Efficient Discovery of Heterogeneous Treatment Effects in Randomized Experiments via Anomalous Pattern Detection

Edward McFowland III[*] Sriram Somanchi[†] Daniel B. Neill[‡]

The randomized experiment is employed heavily across the social sciences as an important tool for scientific discovery, by enabling causal inference–i.e., estimating the causal impact of a particular stimulus, treatment or intervention. It is clear that we now live in a world awash with large quantities of fine-grained data, and following this trend, there is a growing interest in, and ability to conduct, large experiments. For example, web-facing organizations–e.g, Google, Microsoft, Amazon, Facebook, etc.–conduct hundreds of large-scale online experiments daily to measure advertisement effectiveness, guide product development, expedite service adoption, and understand user behaviors [11, 10]. Therefore, in the work we propose an algorithm to enable such organizations–and social science more broadly–to efficiently identify which subpopulations, in a randomized experiment, have the largest (distributional) treatment effects.

The increasing popularity of large-scale experiments has also resulted in widespread interest in discovering heterogeneous treatment effects. The sample population is typically diverse and therefore may experience widely varying responses to an intervention. A portion of this variability will be a result of random noise, while some portion of it results from systematic differences in the population, potentially captured in the observed covariates. The ability to *discover* this systematic variability is of great interest as it can lead to treatment personalization and improved outcomes. However, there is equally great challenge, as there are exponentially many subpopulations to consider, which may result in poor estimation, overfitting noise vis-à-vis multiple hypothesis testing, and simply may be computationally intractable. As a result, there is an increasing vibrant literature using machine learning methods to provide data-driven approaches for modeling and *estimating* heterogenous treatment effects in randomized experiments. One category of methods include those based on sparse (regularized) regression models [14, 8, 13, 16]. However, these methods require the researcher to specify which covariates and treatment interactions should be included in the model, which compromise their ability to truly *discover* previously unknown treatment patterns in subpopulations. If there is no theory or prior knowledge guiding a precise specification, it is not uncommon for researchers to attempt multiple specifications, which can quickly devolve into a unprincipled search. In an attempt to learn a flexible nonparametric data-driven model, the literature has also proposed the use of regression trees [9, 12, 1]. Although a regression tree can adaptively approximate even complex functions, its effectiveness can be severely compromised in many settings as a result of its greedy partitioning: tree models can be unstable and can struggle to estimate functions where a small proportion of the covariates (or covariate values) constitute the influential interactions [3]. There have been subsequent improvements on the single tree model which propose the use of ensemble methods for treatment effect estimation including Bayesian Additive Regression Trees (BART) [6, 4, 5], Random Forrest [2, 15], both of which improve upon the single regression tree by combining the predictions of a collection of trees. However, ensemble methods lose interoperability of natural groupings (e.g., specific combinations of covariates or clearly defined leaves) which is important for identifying affected subpopulations.

In addition to these category specific challenges of the individual methods, there are additional common limitations that are present throughout the literature; therefore, we design a method Treatment Effect Subset Scan (TESS) in an attempt to fill the gaps. For example, there is essentially a uniform focus on estimating the average treatment effect, defined as the average difference between the treatment and control outcome distributions. Test based on investigating the first moment of the distributions may not capture the entire effect, and in some cases fail to identify an effect at all, as they they will have little power against an alternative that has changes in higher order moments. Therefore TESS

[*]University of Minnesota
[†]University of Notre Dame
[‡]Carnegie Mellon University

is powered for treatment effects that arbitrarily affect the distribution of outcomes between the control and treatment group. Beyond the literature's focus on treatment effects as mean shifts, there seems to be no focus on how to *discover* the most interesting or extreme subpopulations: typically, after learning a model to represent the data, the researcher is required to preform manual model inspection or manual querying for the subpopulations with the largest treatment effects. Moreover, only the subpopulations defined by the pre-specified regression or picked out by the leafs of a tree are investigable; additionally, only a minority of algorithms provide a sense of the statistical significance for the discovered subpopulation. Therefore, TESS is designed to *discover* the subpopulations with the most statistically significant effects, while searching over the space of all subpopulations parsimoniously.

The similarity in the current literature stems directly from the continued utilization, or modification, of standard statistical machine learning methods to produce a model that estimates the average treatment effect as a function of the observed covariates. At the core of these methods is an attempt to optimize statistical risk–i.e., minimizing the expected loss–for the *entire* response surface, as a function of the observable covariates. Conversely, TESS is based on Anomalous Pattern Detection [7] and therefore optimizes effect maximization. Unknowingly, minimizing risk can be a impediment to maximizing effect, as flexible model (e.g., a tree) may exclude the most interesting subpopulation from being considered or explicitly estimated. In some contexts learning a "good" overall model of the treatment effect response surface is desirable; however, in many cases, the identification of potentially novel subpopulations is more critical while model learning is simply a (presumably necessary) step toward this goal. For these cases it seems more prudent and efficient to circumvent this first step–and therefore all of the restrictions, challenges, and assumptions imposed by it–and solve the subpopulation identification problem by framing it as one of discovery. Therefore, our Treatment Effect Subset Scan algorithm is designed to efficiently maximize a goodness-of-fit statistic over the space of all subpopulations, allowing it to find the subpopulation whose distribution of outcomes is the most changed as a result of a treatment. Furthermore, goodness-of-fit statistics are parameterized by the empirical distribution functions of the data, which imposes minimal restrictions on the underlying data generating process.

In addition to the proposal of a novel algorithm, through theorems (and supporting lemmas) we provide theoretical guarantees for our procedure. We begin by defining $F(S)$ as the score or the distributional effect (i.e., the distributional divergence between treatment and control groups) of subpopulation S. Additionally, S^* is the subpopulation identified by TESS as the most affected by the treatment–i.e., $S^* = \arg\max_S F(S)$–and $t(n, \alpha)$ is some critical value based on the number of experimental units n and a desired (hypothesis) test level α. When rejection of the null hypothesis–i.e., the treatment has no affect on any subpopulation–occurs if $F(S^*) > t(n, \alpha)$, then we demonstrate that $\lim_{n\to\infty} P_{H_0}\left(Reject\ H_0\right) \leq \alpha$ and $\lim_{n\to\infty} P_{H_1}\left(Reject\ H_0\right) = 1$ for a particular class of "reasonable alternatives". These results imply that TESS provides valid hypothesis testing, with full asymptotic power. Additionally, we provide guarantees on the optimality of our efficient (i.e. non-exhaustive) search over subpopulations and sufficient conditions for guaranteeing the optimality of the identified subpopulation. More precisely, if S^T is the subpopulation truly the most affected by the treatment, we provide conditions such that $S^* \supseteq S^T$ and $S^* \subseteq S^T$, which when combined implies $S^* = S^T$.

In addition to this paper's methodological and theoretical contributions, we also evaluate the efficacy of TESS through simulations and an exploratory analysis on real-world data. More specifically, we use TESS to explore data from the well-known Tennessee Star program evaluation study [17], and find an intuitive subpopulation that appears to experience a significant effect from a treatment that was ruled ineffective [17]. The Tennessee Star study explored the effects of classroom size–i.e., small, regular, and regular with an additional teachers aide–on student test scores from kindergarten to third grade. Although there was strong evidence of the benefit of small classrooms, there is no such evidence for regular classrooms with a teacher aide. However, TESS discovered that a subpopulation of students–i.e, those in second or third grade, in inner-city or urban schools, whose teacher had more than ten years of experience–exhibited large benefits from having a teachers aide present. On a potentially related note, second and third grade where the years when a random subset of teachers (and their aids if applicable) received additional summer training; this training was also deemed ineffective [17]. Although the focal point of our work is not estimating education production functions, it does present evidence that a previously believed ineffective treatment may have been effective for a particularly vulnerable subpopulation. This identified subpopulation is intuitive and believable, but clearly should investigated more directly and rigorously by researcher specifically interested in education production functions. This result, however, does provide a sense of how TESS can be utilized as a tool for generating hypothesis, to be further explored and tested. In many contexts it is rare to know a priori which hypothesis are relevant and supported by

data, and the use of traditional methods (e.g., regression) would suffer from many of the limitations described above, including potentially excluding the significant populations from even being considered. Our Treatment Effect Subset Scan algorithm improves upon these limitations by using Anomalous Pattern Detection to efficiently discover the subpopulations that are highly affected by a given treatment, bringing them to the attention of the researcher.

References

[1] S. Athey and G. Imbens. Recursive partitioning for heterogeneous causal effects: Table 1. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, Jul 2016.

[2] J. C. Foster, J. M. G. Taylor, and S. J. Ruberg. Subgroup identification from randomized clinical trial data. *Statistics in medicine*, 30(24):2867–2880, Aug 2011.

[3] J. H. Friedman. Multivariate Adaptive Regression Splines. 19(1):1–67, Mar 1991.

[4] D. P. Green and H. L. Kern. Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees. *Public opinion quarterly*, 76(3):491–511, Sep 2012.

[5] J. Grimmer, S. Messing, and S. J. Westwood. Estimating Heterogeneous Treatment Effects and the Effects of Heterogeneous Treatments with Ensemble Methods. 2013.

[6] J. L. Hill. Bayesian Nonparametric Modeling for Causal Inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, Jan 2011.

[7] E. M. III, S. Speakman, and D. B. Neill. Fast generalized subset scan for anomalous pattern detection. *The Journal of Machine Learning Research*, 14(1):1533–1561, Jun 2013.

[8] K. Imai and M. Ratkovic. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470, Mar 2013.

[9] K. Imai and A. Strauss. Estimation of Heterogeneous Treatment Effects from Randomized Experiments, with Application to the Optimal Planning of the Get-Out-the-Vote Campaign. *Political Analysis*, 19(1):1–19, Jan 2011.

[10] R. Kohavi, A. Deng, B. Frasca, T. Walker, Y. Xu, and N. Pohlmann. Online controlled experiments at large scale. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, New York, NY, USA, 2013. ACM.

[11] R. Kohavi, R. Longbotham, D. Sommerfield, and R. M. Henne. Controlled experiments on the web: survey and practical guide. *Data Mining and Knowledge Discovery*, 18(1):140–181, jul 2009.

[12] X. Su, C.-L. Tsai, H. Wang, D. M. Nickerson, and B. Li. Subgroup Analysis via Recursive Partitioning. *The Journal of Machine Learning Research*, 10:141–158, Dec 2009.

[13] L. Tian, A. A. Alizadeh, A. J. Gentles, and R. Tibshirani. A Simple Method for Estimating Interactions Between a Treatment and a Large Number of Covariates. *Journal of the American Statistical Association*, 109(508):1517–1532, Dec 2014.

[14] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996.

[15] S. Wager and S. Athey. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *arXiv.org*, Oct 2015.

[16] H. I. Weisberg and V. P. Pontes. Post hoc subgroups in clinical trials: Anathema or analytics? *Clinical trials*, 12(4):357–364, Aug 2015.

[17] E. Word, J. Johnston, H. P. Bain, B. D. Fulton, J. B. Zaharias, C. M. Achilles, M. N. Lintz, J. Folger, and C. Breda. The State of Tennessee's Student/Teacher Achievement Ratio (STAR) Project. *Nashville Tennessee State Department of Education*, pages 1–30, 1990.