Observational Studies () Submitted ; Published

Selective Inference for Effect Modification: An Empirical Investigation

Qingyuan Zhao **qyzhao@wharton.upenn.edu**
Department of Statistics
The Wharton School, University of Pennsylvania
Philadelphia, PA 19104, USA

Snigdha Panigrahi **psnigdha@umich.edu**
Department of Statistics
University of Michigan
Ann Arbor, MI 48109, USA

Abstract

We demonstrate a selective inferential approach for discovering and making confident conclusions about treatment effect heterogeneity. Our method consists of two stages. First, we use Robinson's transformation to eliminate confounding in the observational study. Next we select a simple model for effect modification using lasso-regularized regression and then use recently developed tools in selective inference to make valid statistical inference for the discovered effect modifiers. We analyze the Mindset Study data-set provided by the workshop organizers and compare our approach with other benchmark methods.

Keywords: Lasso, Semiparametric regression, Selective sampler, Variable selection.

1. Methodology and Motivation

1.1 Motivation

In the 2018 Atlantic Causal Inference Conference (ACIC 2018), we were kindly invited to participate in a workshop titled "Empirical Investigation of Methods for Heterogeneity". The workshop organizers provided an observational dataset simulated from the National Study of Learning Mindsets (Mindset Study hereafter) and tasked the participants to analyze how treatment effect of the mindset intervention varies among students in the study. This workshop, in the words of the organizers, "is not intended to be a 'bake off' but rather an opportunity to understand the strengths and weaknesses of methods for addressing important scientific questions". More specifically, the organizers sought answers for the following three research questions about the Mindset Study:

Question 1: Is the intervention effective in improving student achievement?

Question 2: Do two hypothesized covariates (`X1` and `X2`) moderate the treatment effect?

Question 3: Are there other covariates moderating the treatment effect?

In this report, we will attempt to answer these questions using a method proposed in our earlier paper (Zhao et al., 2017) which neatly combines Robinson's transformation

(Robinson, 1988) to remove confounding and the recently developed selective inferential framework (Taylor and Tibshirani, 2015) to discover and make confident conclusions about effect modifiers (covariates moderating the treatment effect).

Effect modification or treatment effect heterogeneity is an old topic in statistics but has gained lots of attention in recent years, possibly due to the increased complexity of empirical datasets and the development of new statistical learning methods that are much more powerful at discovering effect modification. Though the literature on this topic is massive, an executive summary must include three related but different formulations of this problem:

1. What is the optimal treatment assignment rule for future experimental objects?
2. What is the conditional average treatment effect (CATE) as a function of the covariates?
3. What are the potential effect modifiers and how certain are we about them?

See Zhao et al. (2017) for more discussion and references. It is obvious that the questions of the workshop organizers fall into the third category. In fact, we believe this is quite common in practice. Empirical researchers often want to use observational or experimental data to test existing scientific hypotheses about effect modification, generate new hypotheses, and gather information for intelligent decision making. However, prior to Zhao et al. (2017), majority of the statistical methods in the third category focused on discovering potential effect modifiers with little attention targeted towards providing statistical inference (such as confidence intervals for the discovered covariates). When the goal is to calibrate the strengths of effect modifiers in such problems, the researcher often relies on sample splitting, where some of the samples are used for discovery and the remaining samples are used for inference (Athey and Imbens, 2016). However, sample splitting does not optimally utilize the information in the discovery samples and often results in loss of power. Instead, the selective inference framework described in this paper does not waste any data, as it leverages on a conditional approach that only discards the information used in model selection (Lee et al., 2016; Fithian et al., 2014).

1.2 Main method

To introduce the methodology let's first fix some notations. Let Y be the observed outcome (a continuous measure of academic achievement), Z be the binary intervention (0 for control and 1 for treated), and $\boldsymbol{X} = (X_1, \ldots, X_p)$ be the covariates ($p = 10$ in the Mindset Study). Furthermore, denote $Y(0)$ and $Y(1)$ as the two potential outcomes, thus $Y = Y(Z)$. We assume that there are no unmeasured confounders throughout the paper, i.e. $Y(z) \perp\!\!\!\perp Z \,|\, X$ for $z = 0, 1$.

Below we will elaborate the two-step method proposed in Zhao et al. (2017):

Step 1 (Robinson's transformation): Use machine learning methods to estimate $\mu_z(\boldsymbol{x}) = \mathbb{E}[Z|\boldsymbol{X} = \boldsymbol{x}] = \mathbb{P}(Z = 1|\boldsymbol{X} = \boldsymbol{x})$ (the "propensity' score") and $\mu_y(\boldsymbol{x}) = \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$. Let the estimates be $\hat{\mu}_y(\boldsymbol{x})$ and $\hat{\mu}_z(\boldsymbol{x})$. In `R`, there are many off-the-shelf implementations available to learn $\mu_y(\boldsymbol{x})$ and $\mu_z(\boldsymbol{x})$ without any ex ante model specification. It is helpful to use an algorithm called "cross-fitting" in this step for the purpose of

proving theoretical properties (Schick, 1986; Chernozhukov et al., 2018). Cross-fitting is only implemented for the post-workshop analysis. See Section 3.1 for more detail.

Notice that it is straightforward to show (see Zhao et al., 2017) that the CATE $\Delta(\boldsymbol{x}) = \mathbb{E}[Y(1) - Y(0)|\boldsymbol{X} = \boldsymbol{x}]$ satisfies

$$\mathbb{E}[Y - \mu_y(\boldsymbol{X}) \,|\, Z, \boldsymbol{X}] = (Z - \mu_z(\boldsymbol{X}))\Delta(\boldsymbol{X}) \tag{1}$$

Step 2 (Statistical inference): By approximating the CATE using a linear model, $\Delta(\boldsymbol{x}) \approx \beta_0 + \boldsymbol{x}^T\boldsymbol{\beta}$, equation (1) implies that

$$Y - \hat{\mu}_y(\boldsymbol{X}) \approx (Z - \hat{\mu}_z(\boldsymbol{X}))(\beta_0 + \boldsymbol{X}^T\boldsymbol{\beta}) + \text{approximation error} + \text{noise}.$$

This motivates us to treat $\tilde{Y} = Y - \hat{\mu}_y(\boldsymbol{X})$ as the (transformed) response and $\tilde{\boldsymbol{X}} = (Z - \hat{\mu}_z(\boldsymbol{X}))\boldsymbol{X}$ as the (transformed) predictors. We can then use different specifications of $\Delta(\boldsymbol{x})$ to answer the three questions posted by the workshop organizers:

Step 2.1 (answering Question 1): Model CATE by just an intercept term: $\boldsymbol{\Delta}(\boldsymbol{x}) \approx \beta_0$. In `R`, we can report the results of the linear regression $\texttt{lm}(\tilde{Y} \sim \tilde{Z})$ where $\tilde{Z} = Z - \mu_z(\boldsymbol{X})$.

Step 2.2 (answering Question 2): Suppose $\boldsymbol{X}_{\mathcal{M}}$ are the hypothesized effect modifiers (`X1` and `X2` in the Mindset Study). We can model CATE by an intercept and $\boldsymbol{X}_{\mathcal{M}}$: $\boldsymbol{\Delta}(\boldsymbol{x}) \approx \beta_0 + \boldsymbol{X}_{\mathcal{M}}^T\boldsymbol{\beta}_{\mathcal{M}}$. The coefficient $\boldsymbol{\beta}_{\mathcal{M}}$ can be interpreted as the coefficient in the best linear approximation to the actual $\Delta(\boldsymbol{x})$. More precisely, it is defined as (see Zhao et al., 2017):

$$(\beta_0, \boldsymbol{\beta}_{\mathcal{M}}) = \arg\min \; \mathbb{E}_n\Big[(Z - \mu_z(\boldsymbol{X}))^2(\Delta(\boldsymbol{X}) - \beta_0 - \boldsymbol{X}_{\mathcal{M}}^T\boldsymbol{\beta}_{\mathcal{M}})^2\Big], \tag{2}$$

where $\mathbb{E}_n$ stands for averaging over the n samples. In `R`, we can report the results of the linear regression $\texttt{lm}(\tilde{Y} \sim \tilde{Z} + \tilde{Z} : \boldsymbol{X}_{\mathcal{M}})$.

Step 2.3 (answering Question 3): Use lasso regularized regression in Tibshirani (1996) (or potentially other automated variable selection methods) to select a subset of covariates $\hat{\mathcal{M}} \subseteq \{1, 2, \ldots, p\}$. More specifically, $\hat{\mathcal{M}}$ contains positions of non-zero entries in the solution to the following problem:

$$\text{minimize} \quad \frac{n}{2}\,\mathbb{E}_n\Big[\tilde{Y} - \tilde{Z}(\beta_0 + \boldsymbol{\beta}^T\boldsymbol{X})\Big]^2 + \lambda\|\boldsymbol{\beta}\|_1. \tag{3}$$

Then we can use the existing selective inference methods to make inference about the linear submodel $\Delta(\boldsymbol{x}) \approx \beta_0 + \boldsymbol{x}_{\hat{\mathcal{M}}}^T\boldsymbol{\beta}_{\hat{\mathcal{M}}}$ that is selected using the data. The estimand $(\beta_0, \boldsymbol{\beta}_{\hat{\mathcal{M}}})$ is defined in the same way as (2) by treating $\hat{\mathcal{M}}$ as fixed.

The central idea behind the selective inferential methods is to base inference upon a conditional likelihood that truncates the usual (pre-selection) likelihood to the realizations of data that can lead to the same selection event. Lee et al. (2016) proposed the first method along this conditional perspective to overcome the bias encountered in inferring about a data-adaptive target. Assuming Gaussian noise in a linear regression setting, Lee et al. (2016) derived a pivotal statistic that

can be computed in closed-form and has a truncated Gaussian law for a class of polyhedral selection rules including the lasso (3). An implementation of this method can be found in the `selectiveInference` `R` package (Tibshirani et al., 2017). In principle, more sophisticated selective inference can be used in this step too. We will explore them in Section 3.

Compared to other methods, the approach outlined above has several appealing properties. First, the nuisance parameters—$\mu_z(\boldsymbol{x})$ and $\mu_y(\boldsymbol{x})$—are estimated by flexible machine learning methods. Because Robinson's transformation is used, each nuisance parameter only needs to be estimated at rate faster than $n^{-1/4}$ to ensure asymptotic validity of the non-selective or selective inference in Step 2 (Zhao et al., 2017). This echos the suggestion of combining machine learning methods and doubly robust estimation by van der Laan and Rose (2011); Chernozhukov et al. (2018). Second, all the the scientific questions raised by workshop organizers can be answered in the same manner. The data analyst only needs to change the specification of the model for $\Delta(\boldsymbol{x})$. Third, when answering Question 3, an effective variable selection procedure (such as lasso) can often find an interpretable model that includes most of the important effect modifiers. Selective inference can then provide valid statistical significance and confidence interval for the selected effect modifiers. Lastly, the implementation of this procedure is straightforward by harvesting existing softwares of machine learning methods and selective inference. We refer the reader to Zhao et al. (2017) for a more detailed discussion on the strengths and weaknesses of our approach.

1.3 Alternative methods

To provide a more comprehensive picture of our selective inference approach (referred to as method "lasso" below), we decided before seeing any real data in the Mindset Study that we would also use four benchmark methods considered in the applied example in Zhao et al. (2017). These alternative methods are:

Method "naive": This method simply fits a linear model with all the treatment by covariate interactions (and of course all the main effects). In `R`, we can simply use `lm(`$Y \sim Z * X$`)` which is equivalent to `lm(`$Y \sim Z + X + Z : X$`)`. To investigate effect modification, we can just report results for the interactions. This method is called "naive" because the linear model may be misspecified and may be insufficient for removing confounding.

Method "marginal": After Robinson's transformation (Step 1 above), this method fits univariate linear regressions `lm(`$\tilde{Y} \sim \tilde{Z} + \tilde{Z} : X_j$`)` for $j = 1, \ldots, p$. This is a special case of Step 2.2 with fixed model $\mathcal{M} = \{j\}$.

Method "full": After Step 1, this method fits a full linear model `lm(`$\tilde{Y} \sim \tilde{Z} + \tilde{Z} : \boldsymbol{X}$`)`. This is a special case of Step 2.2 with fixed model $\mathcal{M} = \{1, 2, \ldots, p\}$.

Method "snooping": This method is similar to method "lasso" except for the very last step. Instead of selective inference, it directly reports the results of `lm(`$\tilde{Y} \sim \tilde{Z} + \tilde{Z} : \boldsymbol{X}_{\hat{\mathcal{M}}}$`)` treating $\hat{\mathcal{M}}$ as given rather than learned from the data. This method is used as a straw man to illustrate that ignoring model selection (aka "data snooping") may lead to over-confident inference.

2. Workshop results

2.1 Implementation details

In our workshop analysis, we used the random forest (Breiman, 2001) to estimate the nuisance parameters in Step 1. In particular, we used the "honest" forest implementation in the `grf` package (Athey et al., 2018) with `tune.parameters = TRUE` (so some parameters will be tuned by cross-validation) and all other options set to default. In Step 2, categorical covariates are transformed to dummy variables. For example, `XC` (with five levels: 0, 1, 2, 3, 4) is transformed to `XC-1`, `XC-2`, `XC-3`, `XC-4`. In Step 2.3, we used the theoretical value $\lambda = 1.1 \times \mathrm{E}[\|\tilde{\mathbf{X}}^T \boldsymbol{\epsilon}\|_\infty]$ (Negahban et al., 2012) for model selection, where $\boldsymbol{\epsilon}$ is the vector of noise in the outcome. In the real data analysis λ is computed by simulating $\epsilon \overset{i.i.d.}{\sim} \mathrm{N}(0, \hat{\sigma}^2)$ where $\hat{\sigma}^2$ is the estimated noise level, see Lee et al. (2016). We then used the `fixedLassoInf` function in the `selectiveInference` package to make the selective inference. Details of the implementation can be found in the supplementary `R` markdown file.

2.2 Results

By simply specifying an intercept term for $\Delta(\boldsymbol{x})$ as described in Step 2.1, the (weighted) average treatment effect is estimated to be 0.256 with confidence interval [0.235, 0.277] (this does not exactly estimate the average treatment effect because of the regression setup, see equation (2) above). Thus the mindset intervention is indeed effective.

Our results for effect modification are summarized in Figure 1. Notice that although all the methods are plotted in the same figure for the ease of visualization, they may be fitting different linear approximations to $\Delta(\boldsymbol{x})$ and the coefficients for the same covariate may have different meanings. Several covariates (`X1`, `X5`, `XC-4`) are significant using method "marginal" but non-significant using method "full", indicating they may be correlated with the actual effect modifier(s). We find the full model difficult to interpret because it consists of all the covariates. The lasso-regularized regression selects two covariates, `X1` and `XC-3`, as potential effect modifiers, and the application of selective inference shows that `XC-3` is statistically significant even after adjusting for the model selection. In contrast, the "snooping" inference that ignores the bias from model selection would incorrectly declare that `X1` is also statistically significant.

To summarize, our workshop analysis suggests that: `X1` is possibly an effect modifier but more data is possibly needed before a decisive conclusion can be made; `X2` does not moderate the treatment effect; `XC3` is an important effect modifier that the data supports. In fact, with selective inference, we are able to estimate the strength of the effect modifier `XC3` through both interval and point estimates.

3. Post-workshop analysis

3.1 More advanced methods

A major objection to the polyhedral pivot in Lee et al. (2016) is that the selective confidence intervals are often excessively long. For example, in Figure 1 (method "lasso"), the confidence interval of `X1` is very asymmetric: most of the confidence interval lies above 0 but the point estimate is indeed negative. More radical example of this kind can be found

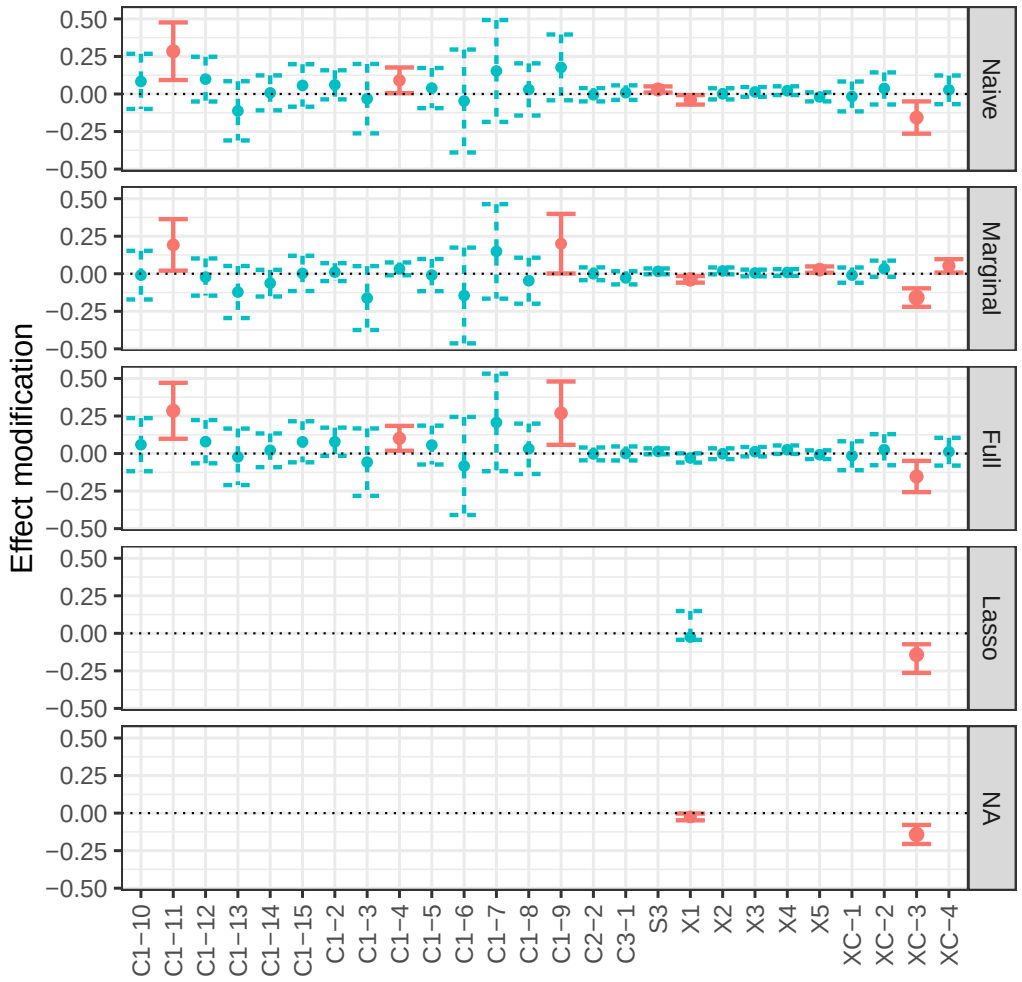

Figure 1: Workshop results: This figure plots the 95% confidence intervals of effect modification by the covariates (red solid intervals do not cover 0).

in Table 1 below. This problem is due to the ill-behavior of the polyhedral pivot when the observed data lies close to the selection boundary. Such phenomenon was observed in the original article by Lee et al. (2016). More recently Kivaranovic and Leeb (2018) has proven that the expected length of the selective confidence interval constructed this way is infinity.

3.1.1 Randomized response

To mitigate this problem, Tian and Taylor (2018) proposed to randomize the response before model selection, thereby smoothing out the selection boundary. This also allows the statistician to reserve more information in the data during the selection stage, leading

to increased power in the inference stage. With moderate amount of injected noise, the increase of inferential power also does not compromise the ability of model selection.

In the effect modification problem, this "randomized lasso" algorithm can be directly applied in Step 2.3 by replacing (3) with the following optimization problem:

$$\text{minimize} \quad \frac{n}{2}\,\mathbb{E}_n\Big[\tilde{Y} - \tilde{Z}(\beta_0 + \boldsymbol{\beta}^T\boldsymbol{X})\Big]^2 + \lambda\|\boldsymbol{\beta}\|_1 - \boldsymbol{\omega}^T\boldsymbol{\beta}, \tag{4}$$

where $\boldsymbol{\omega} \sim \mathcal{N}(0, \tau^2\boldsymbol{I}_p)$ is the injected Gaussian noise. Note that the randomized Lasso has two tuning parameters, one being the amount of ℓ_1-penalty λ and the second one being the amount of injected noise which is measured by τ^2. In our analysis below we will use the same penalty λ as before and set $\tau^2 = \hat{\sigma}^2/2$.

The polyhedral lemma of Lee et al. (2016) no longer applies to randomized lasso because selection now depends on both the data and the injected noise ω. To construct selective confidence intervals after randomized lasso, Tian Harris et al. (2016) proposed to use Monte Carlo and developed a general selective sampler to sample realizations of data truncated to the randomized selection region. To obtain a point estimate of the coefficient, Panigrahi et al. (2016) and Panigrahi and Taylor (2018) introduced the "selection-adjusted" maximum likelihood estimate (selective MLE) that maximizes the conditional likelihood given the selection event. These latest selective inference methods are implemented as Python software available at `https://github.com/selective-inference/Python-software`.

3.1.2 Switching the target of selective inference

In our workshop analysis, the target of selective inference is the partial regression coefficient $\boldsymbol{\beta}_{\hat{\mathcal{M}}}$ defined in (2). Alternatively, one might be interested in the full regression coefficient $\big(\boldsymbol{\beta}_{\{1,2,\ldots,p\}}\big)_{\hat{\mathcal{M}}}$ which contains entries of $\boldsymbol{\beta}_{\{1,2,\ldots,p\}}$ that correspond to the selected covariates $\boldsymbol{X}_{\hat{\mathcal{M}}}$. In other words, instead of targeting all the full regression coefficients as in method "full" above, this approach focuses only on certain selected entries. The selective inference framework in Lee et al. (2016) and Tian Harris et al. (2016) can be effortlessly applied to full regression coefficients because they, like partial regression coefficients, can be written as linear functions of the underlying parameters (in our case $\boldsymbol{\Delta}(\boldsymbol{x})$).

3.1.3 Cross-fitting

Cross-fitting (Schick, 1986; Chernozhukov et al., 2018) is a general algorithm in semiparametric inference to eliminate the dependence of nuisance parameter estimates on the corresponding data point (e.g. dependence of $\hat{\mu}_t(\boldsymbol{X}_i)$ on T_i). In our case, it simply amounts to split the data into two halves and estimating $\mu_t(\boldsymbol{X}_i)$ and $\mu_y(\boldsymbol{X}_i)$ in Step 1 using models trained using the half of the data that does not contain the i-th data point. We implemented this algorithm for our post-workshop analysis. Cross-fitting is useful for proving theoretical properties of the semiparametric estimator. In practice we rarely find that the usage of cross-fitting drastically changes the results.

3.2 Results

Table 1 shows the post-workshop analysis results. There are in total four analyses, targeting partial or full coefficients and using the polyhedral pivot for lasso or selective sampler for

Table 1: Results of different selective inference methods in the Mindset Study dataset.

Target	Selective inference Method	Covariate	Estimate	CI	p-value
Partial	Lasso + polyhedral pivot	C1-11	0.223	[-∞, -2.164]	0.005
		XC-3	-0.141	[-0.156, ∞]	0.234
		X1	-0.025	[-0.042, 0.203]	0.736
	Randomized lasso + sampler	S3	-0.013	[-0.060, 0.017]	0.908
		C1-4	-0.000	[-0.106, 0.082]	0.675
		C1-11	0.121	[-0.339, 0.428]	0.400
		XC-3	-0.151	[-0.265, -0.045]	0.004
		X4	0.002	[-0.063, 0.046]	0.596
Full	Lasso + polyhedral pivot	C1-11	0.284	[-∞, -2.103]	0.005
		XC-3	-0.148	[-0.180, 4.257]	0.256
		X1	-0.031	[-4.345, 0.564]	0.872
	Randomized lasso + sampler	S3	-0.011	[-0.051, 0.016]	0.214
		C1-4	0.061	[-0.092, 0.180]	0.185
		C1-11	0.180	[-0.375, 0.505]	0.568
		XC-3	-0.139	[-0.293, 0.002]	0.052
		X4	0.011	[-0.046, 0.053]	0.819

randomized lasso. Thus the first analysis in Table 1 (lasso + polyhedral pivot) is the same as method "lasso" in Figure 1 besides we used cross-fitting here. The randomized lasso selects three more covariates in the post-workshop analysis. This is typically the case due to the injected noise. However, all selected covariates besides `XC-3` are not statistically significant in the post-selection inference, suggesting that they are probably not effect modifiers.

The biggest advantage of using the randomized lasso and selective sampler is shorter selective confidence interval (CI). For example, For `XC-3`, the CI is reduced from $[-0.156, \infty]$ to $[-0.265, -0.045]$. A careful reader might have noticed that in first row of Table 1, the naive point estimate for `C1-11` obtained by regressing `Y` on the selected covariates—`C1-11`, `XC-3`, and `X1`—is not covered by the CI. This can happen if the data is very close to the decision boundary, see Lee et al. (2016, Fig. 5). The selective MLE point estimates (for randomized lasso) are always covered by the CIs and close to the center of the CIs in Table 1. Switching the inferential target from partial coefficients to full coefficients does not seem to change the results by much. This is likely due to the lack of strong effect modifiers and the lack of dependence between the covariates. In the full model, the covariate `XC-3` is not significant at level 0.05. One possible explanation is that using a selected model often add power to the analysis when the data can be accurately described by a sparse generative model (as opposed to fitting a full model). These observations demonstrate the practical benefits of using the randomized lasso and selective MLE.

4. Discussion

In this paper we have presented a comprehensive yet transparent approach based on Zhao et al. (2017) to analyze treatment effect heterogeneity in observational studies. The same procedure can be applied to randomized experiments as well, and Zhao et al. (2017) has shown that in this case it is sufficient to estimate μ_y consistently in order for the polyhedral

pivot to be asymptotically valid. The proposed procedure can be easily implemented using existing machine learning packages (to estimate μ_t and μ_y) and selective inference softwares. The `R` and `Python` code for our analyses are attached with this report.

We want to re-emphasize some points made in Zhao et al. (2017) about when selective inference is a good approach for analyzing effect modification. Compared to classical statistical analysis, the selective inference framework makes it possible to use the same data to generate new scientific questions and then answer them. This is not useful for the inference of the average treatment effect because it is a deterministic quantity independent of any model selection. However, selective inference can be tremendously useful for effect modification especially when the analyst wants to discover effect modifiers using the data and make some confident conclusions about their effect sizes. We believe that this is indeed the motivation behind the workshop organizers' Question 3, making selective inference a very appealing choice of analyzing datasets like the Mindset Study.

On the other hand, since part of the information in the data is reserved for post-selection inference, the selective inference framework is sub-optimal at making predictions (in our case, estimating $\Delta(\boldsymbol{x})$). There is a long list of literature on estimating the optimal treatment regime or the CATE from the data. This has become a hot topic recently due to the availability of flexible machine learning methods. We refer the reader to Zhao et al. (2017) for some references in this direction. When prediction accuracy is the foremost goal, these machine learning methods should be preferred to selective inference.

Berk et al. (2013) proposed an alternative post-selection procedure that constructs universally valid confidence intervals regardless of the model selection algorithm. However this may be overly conservative when the selection algorithm is pre-specified by the data analyst (for example, the lasso with a fixed λ). Small (2018) discussed connections of this alternative approach to observational studies.

The application in effect modification also suggests new research directions for selective inference. For example, during the workshop several participants attempted to describe the effect modification using decision trees. Results presented in this way are easy to interpret and may have immediate implications in decision making. With the nodes and cutoffs selected in a data-adaptive fashion, this poses yet another post-selection inference problem. Reserving a hold-out data set for a confirmatory analysis on the effects may lead to a loss of power that can be potentially avoided with selective inference. Obtaining optimal inference post exploration via regression trees is an interesting direction for future work.

References

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *Annals of Statistics*, to appear, 2018.

Susan C Athey and Guido W Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao. Valid post-selection inference. *Annals of Statistics*, 41(2):802–837, 2013.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

William Fithian, Dennis Sun, and Jonathan Taylor. Optimal inference after model selection. *arXiv:1410.2597*, 2014.

Danijel Kivaranovic and Hannes Leeb. Expected length of post-model-selection confidence intervals conditional on polyhedral constraints. *arXiv:1803.01665*, 2018.

Jason D Lee, Dennis L Sun, Yuekai Sun, and Jonathan E Taylor. Exact post-selection inference, with application to the lasso. *Annals of Statistics*, 44(3):907–927, 2016.

Sahand Negahban, Bin Yu, Martin J Wainwright, and Pradeep K Ravikumar. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

Snigdha Panigrahi and Jonathan Taylor. Scalable methods for bayesian selective inference. *Electronic Journal of Statistics*, 12(2):2355–2400, 2018.

Snigdha Panigrahi, Jonathan Taylor, and Asaf Weinstein. Bayesian post-selection inference in the linear model. *arXiv:1605.08824*, 2016.

Peter M Robinson. Root-n-consistent semiparametric regression. *Econometrica*, 56(4):931–954, 1988.

Anton Schick. On asymptotically efficient estimation in semiparametric models. *Annals of Statistics*, 14(3):1139–1151, 1986.

Dylan S Small. Larry Brown: Remembrance and connections of his work to observational studies. *Observational Studies*, 4:250–259, 2018.

Jonathan Taylor and Robert J Tibshirani. Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112(25):7629–7634, 2015.

Xiaoying Tian and Jonathan Taylor. Selective inference with a randomized response. *Annals of Statistics*, 46(2):679–710, 2018.

Xiaoying Tian Harris, Snigdha Panigrahi, Jelena Markovic, Nan Bi, and Jonathan Taylor. Selective sampling after solving a convex problem. *arXiv:1609.05609*, 2016.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

Ryan Tibshirani, Rob Tibshirani, Jonathan Taylor, Joshua Loftus, and Stephen Reid. *selectiveInference: Tools for Post-Selection Inference*, 2017. URL `https://CRAN.R-project.org/package=selectiveInference`. R package version 1.2.4.

Mark J. van der Laan and Sherri Rose. *Targeted Learning*. Springer, 2011.

Qingyuan Zhao, Dylan S Small, and Ashkan Ertefaie. Selective inference for effect modification via the lasso. *arXiv:1705.08020*, 2017.