Estimation of average treatment effect for ACIC data challenge

Wanwan Xu

Univeristy of Connecticut, Department of Statistics

Abstract

This is a data-driven project based on ACIC 2019 data challenge, the main objective is to estimate the population average treatment effect (ATE). The performance of regression estimator, inverse propensity weighted estimator, augmented inverse propensity weighted estimator and calibration estimator is compared on test datasets as well as some theoretical properties. The final result for all 3,000 sets of training data is also presented.

Keywords: average treatment effect, counterfactual, causal inference

1 Introduction

Treatment effect has always been of interest, as it quantitatively show the causal effect of a binary variable on an outcome variable. This variable, although with name 'treatment", is now used much more generally than a medical procedure or experimental drug. In economics, the causal effect of a subsidized training program [1] is widely analyzed, so are the effects of active labor market programs such as job search assistance or classroom teaching program [9]. Since treatment effect is defined as the difference between two treatment groups, while in reality, one unit or one observation is only related to one of the treatment assignment, it's impossible to measure causal effects at the individual level, therefore we focus on the expectation, i.e. average treatment effect(ATE):

$$\text{ATE} = E[Y_{1i} - Y_{0i}] = E[E(\mathbf{Y}|\mathbf{A} = 1, \mathbf{V}) - E(\mathbf{Y}|\mathbf{A} = 0, \mathbf{V})]$$

where $\mathbf{Y}$ is the response variable, $\mathbf{A}$ is treatment variable and $\mathbf{V}$ is the set of covariates.

A typical concern for inferring causality in an observational study is confounding, treatment exposure may be associated with covariates that are also associated with potential response. For example, experiment individual characteristics such as demographic factors can be related to both the treatment selection and the outcome of interest. In order to perform unbiased comparison despite the confounding factors, the inferences with causal interpretation need to be adjusted. There are three broad classes of strategies: (1) Outcome regression [11]. The model of outcome $\mathbf{Y}$ given covariates $\mathbf{V}$ need to be specified, it follows similar procedure of calculating counterfactuals. (2) Propensity score estimation [14]. This method requires the construction of weights for individual observations via using estimated propensity scores [13], then use the average of within stratum effects as ATE [15]. (3) Non-parametric estimation [4]. The authors applied the class of survey calibration estimators on the estimation of average treatment effects, they also showed that a globally efficient estimator can be adapted from it.

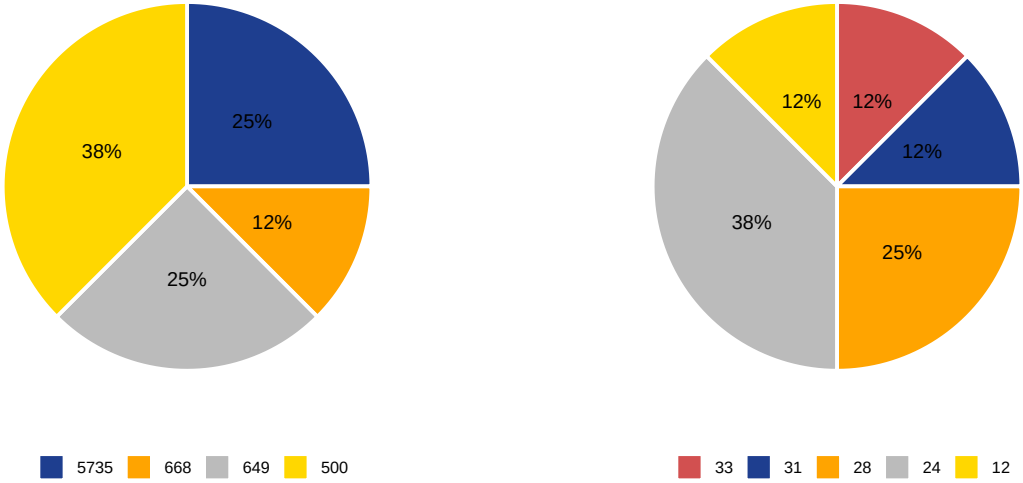
This report focuses on applying several estimation methods described above to ACIC data challenge low dimension track datasets. The report is organized as follows. In Section 2, we introduce the data structure and explanation. The methods applied is briefly introduced in Section 3, the comparison based on test dataset is presented in Section 4 as well as the results for training dataset. Some final remarks and future directions are given in Section 5.

2 Dataset

2.1 Introduction

Two tracks of data (low dimension and high dimension) are included in the ACIC data challenge, each track have 3200 csv files drawn from 32 unique data generating processes (DPG). In this report, we focus only on low dimensional track. Within each csv file, there are:

- response variable $\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_n\}$, half are continuous while the other half is discrete;



(a) number of observations

(b) number of covaraibles

Figure 1: Frequency plot low track dataset

- binary treatment indicator $\mathbf{A} = \{A_1, A_2, \ldots, A_n\}$;
- pre-treatment covariates $\mathbf{V} = \{V_1, \ldots, V_p\}$, which includes the mixture of continuous, binary, and categorical variables.

Aside from the mixture of data type, number of covariates and number of observations may also be different for each worksheet as shown in Figure 1. The description site of data challenge also provided some more information. For instance, there are no mediators among covariates, that is we only need to consider direct effect from covariates. Also, all the covaraites also include all the cofounders, we don't need to consider hidden variables. However, selection variables may still exists, given some of the covaraites may be related to the treatment factor $\mathbf{A}$.

2.2 Objectives

The measure of estimation is the population average treatment effect (ATE) and 95% confidence interval for each dataset. For individual $i \in \{1, 2, \ldots, n\}$, denote Y_{0i} as the outcome variable for individual i if he/she is not treated, y_{1i} as the value of the outcome variable for individual i if he/she is treated. The treatment effect for individual i is given by $Y_{1i} - Y_{0i} = \beta_i$. In the general case, there is no reason to expect this effect to be constant across individuals. The average treatment effects is given by $\text{ATE} = E[Y_{1i} - Y_{0i}]$, where the expectation is with respect to the distribution of covariates.

However, in observational data, we can only observe one of the two for each individual, thus estimation techniques are needed. Under this circumstances, one of the y_{0i}, y_{1i} will be replaced by counterfactual output. Since the causal assumptions of consistency and strong ignorability are guaranteed, the target statistical estimate can be rewritten as: $\text{ATE} = E[E(\mathbf{Y}|\mathbf{A} = 1, \mathbf{V}) - E(\mathbf{Y}|\mathbf{A} = 0, \mathbf{V})]$.

3 Method

3.1 Basic models

As presented before, I compared a few different methods on the ATE estimationATE $= E[E(\mathbf{Y}|\mathbf{A} = 1, \mathbf{V}) - E(\mathbf{Y}|\mathbf{A} = 0, \mathbf{V})]$, including:

Outcome model (REG):

Lots of traditional causal estimation relies on the formulation of a regression model for the outcome variable $\mathbf{Y}$. In other words, the focus is on the estimation of $E[\mathbf{Y}|\mathbf{A}, \mathbf{V}]$. Given ATE is defined with respect to the distribution of covariates, the empirical distribution will be an easy estimate of $F_{\mathbf{V}}$. Thus the regression estimate can be written as: $\widehat{\text{ATE}}_{REG} = \frac{1}{n}\sum_{i=1}^{n}\{\hat{E}(\mathbf{Y}|\mathbf{A} = 1, V_i) - \hat{E}(\mathbf{Y}|\mathbf{A} = 0, V_i)\}$

Propensity score model (IPW):

Due to the limitation of regression estimator, another widely applied method for estimating ATE relies on a model for treatment assignment instead a regression model for the outcome was proposed. Suppose the true

probability of assigning treatment were known, this could be used to define propensity scores, and further be used for matching or weighting covariates. A well-known weighting estimator is the IPW estimator: $\widehat{\text{ATE}}_{IPW} = \left\{\sum_{i=1}^n \frac{A_i}{\hat{\pi}(\mathbf{V}_i)}\right\}^{-1} \sum_{i=1}^n \frac{A_i Y_i}{\hat{\pi}(\mathbf{V}_i)} - \left\{\sum_{i=1}^n \frac{1-A_i}{1-\hat{\pi}(\mathbf{V}_i)}\right\}^{-1} \sum_{i=1}^n \frac{(1-A_i)Y_i}{1-\hat{\pi}(\mathbf{V}_i)}$

AIPW model:

Based on the previous work of IPW estimation, one improvement including fully utilizing the information in the conditioning set $\mathbf{V}$ is AIPW [6]. Aside from the information about treatment assigning probability, $\mathbf{V}$ also carry the predictive information about the response variable $\mathbf{Y}$:

$$\widehat{\text{ATE}}_{AIPW} = \frac{1}{n}\sum_{i=1}^n \Bigg\{ \left[\frac{A_i Y_i}{\hat{\pi}(\mathbf{V}_i)} - \frac{(1-A_i)Y_i}{1-\hat{\pi}(\mathbf{V}_i)}\right] - \frac{(A_i - \hat{\pi}(\mathbf{V}_i))}{\hat{\pi}(\mathbf{V}_i)(1-\hat{\pi}(\mathbf{V}_i))} + 1 - \hat{\pi}(\mathbf{V}_i))\hat{E}(Y_i|A_i=1,\mathbf{V}_i) + \hat{\pi}(\mathbf{V}_i)\hat{E}(Y_i|A_i=0,\mathbf{V}_i)] \Bigg\}$$

AIPW estimator has also been shown to posses double-robustness. As long as one of the models are consistent, the final estimator will be consistent.

Non-parametric model (CAL):

Previous few estimate methods both require estimation of a propensity score function, and an outcome regression function, IPW and AIPW is globally semiparametric efficient when the usual maximum likelihood propensity score estimator is used. Chan [4] proposed a wide class calibration weights to attain the moments of observed covariates among the treated, the control, and the combined group. Furthermore, these empirical weights can be applied for globally efficient non-parametric inference of ATE. Chan and Yam reformulated the problem into optimization framework: Minimize $\sum_{i=1}^n D(w_i, 1)$ subject to:

$$\frac{1}{n}\sum_{i=1}^n A_i w_i u(\mathbf{V}_i) = \frac{1}{n}\sum_{i=1}^n u(\mathbf{V}_i),\ \frac{1}{n}\sum_{i=1}^n (1-A_i) w_i u(\mathbf{V}_i) = \frac{1}{n}\sum_{i=1}^n u(\mathbf{V}_i).$$

Based on the similar formula and sequentially motivation, it's reasonable to ask what is in common and what is different among these methods. Denote $\pi(\mathbf{V}) = P(\mathbf{A}=1|\mathbf{V})$ as propensity score, $m_1(\mathbf{v}) = E(\mathbf{Y}(1)|\mathbf{V}=\mathbf{v})$ and $m_0(\mathbf{v}) = E(\mathbf{Y}(0)|\mathbf{V}=\mathbf{v})$ as conditional mean function, we have:

$$\begin{aligned}
\text{ATE} &= E[\frac{\mathbf{AY}}{\pi(\mathbf{V})} - \frac{(1-\mathbf{A})\mathbf{Y}}{1-\pi(\mathbf{V})}] \qquad \text{(IPW)} \\
&= E[m_1(\mathbf{V}) - m_0(\mathbf{V})] \qquad \text{(REG)} \\
&= E[\frac{\mathbf{AY}}{\pi(\mathbf{V})} - \frac{\mathbf{A}-\pi(\mathbf{V})}{\pi(\mathbf{V})} m_1(\mathbf{V}) \\
&\quad - \frac{(1-\mathbf{A})\mathbf{Y}}{1-\pi(\mathbf{V})}] - \frac{\mathbf{A}-\pi(\mathbf{V})}{1-\pi(\mathbf{V})} m_0(\mathbf{V})] \qquad \text{(AIPW)}
\end{aligned}$$

while CAL use the property: $E[\frac{\mathbf{A}u(\mathbf{V})}{\pi(\mathbf{V})}] = E[\frac{(1-\mathbf{A})u(\mathbf{V})}{1-\pi(\mathbf{V})}] = E[u(\mathbf{V})]$.

Table 1: Summary of the basic models

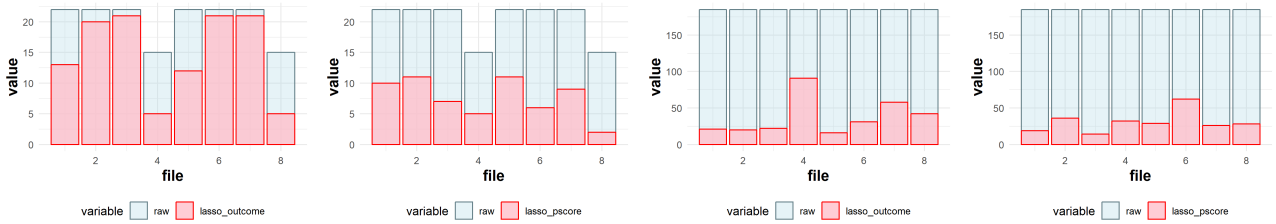
	REG	IPW	AIPW	CAL
Variables for Outcome	✓	✗	✓	✓
Variables for Pscore	✗	✓	✓	✓
Model for Outcome	✓(specific)	✗	✓(df)	✗
Model for Pscore	✗	✓(df)	✓(df)	✗

3.2 Two-Step modification

As shown in Table 1, all basic models require pre specification of some formulas, then what formula should we choose and which variables to include in these formulas is our next question. Borrowing the idea of Lasso regression, we propose a two step modification start by selecting the 'important" variables first by LASSO, then apply previous methods. Figure 2 shows the selected variables for the outcome model and propensity score model for both low dimension and high dimension track. We can see there are some non-significant features in the provided data set. The features selected will be carried along to the next modeling step.

3.3 Causal Forest

We also considered causal forest [2] as another way to perform variable selection. For regular CART regression tree with samples (V_i, Y_i), we start by recursively splitting the feature space until it's partitioned into a set of



(a) outcome (low) (b) propensity score (low) (c) outcome (high) (d) propensity score (high)

Figure 2: Lasso selected variables for both data tracks

leaves L, each of which only contains a few training samples. Then, given a test point v, first identify which leaf $L(v)$ contain v, then prediction is:

$$\hat{\mu}(v) = \frac{1}{|\{i : V_i \in L(v)\}|} \sum_{\{i:V_i \in L(v)\}} Y_i$$

Analogously, our dependent variables are (Y_i, A_i), with ignorability, $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp A_i | V_i$,and recall

$$\text{ATE} = E[E(\mathbf{Y}|\mathbf{A} = 1, \mathbf{V}) - E(\mathbf{Y}|\mathbf{A} = 0, \mathbf{V})]$$

we have the causal forest predicted for ATE:

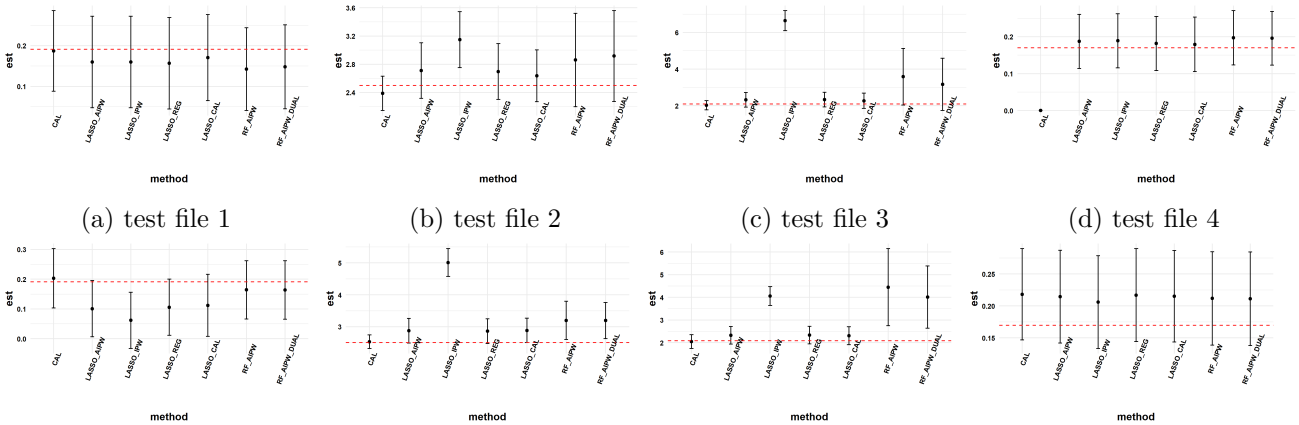
$$\hat{ATE}(v) = \frac{1}{|\{i : A_i = 1, V_i \in L\}|} \sum_{\{i:A_i=1,V_i \in L(v)\}} Y_i - \frac{1}{|\{i : A_i = 0, V_i \in L\}|} \sum_{\{i:A_i=0,V_i \in L(v)\}} Y_i$$

The authors show the estimate is point wise consistent: $\hat{\tau} \to \tau$ in Probability, and have an asymptotically Gaussian and centered sampling distribution: $(\hat{\tau} - \tau)/\sqrt{Var(\hat{\tau})} \to \mathcal{N}(0, 1)$. Similar to before, we will show results for directly apply causal forest, and apply the two stage causal forest algorithm. For two stage, firstly fit the separate regression/classification tree for the outcome and treatment, then use selected variable from those trees to build a new causal tree to estimate ATE.

4 Result

4.1 Training dataset

Firstly, we show the results of all basic methods and the two-step methods for low dimensional data in Figure 3. Considering the big difference between total number of variables and the number of selected variables in high dimension track, the comparison is also shown in 4. From the low dimension track, we can see the two step IPW



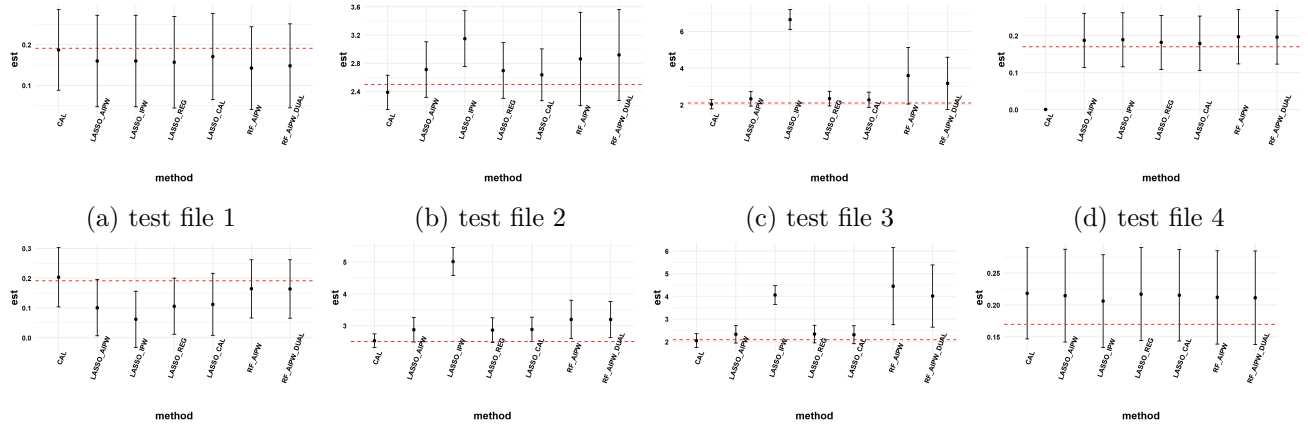
(a) test file 1 (b) test file 2 (c) test file 3 (d) test file 4

(e) test file 5 (f) test file 6 (g) test file 7 (h) test file 8

Figure 3: Estimation of ATE with 95% confidence interval on the low track files

is quite unstable, which matches our prediction given it failed to consider the propensity score part. Similar

conclusion holds for high dimension track as well. Table **??** is a summary for Figure 3. Y indicates the true



(a) test file 1 (b) test file 2 (c) test file 3 (d) test file 4

(e) test file 5 (f) test file 6 (g) test file 7 (h) test file 8

Figure 4: Estimation of ATE with 95% confidence interval on the files of high track

ATE is contained in the 95% confidence interval of that specific method, the annotated N indicates the method fail to capture true ATE. We can see the non-parametric estimation CAL and the two step causal forest with AIPW formula performs best. Aside from the TRUE/FALSE answer of whether the true ATE is included in

	CAL	LASSO AIPW	LASSO IPW	LASSO REG	LASSO CAL	RF AIPW	RF_AIPW DUAL
1	Y	Y	Y	Y	Y	Y	Y
2	Y	Y	N*	Y	Y	Y	Y
3	Y	Y	N*	Y	Y	N*	Y
4	N*	Y	Y	Y	Y	Y	Y
5	Y	N*	N*	Y	Y	Y	Y
6	Y	Y	N*	Y	Y	N*	N*
7	Y	Y	N*	Y	Y	N*	N*
8	Y	Y	Y	Y	Y	Y	Y

Table 2: Summary table of 95% confidence interval for all methods

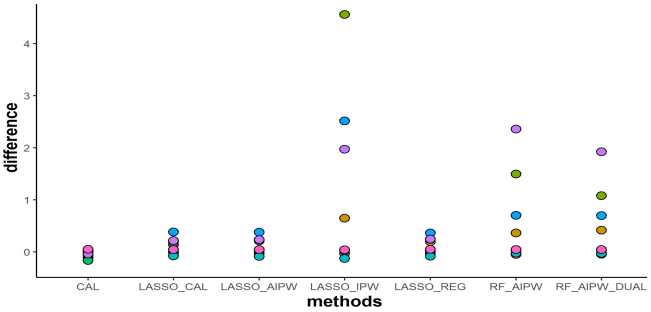


Figure 5: Comparison of the difference with true ATE for all methods

the 95% confidence interval, the actual difference of point estimate and true ATE is also of our interest. The results of differences between True ATE and point estimate based on all 8 methods are presented in Figure 5. We can see LASSO_IPW have largest differences compare to other methods, which confirms the previous conclusion that LASSO_IPW fails to calculate the ATE. Another interesting thing is that even the causal forest methods contain the true ATE in most cases, it also have fairly large point estimate difference in some cases, which can be explained by the large covariance induced in the method.

4.2 Test dataset

We perform the above methods on all 3,200 testing files. Since there is no true ATE to compare to, and all we do have the information that those files are generated from 32 separate DPGs and each DGP have the same type ATE. In other words, there should be at most 32 unique true ATEs, thus one way to check the quality of the results for test dataset is to cluster the estimated ATE, hopefully we can see a clear division.

Figure 6 is the result from K-means clustering with pre-specified 32 clusters. We can see that it's not easy to tell if the result makes sense, similarly for the univariate clustering methods as shown in Figure 7.

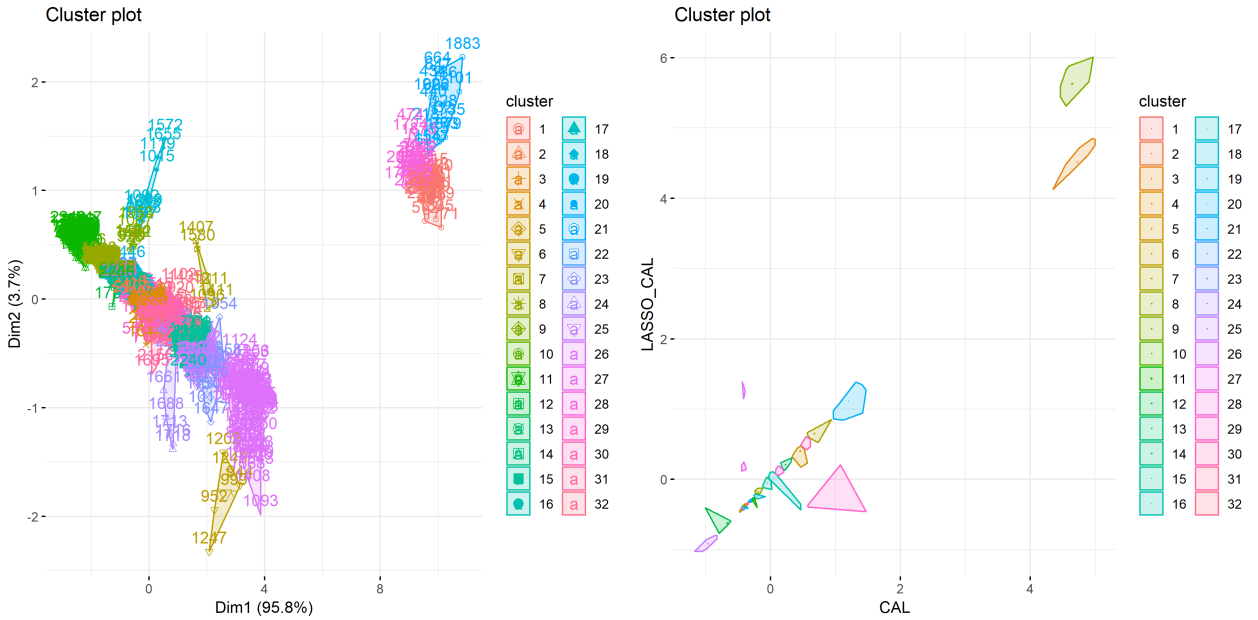


Figure 6: K-means clustering result for all 3200 low track csv files, left side is for all estimation method, right side is for CAL and LASSO_CAL

Without specifying the cluster number, there were 27 clusters detected, 14 was identified for both LASSO_CAL and LASSO_AIPW methods, and both Random forest methods detect 8 clusters. However, these results serve no more than a possible measure of how well is the performance, especially without more knowledge of what the true ATEs are.

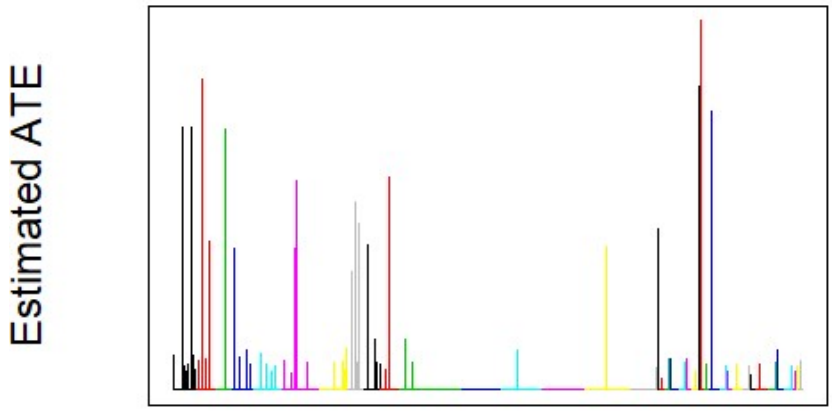


Figure 7: Univariate clustering result for LASSO_CAL method only, detected 14 clusters.

5 Discussion

In this method we briefly compare seven methods, which can be characterized as either the basic model or the two-step. Based on the training set, calibration weights (CAL), as a non-parametric estimation process, outperforms other methods in most cases. In general, the two step methods performed better than the basic model. There are still a few questions left , including the actual algorithm performance on testing set, which is not accessible until the meeting ends. As for the methodology, other possible extensions including the usage of non-linear model for both outcome and propensity score, or add in the interaction terms into the modeling. Other variable selection methods, like elastic net(ENET), may also be applied.

References

[1] Ashenfelter, Orley. "Estimating the effect of training programs on earnings." The Review of Economics and Statistics (1978): 47-57.

[2] Athey, Susan, Julie Tibshirani, and Stefan Wager. "Generalized random forests." The Annals of Statistics 47.2 (2019): 1148-1178.

[3] Chambers, John M., and Trevor J. Hastie, eds. Statistical models in S. Vol. 251. Pacific Grove, CA: Wadsworth & Brooks/Cole Advanced Books & Software, 1992.

[4] Chan, Kwun Chuen Gary, Sheung Chi Phillip Yam, and Zheng Zhang. "Globally efficient nonparametric inference of average treatment effects by empirical balancing calibration weighting." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 78.3 (2016): 673-700.

[5] Chen, Xiaohong, Han Hong, and Alessandro Tarozzi. "Semiparametric efficiency in GMM models with auxiliary data." The Annals of Statistics 36.2 (2008): 808-843.

[6] Glynn, Adam N., and Kevin M. Quinn. "An introduction to the augmented inverse propensity weighted estimator." Political analysis 18.1 (2010): 36-56.

[7] Imbens, Guido W. "Nonparametric estimation of average treatment effects under exogeneity: A review." Review of Economics and statistics 86.1 (2004): 4-29.

[8] King, Gary, and Langche Zeng. "The dangers of extreme counterfactuals." Political Analysis 14.2 (2006): 131-159.

[9] LaLonde, Robert J. "Evaluating the econometric evaluations of training programs with experimental data." The American economic review (1986): 604-620.

[10] Lunceford, Jared K., and Marie Davidian. "Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study." Statistics in medicine 23.19 (2004): 2937-2960.

[11] Oaxaca, Ronald. "Male-female wage differentials in urban labor markets." International economic review (1973): 693-709.

[12] Pearl, Judea. "Causality: models, reasoning, and inference." IIE Transactions 34.6 (2002): 583-589.

[13] Robins, James M., Miguel Angel Hernan, and Babette Brumback. "Marginal structural models and causal inference in epidemiology." (2000): 550-560.

[14] Rosenbaum, Paul R., and Donald B. Rubin. "The central role of the propensity score in observational studies for causal effects." Biometrika 70.1 (1983): 41-55.

[15] Rosenbaum, Paul R., and Donald B. Rubin. "Reducing bias in observational studies using subclassification on the propensity score." Journal of the American statistical Association 79.387 (1984): 516-524.