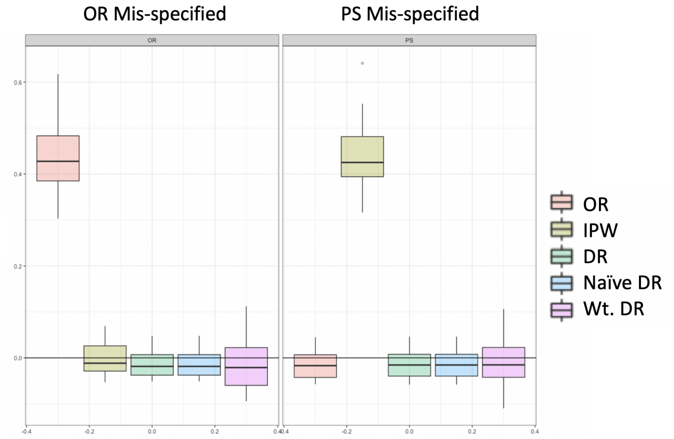


Figure: Comparison of five estimators in low-dimensional case. All models correctly specified, no structured between-site heterogeneity. $K = 50, n = 100, p = 10$. All estimators are unbiased, Weighted DR is inefficient when the target is all sites.

Misspecifying Models



OR Mis-specified

PS Mis-specified

Figure: Purposefully misspecifying either the outcome regression or the propensity score model by not including two important covariates, causing bias in the OR and IPW estimators respectively. $n = 100$, $p = 100$, $\beta_A = 10$ in all $K = 50$ sites.

Structured Between-Site Heterogeneity

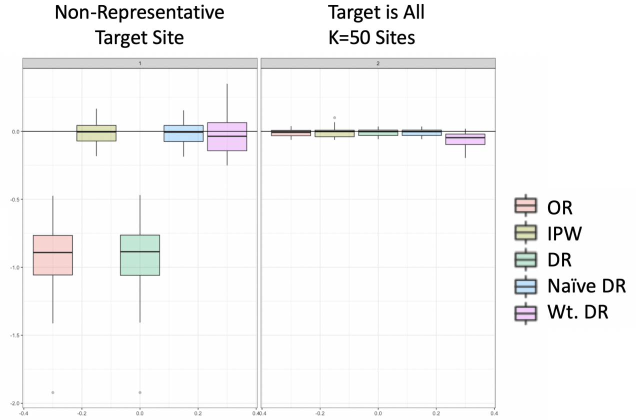


Figure: Covariate heterogeneity between 5-site target and other 45 sites with covariate-dependent TATE leads to bias in the OR and DR estimators. $n = 100, p = 100$

High Dimensional TATE

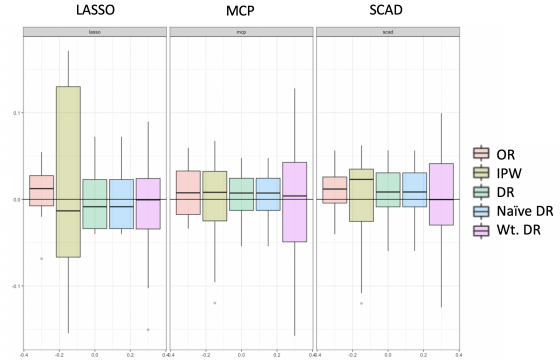


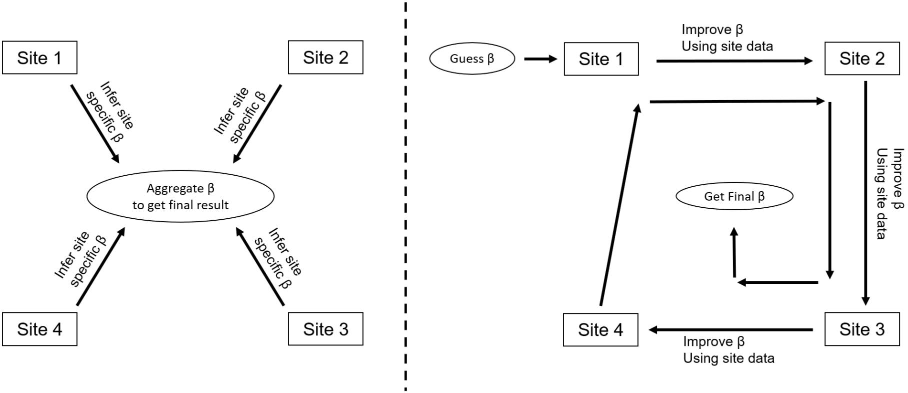
Figure: Comparing five estimators across three regularization methods. Note that the LASSO failed to converge more often than the other methods in this setting, when $n = 100$ and $p = 500$ with $p^* = 5$ non-zero entries in the true β vector. $\beta_A = 10$

Approach 1 Discussion

- Two-part framework allows for dimension-reduction first and then distributed learning while maintaining patient privacy
- Estimators perform roughly as expected
- Weighted doubly robust estimator should allow for patients similar to target site population to contribute to estimating TATE...
  - ...but performance is dependent on aggregation method (needs more fine-tuning)
- Further research needed to assess efficiency rigorously
- Our method only allows central site to perform variable selection

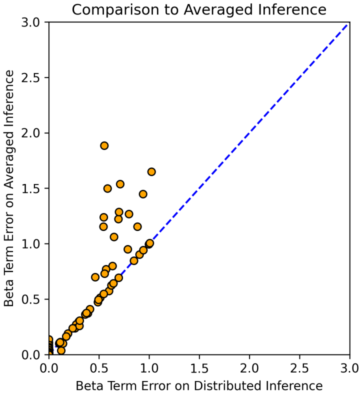
Approach 2

We can't move the data, so let's move the estimators to the data.



Each site uses the PyTorch Adamax optimizer to take 10 steps to improve β before passing it on.

Regression Comparison



The x axis is the β term prediction error for our method. The y axis is the β term prediction error for the naive method (left on previous slide).

Treatment Effect Inference

- Run the distributed LASSO, SCAD, or MCP with cross validation on all the data to select parameters meaningful to the treatment and outcome.
- Take only the selected parameters, add the treatment indicator variable A, and run a distributed linear regression on outcome Y
- The coefficient of A is an estimator of the average treatment effect
- Results on testing was a less than 4% error on estimating the average treatment effect

Approach 2 Discussion & Future Directions

- We have a method that performs regression without compromising privacy.
- This method works and is superior to naive methods with prebuilt packages.
- Using this method to infer average treatment effect works.

Future Directions

- It would be nice to infer targeted average treatment effect with this method.
- We could use distributed regression for variable selection and go back to approach 1.
- We could also weight people by similarity to the target and adjust their contribution to the loss accordingly.

Selected References

H. Battey, J. Fan, H. Liu, J. Lu, and Z. Zhu. Distributed testing and estimation under sparse high dimensional models. *Annals of statistics*, 46(3):1352, 2018.

G. A. Brat, G. M. Weber, N. Gehlenborg, P. Avillach, N. P. Palmer, L. Chiovato, J. Cimino, L. R. Waitman, G. S. Omenn, A. Malovini, et al. International electronic health record-derived covid-19 clinical course profiles: the 4ce consortium. *medRxiv*, 2020.

Y. Chen, G. Dong, J. Han, J. Pei, B. W. Wah, and J. Wang. Regression cubes with lossless compression and aggregation. *IEEE Transactions on Knowledge and Data Engineering*, 18(12):1585–1599, 2006.

R. Duan, M. R. Boland, Z. Liu, Y. Liu, H. H. Chang, H. Xu, H. Chu, C. H. Schmid, C. B. Forrest, J. H. Holmes, et al. Learning from electronic health records across multiple sites: A communication-efficient and privacy-preserving distributed algorithm. *Journal of the American Medical Informatics Association*, 27(3):376–385, 2020a.

R. Duan, C. Luo, M. H. Schuemie, J. Tong, J. C. Liang, H. H. Chang, M. R. Boland, J. Bian, H. Xu, J. H. Holmes, et al. Learning from local to global-an efficient distributed algorithm for modeling time-to-event data. *bioRxiv*, 2020b.

M. I. Jordan, J. D. Lee, and Y. Yang. Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, 2018.

J. D. Lee, Q. Liu, Y. Sun, and J. E. Taylor. Communication-efficient sparse regression. *The Journal of Machine Learning Research*, 18(1):115–144, 2017.

C.-L. Lu, S. Wang, Z. Ji, Y. Wu, L. Xiong, X. Jiang, and L. Ohno-Machado. Webdisco: a web service for distributed cox model learning without patient-level data sharing. *Journal of the American Medical Informatics Association*, 22(6):1212–1219, 2015.

J. Qin. Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, 85(3):619–630, 1998.

Y. Wu, X. Jiang, J. Kim, and L. Ohno-Machado. Grid binary logistic regression (glore): Building shared models without sharing data. *Journal of the American Medical Informatics Association*, 19(5):758–764, 2012.