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

Edward McFowland III  
Information and Decision Sciences,  
Carlson School of Management,  
University of Minnesota

Sriram Somanchi  
IT, Analytics, and Operations,  
Mendoza College of Business,  
University of Notre Dame

Daniel B. Neill  
Event and Pattern Detection Laboratory,  
H.J. Heinz III College,  
Carnegie Mellon University

Abstract

In the recent literature on estimating heterogeneous treatment effects, each proposed method makes its own set of restrictive assumptions about the intervention’s effects and which subpopulations to explicitly estimate. Moreover, the majority of the literature provides no mechanism to identify which subpopulations are the most affected—beyond manual inspection—and provides little guarantee on the correctness of the identified subpopulations. Therefore, we propose Treatment Effect Subset Scan (TESS), a new method for discovering which subpopulation in a randomized experiment is most significantly affected by a treatment. We frame this challenge as a pattern detection problem where we efficiently maximize a nonparametric scan statistic over subpopulations. Furthermore, we identify the subpopulation which experiences the largest distributional change as a result of the intervention, while making minimal assumptions about the intervention’s effects or the underlying data generating process. In addition to the algorithm, we demonstrate that the asymptotic Type I and II error can be controlled, and provide sufficient conditions for detection consistency—i.e., exact identification of the affected subpopulation. Finally, we validate the efficacy of the method by discovering heterogeneous treatment effects in simulations and in real-world data from a well-known program evaluation study.

Keywords: causal inference, program evaluation, algorithms, distributional average treatment effect, treatment effect subset scan, heterogeneous treatment effects
1 Introduction

The randomized experiment is employed across many empirical scientific disciplines as an important tool for scientific discovery, by estimating the causal impact of a particular stimulus, treatment or intervention. From bioinformatics to behavioral economics, large-scale experiments are being used for data-driven discovery of new biological phenomena \[1, 2\] and to inform policy in areas including poverty, education, health, microfinance, and governance \[14\]. Furthermore, web-facing organizations—e.g., Google, Microsoft, Amazon, Facebook, and eBay—conduct hundreds of large-scale online experiments daily to measure advertisement effectiveness, guide product development, expedite service adoption, and understand user behaviors \[23\]. The increasing popularity of large-scale experiments has resulted in a widespread interest in discovering more fine-grained truths about experimental units, most prominently in the form of heterogeneous treatment effects.

Heterogeneous treatment effects (HTE) describe the variability in individuals’ response to an intervention within the sampled population. A portion of this variability may result from systematic differences in the population, potentially captured in the observed covariates. Each combination of covariate values defines a characteristic profile, and a collection of profiles represents a subpopulation—e.g., the subpopulation gender = “male” or the more specific subpopulation gender = “female” & race = “A”. Discovering heterogeneity can be challenging because there are exponentially many subpopulations—with respect to the number of observable covariates—to consider, potentially resulting in multiple hypothesis testing issues and raising questions of unprincipled post-hoc investigation: searching for a fortuitously statistically significant result \[7, 40\]. Although these challenges and concerns are valid, uncovering affected subpopulations can lead to important scientific progress. In a “step toward a new frontier of personalized medicine” \[29\] the FDA approved the first race-specific drug, whose impact on African-American subjects was first discovered post-hoc from more general experiments \[12, 13\]. Conversely, the Perry preschool experiment found extremely significant educational and life outcomes for pre-school education \[4, 9, 30\], while a re-analysis focused on heterogeneity and multiple hypothesis testing concludes that only girls experience these benefits \[3\]. The original Perry preschool results were fundamental to the creation of the Head Start pre-school program \[4\], a national social program that provides, among other services, early childhood education to low-income children. If large-scale medical and policy decisions are made as a result of such experiments, then it is clear that identifying whether there is heterogeneity in treatment effects should be an integral component of the analysis.

In this work we propose a novel computationally efficient framework—Treatment Effect Subset Scanning (TESS)—for discovering which subpopulations in a randomized experiment are the most significantly affected by a treatment. The contributions of this work can be summarized as follows:

- Our TESS algorithm enables efficient discovery of subpopulations where the individuals affected by the treatment have observed outcome distributions that are unexpected given the distributions of their corresponding control groups. Unlike the standard approaches, TESS frames the challenge of discovering causal effects in subpopulations as one of *anomalous pattern detection*, and provides a computationally efficient approach for finding conditionally optimal subpopulations. Therefore, TESS is a novel contribution to the burgeoning literature on subset scanning, providing conditions under which the linear-time subset scanning property \[25\] can be exploited in the context of high-dimensional tensors, and thus extending the applicability of this efficient optimization approach beyond the standard low-dimensional context \[25, 26, 32\].

- We formalize the objective of identifying subpopulations with significant treatment effects by proposing a new treatment effect estimand (§3). This estimand allows for identification of nuanced distributional treatment effects, as opposed to standard mean shifts, and contains the popular HTE estimands as special cases.

- We provide theoretical results on the detection properties of TESS. When the maximum subpopulation score identified by TESS is used as a test statistic for the presence of HTEs, we demonstrate the conditions under which the Type I (Theorem 2) and Type II (Theorem 3) errors can jointly be controlled asymptotically. Furthermore, we provide sufficient conditions on how “homogeneous” (Theorem 4) and
“strong” (Theorem 5) the treatment effect must be across the affected subpopulation, such that TESS is guaranteed to detect the precisely correct subpopulation.

• In the process of developing theory for TESS, we prove results for the general nonparametric scan statistic (NPSS), which has been used in the scan statistics literature [11, 25]. We are the first to provide theoretical guarantees on the detection behavior of subset scanning algorithms. Furthermore, our theory is derived for the high dimensional (tensor) context, with nonparametric score functions, and our results directly hold for the lower-dimensional and parametric cases as well.

• Our empirical results (§5.4) provide useful insights to practitioners, revealing a potentially affected subpopulation in the Tennessee STAR study of class size and educational outcomes, for a treatment condition (the use of a teacher’s aide in a class of regular size) that was generally considered ineffective.

These contributions are enabled by structuring the question of causal inference as one of anomalous pattern detection (and effect maximization), rather than model fitting (and risk minimization). In some contexts, the standard approach of learning an overall good model of the treatment effect response surface is desirable; however, in many cases, the identification of affected subpopulations is the primary goal and model learning is simply a step toward this goal. For these cases it seems more prudent and efficient to circumvent this first step and solve the subpopulation identification problem by framing it as one of pattern or subset discovery. Such a framing has not previously been considered in the literature.

The remainder of this work begins with a review of recent statistical learning methods to estimate heterogeneous treatment effects (HTE), and an outline of general gaps that exist in the literature (Section 2). In Section 3 we propose a new class of causal estimands which generalizes the common HTE estimands and helps to address their limitations. Section 4 presents our computationally efficient TESS algorithm, which can identify the subpopulation that experiences the largest distributional change as a result of the treatment, while disregarding provably sub-optimal subpopulations. Additionally, we demonstrate that the probability of committing Type I and II errors can be bounded asymptotically, and we provide sufficient conditions under which our framework will discover the exact subpopulation of interest (§4.5). In Section 5 we demonstrate empirically that our framework exhibits significantly more power to detect subtle signals than current methods, while also providing more precise characterization of the affected subpopulation. We then use TESS to conduct an exploratory analysis of the well-known Tennessee STAR [41] study, discovering previously unidentified treatment effect heterogeneity. Section 6 concludes the paper.

2 Related Work

Heterogeneity in treatment effects is studied across many disciplines. The typical approach is to specify a model of the relationship between variables (usually, linear regression) based on theory or intuition, estimate parameters of the model from data, and test the statistical significance of these parameters. When there is interest in identifying treatment heterogeneity, the researcher is expected to pre-specify the model with the form of heterogeneity included. In the absence of sufficient prior knowledge to guide the precise model specification, it is common to attempt multiple specifications and tests, which can quickly devolve into an unprincipled search. In response, some medical and social science disciplines require pre-analysis plans, which can impede the knowledge discovery process. We argue that these challenges necessitate new data-driven tools that enable the discovery of unknown, and possibly subtle, treatment effects in subpopulations, while avoiding the pitfalls created by multiple testing and post-hoc analysis.

There has been a growing literature using statistical learning methods to provide data-driven approaches for estimating heterogeneous treatment effects in randomized experiments, including both sparse (regularized) regression models and tree-based methods. Recent work has adapted regularization to the causal setting and specifically to treatment effect heterogeneity [21, 35, 10], proposing methods that frame the treatment effect estimation problem as one of L1-regularized (LASSO) model selection [36]. Although these regularized regression methods select and estimate the importance of covariates, they are still subject to the possibly restrictive assumptions and limitations of (linear) regression, including requiring the researcher to specify
which covariate and treatment interactions to include, compromising their ability to discover unexpected treatment patterns in subpopulations.

Other methods [8, 34] select subpopulations and estimate treatment effects using the well-known regression tree, which recursively partitions the data into homogeneous subpopulations that share a subset of covariate profile values and have similar outcomes. 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; they can provide extremely discontinuous approximations of an underlying smooth function, limiting overall accuracy; and they can struggle to estimate functions which exhibit specific properties, including when a small proportion of the covariates constitute the influential interactions [17].

Subsequent improvements on the single tree model propose the use of ensemble methods for treatment effect estimation: including the use of Bayesian Additive Regression Trees [19] and Random Forests [39]. [20] observes that the machine learning methods proposed for model selection each make implicit modeling assumptions whose validity will vary given the specific problem context. Therefore, the authors propose a general ensemble method that brings together various models, where the weights of their estimates are learned from cross-validation. Combining the predictions of multiple models provide more stable and smooth function estimates [39]; however, they lose the interpretability of natural groupings (e.g., specific combinations of covariates or clearly defined leaves) which is important for identifying affected subpopulations.

2.1 Addressing limitations of the prior literature

Our proposed methodology for Treatment Effect Subset Scanning (TESS) differs substantially from the prior literature in two main aspects: identifying general changes in distribution (or specific quantiles) rather than mean shifts, and a focus on detecting the subpopulations most significantly affected by treatment rather than estimating treatment effects for all individuals. First, the stated objective of the majority of methods in the current literature is estimating the average treatment effect (ATE), for the population or some subpopulations. The ATE measures the difference between the means of the treatment and control outcome distributions but cannot identify other changes in distribution. Anscombe’s quartet [5] is a classic example of datasets which have very different distributions but identical first and second moments. In such cases, the ATE will fail to identify an effect, leading to incorrect assumptions about the similarity of the distributions. In other cases, a treatment (such as a policy change which impacts only the very rich or the very poor) may substantially affect various quantile values of the distribution with only slight shifts in the mean. In such cases, estimating the ATE would have low power to identify these distributional changes.

Second, the prior literature on heterogeneous treatment effects is primarily focused on estimating the treatment effect for each individual or for a small set of manually defined subpopulations (e.g., estimating separate effects for males vs. females). To the best of our knowledge, none of these approaches provide a mechanism to automatically detect which subpopulations exhibit the most significant treatment effects. Our TESS framework is explicitly designed for subpopulation discovery, with the twin goals of maximizing 1) detection power, the ability to distinguish between experiments with a subtle heterogeneous treatment effect and those with no treatment effect, and 2) detection accuracy, the ability to precisely identify the affected subpopulation. This allows us to provide theoretical guarantees on the results of discovery as well as improving both detection power and accuracy in practice.

In contrast, the prior literature can be roughly divided into three groups. Methods such as [39] produce separate estimates of the treatment effect for each individual (or set of individuals who are identical on all observed covariates). While such methods can produce a list of treated individuals ranked by estimated treatment effects, this provides little continuity across individuals, with no principled way to identify affected subpopulations or to distinguish significant HTEs from noise. Manually grouping highly affected individuals can easily lead to false positives and incorrect generalizations, as well as low power to detect subtle effects across multiple covariate profiles.

Regression-based methods such as [21, 35] allow manual inspection of the coefficients for each covariate interacted with the treatment dummy. However, such approaches typically assume a small number of prespecified interaction terms and cannot identify other affected subpopulations. The extreme alternative of
adjusting for each subpopulation separately, including a term for every possible combination of covariates interacted with the treatment, would require exponentially many interaction terms, leading to computational intractability as well as statistical challenges (lack of power and multiple testing).

Finally, methods such as causal trees [3] and interaction trees [34] use a greedy top-down approach to create specific partitions of the covariate space (the leaves of the tree) that can be interpreted as subpopulations, enabling manual or automatic identification of those partitions with the largest treatment effects. However, when the affected subpopulation and effect size are small, we do not expect the resulting partitions to correspond well to the subpopulation of interest, since the approach optimizes a global objective function such as statistical risk (average loss) rather than focusing on the most significantly affected subpopulations. This difference in emphasis may allow the tree to precisely estimate treatment effects across the entire population (including effects which are near zero) but have larger errors for the small and significantly affected subpopulations we wish to detect. Poor choice of partitions could exclude the affected subpopulation from being considered or identified, and instead estimate an effect which is the average over this subpopulation of interest and others. These aspects lead to reduced detection power and accuracy in practice, as shown in our results below. Moreover, the instability of tree-based methods may call into question the relevance of the tree-selected subpopulations, while extensions to random forest-based approaches [39] sacrifice the ability to identify subpopulations for more stable and more accurate estimation of individual treatment effects.

In summary, the current state of the heterogeneous treatment effects literature has many gaps: the only effects of interest are mean shifts, these effects are estimated under possibly restrictive modeling assumptions, only a subset of possible subpopulations are considered and represented, discovering the subpopulation with the largest effect requires manual inspection or an exhaustive search over all modeled subpopulations, and there is little guarantee of the optimality of the discovered subpopulations. In contrast, our proposed TESS approach directly searches for the most significantly affected subpopulations, where significance is measured based on the divergence between the empirical distributions of the treatment and control data, thus avoiding restrictive modeling assumptions. We derive statistical theory which provides performance guarantees, and demonstrate state-of-the-art empirical performance on both real and simulated data, as described below.

3 Framework for Distributional Causal Inference

The Treatment Effect Subset Scan framework builds on the widely studied potential outcomes framework (Neyman-Rubin Causal Model), with random treatment assignment, enabling valid causal statements. More precisely, it begins with observing \(N\), a sample of \(n\) independent and identically distributed units from a population of interest \(P\). The units are indexed by \(i \in \{1, \ldots, n\}\), and for each unit there is a binary assignment indicator \(W_i \in \{0, 1\}\), where \(W_i = 0\) indicates assignment to the control group (i.e., the group that did not receive the treatment), while \(W_i = 1\) indicates assignment to the treatment group. Therefore, there exist two potential outcomes for each unit \((Y_i(0), Y_i(1) \in \mathbb{R})\), although only one of these two potential outcomes is observed for each unit. Additionally, each unit is described by \(X_i\), a \(d\)-dimensional vector of covariates which are fixed, known, and unaffected by treatment assignment. Given this sample, we wish to perform causal inference for the (potentially infinite) population \(P\). In particular, there is interest in a causal population estimand that is a function of the potential outcome distributions and covariates:

\[
\tau = \tau(F_{Y(1)}, F_{Y(0)} | X),
\]

which can be approximated with estimators of the finite sample. In particular, we follow the literature and consider finite sample estimators that can be described as row-exchangeable functions of the potential outcomes, treatment assignments, and covariates, for all of the units in \(N\). More specifically, we consider \(\hat{\tau} = \hat{\tau}(Y(0), Y(1), X, W)\), where \(Y(0)\) and \(Y(1)\) are the \(n\)-dimensional column vectors of potential outcomes, \(W\) is the \(n \times d\) matrix of covariates, all of which are indexed by sample units \(i\). In the following subsections we will describe causal estimands that are common in the literature and present the new causal estimands and estimators at the core of our TESS framework.
3.1 Causal Estimands in the Literature

In the heterogeneous treatment effect literature, the most flexible estimand considered thus far is the marginal conditional average treatment effect (MCATE) \( \tau_{\text{MCATE}} \), defined as

\[
\tau_{\text{MCATE}}(x^*) = \int \left( \int y \ dF_{Y(1)|X^*}(y|x^*) - \int y \ dF_{Y(0)|X^*}(y|x^*) \right) dF_{X^*|X^*=x^*}
\]

\[
= \int \mathbb{E} \left[ Y(1) - Y(0) \mid (X^1, X^2, \ldots, X^s = x^s, \ldots, X^d) \right] dF_{X^*|X^*=x^*}
\]

\[
= \mathbb{E} [Y(1) - Y(0) \mid X^* = x^*].
\]

The MCATE estimates the expected difference in potential outcomes for the specific subset \( x^* \) of the covariate profile, marginalized over the remaining unfixed covariates. The MCATE generalizes the prevalent estimands in the literature: the average treatment effect (ATE), \( \tau_{\text{ATE}} = \mathbb{E}[Y(1) - Y(0)] \), and conditional average treatment effect (CATE), \( \tau_{\text{CATE}}(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x] \). Essentially, the MCATE is a weighted average of the CATEs that include \( X^* = x^* \), weighted by the conditional distribution of the remaining unfixed covariates.

3.2 A Distributional Average Treatment Effect Estimand Class

Although MCATE generalizes other estimands from the literature, it is limited to estimating the ATE for a particular covariate profile \( X^* = x^* \). In order to provide a population-level measurement that can capture more general changes in the outcome distribution resulting from treatment, we generalize MCATE to the distributional average treatment effect (DATE), a new class of treatment effect estimands. First, for a given covariate profile \( X = x \), we define \( \tau_{\text{DATE}}(x) \) as an arbitrary function of the cumulative distribution functions (cdfs) of the potential outcomes \( Y(1) \) and \( Y(0) \) given \( X = x \): \( \tau_{\text{DATE}}(x) \equiv \tau(F_{Y(1)|X=x}, F_{Y(0)|X=x}) \) is a scalar which captures the individual-level treatment effect for a covariate profile \( x \). For a set of covariate profiles \( S \), we define \( \tau_{\text{DATE}}(S) \) as a weighted average over the individual profiles:

\[
\tau_{\text{DATE}}(S) = \int_{x \in S} \tau_{\text{DATE}}(x) P(X = x \mid X \in S) \ dx. \tag{1}
\]

We note that \( \tau_{\text{CATE}}(x) \) and \( \tau_{\text{MCATE}}(x^*) \) are special cases of \( \tau_{\text{DATE}}(x) \) and \( \tau_{\text{DATE}}(S) \) respectively, with function \( \tau(F_1, F_0) = \int y dF_1(y) - \int y dF_0(y) \), the difference between the means of the two cdfs, and \( S = \{X : X^* = x^*\} \) consisting of those profiles with the given values for the subset of covariates \( X^* \). Although the DATE class includes CATE and MCATE, it provides more flexible estimation of HTEs by allowing other specifications of \( \tau(F_1, F_0) \) that capture arbitrary comparisons between the potential outcome distributions. DATE also provides a flexible definition of subpopulations, considering an arbitrary set \( S \) of covariate profiles, while MCATE only considers a single value \( x \) for each covariate \( X \in X^* \) and all values for each covariate \( X \in X^{\sim} \). Here we consider subsets \( S \) representing subspaces of the attribute space, i.e., the Cartesian product of a subset of values for each attribute. This is important because a treatment of interest may affect multiple values, e.g., African-Americans or Hispanics who live in New York or Pennsylvania.

While \( \tau_{\text{DATE}}(S) \) is useful to estimate for a given subset \( S \), our primary goal is to identify those subsets which have the most significant treatment effects. To do so, we need a model \( H_0 \) of how the data is generated under the null hypothesis of no treatment effect, i.e., \( F_{Y(1)|X=x} = F_{Y(0)|X=x} \) for all \( x \), and a general measure of divergence, \( \text{Div} : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R} \), where \( \text{Div}(u, v) \geq 0 \) for all \( u, v \) and \( \text{Div}(u, u) = 0 \) for all \( u \). We then define \( \mu_{\text{DATE}}(S) \) to represent the divergence between \( \tau_{\text{DATE}}(S) \) and its expected value under \( H_0 \):

\[
\mu_{\text{DATE}}(S) = \text{Div}(\tau_{\text{DATE}}(S), \mathbb{E}_{H_0}[\tau_{\text{DATE}}(S)]). \tag{2}
\]

For MCATE, we have \( \mathbb{E}_{H_0}[\tau_{\text{MCATE}}(S)] = 0 \), and thus \( \mu_{\text{MCATE}}(S) = \text{Div}(\tau_{\text{MCATE}}(S), 0) \). As described in (3.3), the choice of divergence function depends on our assumptions about the data distribution under both null and alternative hypotheses.
3.3 The Nonparametric Average Treatment Effect Estimand

The class of estimands defined by the DATE is large; therefore, we select an instance of this class—which we define as the nonparametric average treatment effect (NATE)—that utilizes the flexibility provided by the DATE to evaluate a general divergence between two potential outcome distributions. We first define

\[
\tau_{\text{NATE,}_a}(x) = \beta_x(\alpha)
\]

\[
:= F_{Y(1)|X=x}\left( F_{Y(0)|X=x}^{-1}(\alpha) \right),
\]

which maps the quantile value \( \alpha \) of the control potential outcome distribution into the corresponding quantile value \( \beta \) of the treatment potential outcome distribution. The corresponding \( \tau_{\text{NATE,}_a}(S) \) and \( \mu_{\text{NATE,}_a}(S) \) are defined as in (1) and (2) respectively. Under \( H_0 \) the potential outcomes are equal, thus \( \tau_{\text{NATE,}_a}(x) = F_{Y(0)|X=x}\left( F_{Y(0)|X=x}^{-1}(\alpha) \right) = \alpha \), and

\[
\mu_{\text{NATE,}_a}(S) = \text{Div} \left( \int_{x \in S} \beta_x(\alpha)P(X = x \mid X \in S) \, dx, \alpha \right).
\]

Intuitively, \( \mu_{\text{NATE,}_a}(S) \) is a comparison between potential outcome distribution functions, localized to a specific subpopulation \( S \) and quantile value \( \alpha \) of the null distribution.

We also consider the quantity \( \mu_{\text{NATE}}(S) = \max_{\alpha} \mu_{\text{NATE,}_a}(S) \), which maximizes the divergence between treatment and control potential outcome distributions over a desired range of quantile values \( \alpha \). This estimand will identify arbitrary effects of a treatment, over general subpopulations, measured by the maximal divergence between potential outcome distributions. An additional, and critical, component of NATE is allowing the different covariate profiles \( x \in S \) to have different reference distributions: the distribution \( F_{Y(0)|X=x} \) serves as the expectation for the corresponding distribution \( F_{Y(1)|X=x} \). We note that the alternative approach of using a single reference distribution, aggregated from all controls in \( S \), fails when the different covariate profiles being aggregated have different outcome distributions. In this case, marginalization could obfuscate, or even reverse, the true effects that are occurring in these covariate profiles: this phenomena is commonly known as Simpson’s Paradox. NATE avoids this paradox by evaluating the relationship between the outcome distributions for individual covariate profiles before aggregating across the subpopulation.

3.4 Causal Estimators

Given that the estimands \( \tau_{\text{DATE}} \) and \( \mu_{\text{DATE}} \) are defined in terms of the cdfs \( F_{Y(T)|X=x}, T \in \{0, 1\} \), we consider the corresponding finite sample estimators, \( \hat{\tau}_{\text{DATE}} \) and \( \hat{\mu}_{\text{DATE}} \). We assume that each individual unit \( Q_i \) is drawn i.i.d. from \( P \). A unit can be represented by a 4-tuple, \( Q_i = (X_i, Y_i(0), Y_i(1), W_i) \), where \( X_i \) are covariates, \( Y_i(0) \) and \( Y_i(1) \) represent that unit’s potential outcomes under control and treatment conditions respectively, and \( W_i \in \{0, 1\} \) is the treatment indicator. We note that \( Y_i^{\text{obs}} = Y_i(W_i) \) is the unit’s observed outcome, while \( Y_i(1 - W_i) \) is unobserved. We define

\[
\hat{F}_{Y|C=X=x}(y) = \frac{\sum_{Q_i: X_i=x} \mathbb{1}\{W_i = 0\} \mathbb{1}\{Y_i^{\text{obs}} \leq y\}}{\sum_{Q_i: X_i=x} \mathbb{1}\{W_i = 0\}},
\]

and \( \hat{F}_{Y|T=X=x}(y) \) similarly for units with \( W_i = 1 \). These represent the empirical cumulative distribution functions of \( Y_i^{\text{obs}} \) for control individuals with \( X_i = x \) and for treatment individuals with \( X_i = x \) respectively. We can then define:

\[
\hat{\tau}_{\text{DATE}}(x) = \tau(\hat{F}_{Y|T=X=x}, \hat{F}_{Y|C=X=x}),
\]

\[
\hat{\tau}_{\text{DATE}}(S) = \frac{1}{N(S)} \sum_{Q_i: X_i \in U_X(S)} \hat{\tau}_{\text{DATE}}(X_i) = \sum_{x \in U_X(S)} \frac{N(x)}{N(S)} \hat{\tau}_{\text{DATE}}(x),
\]

\[
\hat{\mu}_{\text{DATE}}(S) = \text{Div} \left( \hat{\tau}_{\text{DATE}}(S), \mathbb{E}_{H_0}[\hat{\tau}_{\text{DATE}}(S)] \right),
\]

7
where $U_X(S)$ is the set of unique covariate profiles in $S$, while $N(x)$ and $N(S)$ are the numbers of individuals $Q_i$ with $X_i = x$ and $X_i \in S$ respectively.

To show that $\hat{F}_{Y|X=x}$ and $\hat{F}_{Y|X=x}$ are unbiased estimators of $F_{Y|X=x}$ and $F_{Y|X=x}$ requires additional assumptions about the mechanism by which units are assigned to the treatment or control group. Recall that $W_i$ determines which potential outcome is observed for unit $Q_i$. If $W_i$ is biased in which units it assigns to treatment, then subsequent inferences that do not account for this bias may be inaccurate. Thus we assume unconfoundedness: $Y_i(0), Y_i(1) \perp W_i | X_i$, i.e., potential outcomes are independent of treatment assignment conditional on the covariates. As a consequence, $E[\hat{F}_{Y|X=x}(y)] = F_{Y(0)|X=x}(y)$, and similarly $E[\hat{F}_{Y|X=x}(y)] = F_{Y(1)|X=x}(y)$ (Lemma 1 in Appendix A). Therefore, given a randomized experiment, the empirical cumulative distribution of the control group is an unbiased and strongly consistent estimator of its population cumulative distribution function, and likewise for the treatment group. Noting that $\frac{N(x)}{N(S)} \xrightarrow{a.s.} P(X = x | X \in S)$, it follows that $\hat{\tau}_{DATE}(x)$, $\hat{\tau}_{DATE}(S)$, and $\hat{\mu}_{DATE}(S)$, defined in terms of $\hat{F}_{Y|X=x}$ and $\hat{F}_{Y|X=x}$ as above, are unbiased and strongly consistent finite sample estimators of their population estimands.

### 3.5 Choice of divergence function and test statistic

Having defined the finite sample estimator $\hat{\mu}_{DATE}(S)$ in terms of the divergence $\text{Div}(\cdot, \cdot)$ between $\hat{\tau}_{DATE}(S)$ and its expectation under the null hypothesis, we now consider the choice of divergence function. For the non-parametric average treatment effect estimator, recall that $\hat{\mu}_{NATE}(S) = \max_\alpha \hat{\mu}_{NATE}(S) = \max_\alpha \text{Div}(\hat{\tau}_{NATE}(S), \alpha)$, where $\hat{\tau}_{NATE}(x) = \hat{F}_{Y|X=x}(\hat{F}_{Y|X=x}^{-1}(\alpha))$ maps the $\alpha$ quantile of the control observations with $X = x$ to a corresponding quantile $\beta_x(\alpha)$ of the treatment observations with $X = x$. Let $N_a(x)$ be defined as number of outcomes $y$, that are significant at $\alpha$ level, with covariate profile $X_i = x$.

We now consider two different models of the data generating process, based on the binomial distribution and a normal approximation to the binomial respectively. In each case, we compute the log-likelihood ratio statistic $F(S) = \log \left( \frac{P(\text{Data}|H_1(S))}{P(\text{Data}|H_0)} \right)$, which can be written as the product of the total number of p-values $N(S)$ in subset $S$ and a divergence $\text{Div}\left( \frac{N_a(S)}{N(S)}, \alpha \right)$ between the observed and expected proportions of p-values that are significant at level $\alpha$. For the binomial model, we have:

$$H_0 : N_a(x) \sim \text{Binomial}(N(x), \alpha) \quad \forall x$$

$$H_1(S) : N_a(x) \sim \text{Binomial}(N(x), \beta) \quad \forall x \in U_X(S) \quad \beta \neq \alpha,$$

with the following Berk-Jones (BJ) log-likelihood ratio statistic $F^{BJ}(S)$:

$$F^{BJ}_\alpha(S) = \log \left( \frac{P(\text{Data}|H_1(S))}{P(\text{Data}|H_0)} \right)$$

$$= N_a(S) \log \left( \frac{\beta}{\alpha} \right) + (N(S) - N_a(S)) \log \left( \frac{1 - \beta}{1 - \alpha} \right)$$

$$= N(S) \text{Div}_{KL} \left( \frac{N_a(S)}{N(S)}, \alpha \right),$$

where we have used the maximum likelihood estimate $\beta = \beta_{MLE}(S) = \frac{N_a(S)}{N(S)}$, and $\text{Div}_{KL}(\cdot, \cdot)$ is the Kullback-Liebler divergence, $\text{Div}_{KL}(x, y) = x \log \frac{x}{y} + (1 - x) \log \frac{1 - x}{1 - y}$.

For the normal approximation, we have:

$$H_0 : N_a(x) \sim \text{Gaussian}(N(x)\alpha, \alpha(1 - \alpha)N(x)) \quad \forall x$$

$$H_1(S) : N_a(x) \sim \text{Gaussian}(N(x)\beta, \alpha(1 - \alpha)N(x)) \quad \forall x \in U_X(S) \quad \beta \neq \alpha,$$
with the following normal approximation (NA) log-likelihood ratio statistic:

\[
F_N^N(S) = \log \left[ \frac{P(\text{Data}|H_1(S))}{P(\text{Data}|H_0)} \right]
= \frac{N_\alpha(S)(\beta - \alpha)}{\alpha(1 - \alpha)} + \frac{N(S)(\alpha^2 - \beta^2)}{2\alpha(1 - \alpha)}
= \frac{(N_\alpha(S) - N(S)\alpha)^2}{2N(S)\alpha(1 - \alpha)}
= N(S)\text{Div}_{\chi^2} \left( \frac{N_\alpha(S)}{N(S)}, \alpha \right),
\]

where we have again used the maximum likelihood estimate of \( \beta = \frac{N_\alpha(S)}{N(S)} \), and \( \text{Div}_{\chi^2}(\cdot, \cdot) \) is a scaled \( \chi^2 \) divergence, \( \text{Div}_{\chi^2}(x, y) = \frac{(x - y)^2}{2y(1 - y)} \). The above \( \text{Div} \left( \frac{N_\alpha(S)}{N(S)}, \alpha \right) \) each represent a \( \mu_{\text{NATE}_\alpha}(S) \), exhibiting the desirable properties described in Section 3.4, and the corresponding score functions can be written as \( F_\alpha(S) = N(S)\mu_{\text{NATE}_\alpha}(S) \). We note that in NA, the alternative hypothesis corresponds to a change in the mean of a normal distribution, while the variance remains unchanged. We show in Appendix A that this test statistic is related to many well-known goodness-of-fit statistics such as the Kolmogorov-Smirnov, Cramer-von Mises, Anderson-Darling, and Higher Criticism statistics.

### 4 Treatment Effect Subset Scanning

Treatment Effect Subset Scan (TESS) is a novel framework for identifying subpopulations in a randomized experiment which experience treatment effects—i.e., changes in quantiles of their outcome distribution—built atop the framework for distributional causal inference established in [3], with the divergence estimat of interest described in Section 3.3. Unlike previous methods, TESS structures the challenge of treatment effect identification as an anomalous pattern detection problem—where the objective is to identify patterns of systematic deviations away from expectation—which is then solved by scanning over subpopulations. TESS therefore searches for subsets of values of each attribute for which the distributions of outcomes in the treatment groups are systematically anomalous, i.e., significantly different from their expectation as derived from the control group. More precisely, we define a real-valued outcome of interest \( Y \) and a set of discrete covariates \( X = (X^1, \ldots, X^d) \), where each \( X^j \) can take on a vector of values \( V^j = \{v^j_{\ell}\}_{\ell=1}^{\ell=|V^j|} \). Therefore, we define the arity of covariate \( X^j \) as \( |V^j| \), the cardinality of \( V^j \), and note that any covariate profile \( x \), a realization of \( X \), follows \( x \in \{V^1 \times \ldots \times V^d\} \). We then define a dataset (as in [3]) as a sample \( N \) composed of \( n \) records (units) \( \{R_1, \ldots, R_n\} \), randomly drawn from population \( P \), where each 3-tuple \( R_i = (Y^i, X^i, W_i) \) is described by an observed potential outcome \( Y_i = Y^{i,\text{obs}} \), covariates \( X^i \), and an indicator variable \( W_i \), which indicates if the unit was randomly assigned to the treatment condition; see Table 1 for a demonstrative example. We define the subpopulations \( S \) under consideration to be \( S = \{v^1 \times \ldots \times v^d\} \), where \( v^j \subseteq V^j \). We wish to find the most anomalous subset

\[
S^* = v^1^* \times \ldots \times v^d^* = \arg \max_S F(S)
\]

where \( F(S) \) is commonly referred to in the anomalous pattern detection literature as a score function, to measure the anomalousness of a subset \( S \). In the context of TESS, this function is a test statistic of the treatment effect—i.e., the divergence between the treatment and control group—in subpopulation \( S \) and therefore will be a function of the estimator \( \mu_{\text{NATE}}(S) \).

We accomplish this by first partitioning the experiment dataset into control and treatment groups, and passing the groups to the TESS algorithm. For each unique covariate profile \( x \) in the treatment group, TESS computes the empirical conditional outcome distribution \( \hat{F}_{Y|X=x} \) from the control group, estimating the conditional outcome distribution under the null hypothesis \( H_0 \) that the treatment has no effect on units.
with this profile. Then for each record \( R_i \) in the treatment group, TESS computes an empirical \( p \)-value \( p_i \), which serves as a measure of how uncommon it is to see an outcome as extreme as \( Y_i \) given \( X = X_i \) under \( H_0 \). The ultimate goal of TESS is to discover subpopulations \( S \) with a large amount of evidence against \( H_0 \), i.e., the outcomes of units in \( S \) are consistently extreme given \( H_0 \). Thus, TESS searches for subpopulations which contain an unexpectedly large number of low (significant) empirical \( p \)-values, as such a subpopulation is more likely to have been affected by the treatment.

### 4.1 Estimating Reference Distributions and Empirical \( p \)-values

After partitioning the data into treatment and control groups, the TESS framework estimates the reference distribution for each unique covariate profile in the treatment group. These estimates follow from two assumptions: randomization and a sharp null hypothesis of no treatment effect. Randomization implies that \( \hat{F}_{Y|X} \) is an unbiased and strongly consistent estimator of the unknown \( F_{Y|X} \) under \( H_0 \). Intuitively, under \( H_0 \) the outcomes of the treatment and control groups are drawn from the same distribution, allowing \( \hat{F}_{Y|X} \) to serve as an outcome reference distribution for treatment units with covariate profile \( X = x \). **When \( H_0 \) is true** the outcomes for every unit in the treatment group and the control group, with the same covariate profile, are exchangeable. **When \( H_0 \) is false** the affected treatment outcomes are drawn from an alternative distribution, different than their assumed reference distributions under \( H_0 \). No additional assumptions are made about the relationship between reference distributions.

TESS calculates an empirical \( p \)-value for each treatment unit to obtain a measure of how “anomalous” or unusual a particular unit’s outcome is given its reference distribution. For each unit \( R_i \) in the treatment group, we compute its empirical reference distribution:

\[
\hat{F}_{Y|X}(y_i|x_i) = \frac{1}{N(x_i)} \sum_{y_j \in Y^C(x_i)} 1\{y_j \leq y_i\},
\]

where \( Y^C(x_i) \) is the set of outcomes for control units with covariate profile \( x_i \) and \( N(x_i) = |Y^C(x_i)| \). The empirical \( p \)-value \( p(y_i) \) (or \( p_i \) for notational convenience) is derived from \( \hat{F}_{Y|X} \), as in [25]. Furthermore, the \( p_i \) are guaranteed to be distributed Uniform(0,1) under \( H_0 \), which follows from exchangeability and the probability integral transform. We define the significance of a \( p \)-value, for a significance level \( \alpha \), as \( n_{\alpha}(p(y)) = 1\{p(y) \leq \alpha\} \).

Although we define and estimate \( \hat{F}_{Y|X} \) individually for each unique covariate profile \( x \) using the empirical distribution function, we note that TESS only requires some means of computing a \( p \)-value for each treatment unit. The empirical distribution allows TESS to accommodate arbitrary differences in conditional outcome

\footnote{Traditional empirical \( p \)-values are only asymptotically Uniform(0,1); for \( p \)-value ranges \[25\], \( p \)-values drawn uniformly from each range will be Uniform(0,1), even in finite samples.}

### Table 1

| Record | \( Y \) | \( X^{\text{gender}} \) | \( X^{\text{race}} \) | \( W \) |
|--------|--------|----------------|----------------|-------|
| 1      | 2.35   | Female         | Black          | 1     |
| 2      | 2.06   | Female         | White          | 1     |
| 3      | 2.92   | Male           | Black          | 1     |
| 4      | 2.27   | Male           | White          | 1     |
| 5      | 1.73   | Female         | Black          | 0     |
| 6      | 1.84   | Female         | White          | 0     |
| 7      | 1.7    | Male           | Black          | 0     |
| 8      | 1.59   | Male           | White          | 0     |

Table 1: This table is a demonstrative dataset of \( n = 8 \) records, with a \( d = 2 \) sized vector of covariates, \( X = (X^{\text{gender}}, X^{\text{race}}) \). The first, \( (X^{\text{gender}}) \), can take values in \( V^{\text{gender}} = \{\text{Female}, \text{Male}\} \), and the second \( (X^{\text{race}}) \) can take values in \( V^{\text{race}} = \{\text{Black, White}\} \). A covariate profile \( x \), and realization of \( X \), is an element in the set of all covariate profiles \( V^{\text{race}} \times V^{\text{gender}} = \{\{\text{Female, Black}\}, \{\text{Female, White}\}, \{\text{Male, Black}\}, \{\text{Male, White}\}\} \).
distributions across covariate profiles, enabling general applicability without a priori contextual knowledge. However, in a specific context of interest, it may be possible to combine data across profiles to construct a more general estimate of the conditional probability distributions. Statistical learning offers many options for density estimation, any of which can be utilized in TESS.

### 4.2 Subpopulations

Given \( p \)-values as a measure of the anomalousness of individual treatment units, we now consider how TESS combines these measures to form subpopulations. For intuition, we propose representing the data as a tensor, where each covariate is represented by a mode of the tensor, \( |V^j| \), the arity of the \( j^{th} \) covariate, is the size of the \( j^{th} \) mode. Therefore, each covariate profile \( x \) maps to a unique cell in the tensor, which contains the \( p \)-values of the treatment units that share \( x \) as their covariate profile. As stated above, a subpopulation is \( S = \{v^1 \times \ldots \times v^d\} \), where \( v^j \subseteq V^j \); therefore, an individual cell (i.e., covariate profile \( x \)) is itself a subpopulation: \( S = \{\{x^1\} \times \ldots \times \{x^d\}\} \). For a demonstrative example see Table 2. For a given subpopulation \( S \), we define the quantities

\[
C(S) = \bigcup_{x \in U_X(S)} Y^T(x), \quad N_\alpha(S) = \sum_{y \in C(S)} n_\alpha(p(y)), \quad N(S) = \sum_{y \in C(S)} 1
\]

(7)

where \( Y^T(x) \) is defined similarly to \( Y^C(x) \), but for the treatment group; \( C(S) \) is the union of treatment outcomes in the cells (i.e., covariate profiles) in \( S \), and the corresponding \( p \)-values; \( N(S) \) represents the total number of empirical \( p \)-values contained in \( C(S) \); and \( N_\alpha(S) \) is the number of \( p \)-values in \( C(S) \) that are less than \( \alpha \). Given that the distribution of each \( p \)-value is Uniform(0,1) under the null hypothesis that the treatment has no effect, for a subpopulation \( S \) consisting of \( N(S) \) empirical \( p \)-values, \( E[N_\alpha(S)] = \alpha N(S) \). Under the alternative hypothesis, we expect the outcomes of the affected units to be more concentrated in the tails of their reference distributions; thus, the \( p \)-values for these affected units will be lower. Therefore, subpopulations composed of covariate profiles that are systematically affected by the treatment should express higher values of \( N_\alpha(S) \) for some \( \alpha \). Consequently, a subpopulation \( S \) where \( N_\alpha(S) > \alpha N(S) \) (i.e., with a higher than expected number of low, significant \( p \)-values) is potentially affected by the treatment.

### 4.3 Nonparametric Scan Statistic

TESS utilizes the nonparametric scan statistic \cite{11, 25} to evaluate the statistical anomalousness of a subpopulation \( S \) by comparing the observed and expected number of significantly low \( p \)-values it contains. The general form of the nonparametric scan statistic is

\[
F(S) = \max_\alpha F_\alpha(S) = \max_\alpha \phi(\alpha, N_\alpha(S), N(S)) = \max_\alpha N(S) \mu_{NATE_\alpha}(S),
\]

where \( N_\alpha(S) \) and \( N(S) \) are defined as in (7). Here \( F_\alpha(S) \) is a log-likelihood ratio test statistic of the treatment effect in subpopulation \( S \) which, as shown in \cite{3.5}, is proportional to the divergence \( \mu_{NATE_\alpha}(S) \).

\(^2\)For \( p \)-value ranges, as in \cite{25}, \( N_\alpha(S) \) is more precisely the total probability mass less than \( \alpha \) over the \( p \)-value ranges in \( C(S) \).
We consider “significance levels” $\alpha \in [\alpha_{\text{min}}, \alpha_{\text{max}}]$, for constants $0 < \alpha_{\text{min}} < \alpha_{\text{max}} < 1$. Maximizing $F(S)$ over a range of $\alpha$, rather than a single arbitrarily-chosen $\alpha$ value, enables TESS to detect a small number of highly anomalous p-values, a larger subpopulation with subtly anomalous p-values, or anything in between. The range of $\alpha$ to consider can be specified based on the quantile values of interest, or $(0, 1)$ representing the entire distribution. The choice of $\alpha_{\text{max}}$ describes how extreme a value must be, as compared to the reference distribution, in order to be considered significant. We often choose $\alpha_{\text{min}} \approx 0$, but larger values can be used to avoid returning subsets with a small number of extremely significant p-values.

### 4.3.1 Efficient Scanning

The next step in the TESS framework is to detect the subpopulation most affected by the treatment, i.e., to identify the most anomalous subset of values for each of the $d$ modes of the tensor, or equivalently for each covariate $X^1 \ldots X^d$. More specifically, the goal is to identify the set of subsets $\{v^1, \ldots, v^d\}$ where each element corresponds to values in a tensor-mode (covariate), such that $F(v^1 \times \ldots \times v^d)$ is jointly maximized. The computational complexity of solving this optimization naively is $O(2^{\sum_j |V^j|})$, where $|V^j|$ is the size of mode $j$ (the arity of $X^j$), and is computationally infeasible for even moderately sized datasets. We therefore employ the linear-time subset scanning property (LTSS) \cite{26}, which allows for efficient and exact maximization of any function satisfying LTSS over all subsets of the data.

We begin by noting that, for the score function $F_\alpha(S)$ with a fixed value of $\alpha$: (A1) $\phi$ is monotonically increasing w.r.t. $N_\alpha$, (A2) $\phi$ is monotonically decreasing w.r.t. $N$, and (A3) $\phi$ is convex w.r.t. $N_\alpha$ and $N$. These properties are intuitive because the ratio of observed to expected number of significant p-values $\frac{\alpha}{\gamma}$ increases with the numerator (A1) and decreases with the denominator (A2). Also, a fixed ratio of observed to expected is more significant when the observed and expected counts are large (A3). We now turn to the LTSS property which states that, for a given set of data elements $R = \{R_1 \ldots R_n\}$, a score function $F(S)$ mapping $S \subseteq R$ to a real number, and a priority function $G(R_i)$ mapping a single data element $R_i \in R$ to a real number, the LTSS property guarantees that the only subsets with the potential to be optimal are those consisting of the top-$t$ highest priority records $\{R_{(1)} \ldots R_{(t)}\}_{t \in [1,n]}$. More formally, we restate a theorem from (author?) \cite{26} and add a corollary that extends LTSS to the high-dimensional tensor context:

**Theorem 1.** Let $F(S) = F(X,Y)$ be a function of two additive sufficient statistics of subset $S$, $X(S) = \sum_{R_i \in S} x_i$ and $Y(S) = \sum_{R_i \in S} y_i$, where $x_i$ and $y_i$ depend only on element $R_i$. Assume that $F(S)$ is monotonically increasing with $X(S)$, that all $y_i$ values are positive, and that $F(X,Y)$ is convex. Then $F(S)$ satisfies the LTSS property with priority function $G(R_i) = \frac{y_i}{x_i}$.

**Corollary 1.** Consider the general class of nonparametric scan statistics $F(S) = \max_\alpha F_\alpha(S)$, where the significance level $\alpha \in [\alpha_{\text{min}}, \alpha_{\text{max}}]$, for constants $0 < \alpha_{\text{min}} < \alpha_{\text{max}} < 1$. For a given value of $\alpha$ and $v^{-j} = \{v^1, \ldots, v^{j-1}, v^{j+1}, \ldots, v^d\}$ under consideration, $F_\alpha(S)$ can be efficiently maximized over all subpopulations $S = v^j \times v^{-j}$, for $v^j \subseteq V^j$.

**Proof.** First, we note that number of p-values in every $v^j$ is positive: we only consider the values of a covariate that are expressed by some treatment unit. Thus we have $F_\alpha(S) = \phi(\alpha, N_\alpha(v^j), N(v^j))$, with the additive sufficient statistics $N_\alpha(v^j) = \sum_{y \in C(v^j \times v^{-j})} n_\alpha(p(y))$ and $N(v^j) = \sum_{y \in C(v^j \times v^{-j})} 1$. Since the nonparametric scan statistic is defined to be monotonically increasing with $N_\alpha$ (A1), monotonically decreasing with $N$ (A2), and convex (A3), we know that $F_\alpha(S)$ satisfies the LTSS property with priority function, over the values of mode (covariate) $j$, $G_\alpha(v^j_m) = \frac{\sum_{y \in C(v^j \times v^{-j})} n_\alpha(p(y))}{\sum_{y \in C(v^j \times v^{-j})} 1}$ for $v^j_m \in V^j$. Therefore the LTSS property holds for each value of $\alpha$, enabling each $F_\alpha(S)$ to be efficiently maximized over subsets of values for the $j^{th}$ mode of the tensor, given values for the other $d - 1$ modes.

Essentially, Corollary 1 demonstrates that the nonparametric scan statistic satisfies LTSS in the context of TESS and therefore a single mode of a tensor can be efficiently optimized, conditioned on values of the other modes. Let $U_\alpha(S)$ be the set of unique p-values between $\alpha_{\text{min}}$ and $\alpha_{\text{max}}$ contained in subpopulation $S$. Then the quantity $\max_S F(S) = \max_{\alpha \in U_\alpha(S)} \max_S F_\alpha(S)$ can be efficiently and exactly computed over...
all subsets $S = v^j \times v^{-j}$, where $v^j \subseteq V^j$, for a given subset of values for each of the other modes $v^{-j}$. To do so, consider the set of distinct $\alpha$ values, $U = U_\alpha(V^j \times v^{-j})$. For each $\alpha \in U$ we employ the same logic as described in Corollary 1 to optimize $F_\alpha(S)$: we compute the priority $G_\alpha(v^j_m)$ for each value $(v^j_m \in V^j)$, sort the values based on priority function $G_\alpha(v^j_m)$, and evaluate subsets $S = \{v^j_1, \ldots, v^j_t\} \times v^{-j}$ consisting of the top-$k$ highest priority values, for $t = 1, \ldots, |V^j|$.

TESS then iterates over modes of the tensor, using the efficient optimization steps described above to optimize each mode: $v^j = \arg\max_{v^j \subseteq V^j} F(v^j \times v^{-j}), j = 1 \ldots d$. The cycle of optimizing each mode continues until convergence, at which point TESS has reached a conditional maximum of the score function, i.e., $v^j$ is conditionally optimal given $v^{-j}$ for all $j = 1 \ldots d$. This ordinal ascent approach is not guaranteed to converge to the joint optimum, but with multiple random restarts the combination of subset scanning and ordinal ascent has been shown to locate near globally optimal subsets with high probability [25, 27]. Moreover, if $\sum_{j=1}^d |V^j|$ is large, this iterative procedure makes the ability to detect anomalous subpopulations computationally feasible, without excluding potentially optimal subpopulations from the search space (as a greedy top-down approach may). A single iteration (optimization of mode $j$ of the tensor) has a complexity $O\left(|U| \left(n_t + |V^j| \log |V^j|\right)\right)$, where the $n_t$ term, the number of treatment units, results from collecting the $p$-values for all units in $C\left(V^j \times v^{-j}\right)$ over our sparse tensor; $U = U_\alpha(V^j \times v^{-j})$, with $|U| \leq n_t$; and $O\left(|V^j| \log |V^j|\right)$ is required to sort the values of tensor mode $j$. Therefore a step in the procedure (a sequence of $d$ iterations over all modes of the tensor) has complexity $O\left(\hat{O}d \left(n_t + \hat{V} \log \hat{V}\right)\right)$, where $\hat{U}$ and $\hat{V}$ are the average numbers of $\alpha$ thresholds considered and covariate arity, respectively. Thus the TESS search procedure has a total complexity of $O\left(I \hat{O}d \left(n_t + \hat{V} \log \hat{V}\right)\right)$, where $I$ is the number of random restarts and $\hat{Z}$ is the average number of iterations required for convergence. We note that $\hat{Z}$ is typically very small; $\hat{Z} \leq 5$ across all simulations discussed in §5.

### 4.4 TESS Algorithm

**Inputs:** randomized experiment dataset, $\alpha_{\min}, \alpha_{\max}$, number of iterations $I$.

1. For each unique covariate profile $x$ in the treatment group:
   
   (a) Estimate $\hat{F}_{Y|X=x}$ from the outcomes of the units in the control group that share profile $x$.
   
   (b) Compute the $p$-value $p_i = p(y_i)$ for each treatment unit $i$ with profile $x$ from $\hat{F}_{Y|X=x}$.

2. Iterate the following steps $I$ times. Record the maximum value $F^*$ of $F(S)$, and the corresponding subsets of values for each mode $\{v^{1*}, \ldots, v^{d*}\}$ over all such iterations:

   (a) For each of the $d$ modes, initialize $v^j$ to a random subset of values $V^j$.
   
   (b) Repeat until convergence:
      
      i. For each of the $d$ modes:
         
         A. Maximize $F(S) = \max_{\alpha \in [\alpha_{\min}, \alpha_{\max}]} F_\alpha(v^j \times v^{-j})$ over subsets of values for $j^{th}$ mode $v^j \subseteq V^j$, for the current subset of values of the other $d-1$ modes ($v^{-j}$), and set $v^j \leftarrow \arg\max_{v^j \subseteq V^j} F(v^j \times v^{-j})$.

3. Output $S^* = \{v^{1*}, \ldots, v^{d*}\}$.

### 4.5 Estimator Properties

In the above sections we outline a procedure to efficiently compute $\max_{S \subseteq R} F(S)$, where $R$ represents the space of all rectangular subsets. In this section we treat $\max_{S \subseteq R} F(S)$ as a statistic of the data and aim to show that it has desirable statistical properties. It is known that for data $X_1, \ldots, X_n \overset{iid}{\sim} P$ and the corresponding empirical distribution function $P_n$, $\|P_n - P\|_\infty \overset{a.s.}{\rightarrow} 0$. Many goodness-of-fit statistics $\text{GoF}(P_n, P)$ are equivalent to an empirical process over centered and scaled empirical measures; and empirical process
theory provides tools to control Type I and II error\[15, 18, 31\]. However, in a general sense our goal is to control the behavior of $\max_{S \subseteq \{1, \ldots, N\}} GoF(P_S, P)$, where $P_S$ is the empirical measure given by the subset $S$. It is not obvious whether the desirable properties present for $P_\alpha$ will persist when considering the empirical measure of $P_S$, a non-random subset of the data chosen by our optimization procedure. Given that this context of optimization over subsets is not considered in the current goodness-of-fit literature, we provide various theoretical results in support of our subset scanning algorithm. In the remainder of the section we present the key statements necessary to show our desired properties below, while additional results and all proofs can be found in Appendix A. We begin with:

**Lemma 3.** $F_{BJ}(S) \approx F_{NA}(S)$ as $N(S) \rightarrow \infty$.

This result indicates that, as the number of subjects in a given subpopulation grows, its score under $F_{BJ}$ is well approximated by $F_{NA}$, where both functions are described in [4, 5]. Given the fact that a large class of other goodness-of-fit statistics in the literature are monotonic transformations of $F_{NA}$ (see Appendix A), this result allows us to focus the remainder of our results on $F_{NA}$. Our score function can be considered a test statistic for the following hypothesis test:

$$
H_0 : Y_i(1) \sim F_{Y_i(0)|X}, \quad \forall X_i \in U_X(D)
$$

$$
H_1 (S) : Y_i(1) \neq F_{Y_i(0)|X}, \quad \forall X_i \in U_X(S) \quad \text{and} \
Y_i(1) \sim F_{Y_i(0)|X}, \quad \forall X_i \notin U_X(S), \quad S \in R
$$

(8)

where $D$ is our dataset (or tensor) of treatment units and $R$ is the set of all rectangular subsets of $D$. The null hypothesis is that all of the observed outcomes of treatment units are drawn from the same conditional outcome distribution (given the observed covariates) as their control group. The hypothesis tests which serve as the foundation for the score functions described in [4, 5] are special cases of this more general hypothesis test in (8). Recall that $U_X(D)$ is the set of unique covariate profiles (non-empty tensor cells) in our data, with cardinality $|U_X(D)| = M$; while $S^* = \arg \max_{S \in R} F(S)$ and $S_\alpha^* = \arg \max_{\alpha} F(S)$ represent the most anomalous rectangularly constrained subset and the most anomalous unconstrained subset respectively. We assume $N(x) \geq n$ for all $x \in U_X(D)$, i.e., at least $n$ units belong to each unique covariate profile (non-empty cell) in the data. We consider the case where $M, n \rightarrow \infty$, maximizing $F(S) = \max_\alpha F_\alpha(S)$ over $\alpha \in [\alpha_{\min}, \alpha_{\max}]$ for constants $0 < \alpha_{\min} < \alpha_{\max} < 1$. We can therefore demonstrate:

**Lemma 4.** Under $H_0$ defined in (8), $F_{NA}(S_u^*) \overset{D}{\rightarrow} \max_{Z \in [\phi(Z)^2]} M_{\phi(Z)} \approx 0.202 M$, where $\phi$ and $\Phi$ are the Gaussian pdf and cdf respectively.

Thus, when the null hypothesis is true, the score of the most unconstrained anomalous subset is asymptotically linear in $M$. Our ability to understand the limiting behavior of the $F(S_u^*)$ is built atop LTSS theory which indicates the optimal unconstrained subset will be $S_u^* = \{x(1), \ldots, x(t)\} \in [1, M]$, where $x(t)$ has the $t^{th}$ largest value of the random variable $N(x) \forall x \in U_X(D)$. Next we note that the score maximized over the space of unconstrained subsets upper-bounds the score maximized over the subspace of rectangular subsets: $F(S^*) \leq F(S_u^*)$. Therefore, we have the following result:

**Theorem 2.** Under $H_0$ defined in (8), let $N(x) \geq n \forall x \in U_X(D)$, fix $\epsilon > 0$, and assume $M, n \rightarrow \infty$; then there exist a constant $C \leq \max_Z \frac{M_{\phi(Z)}}{2(1 - \Phi(Z))} \approx 0.202$ and critical value $h(M, \epsilon) = CM + \epsilon$ such that

$$
P_{H_0} \left( \max_{S \in R} F(S) > h(M, \epsilon) \right) \rightarrow 0.
$$

As a direct consequence of Theorem 2, $\max_{S \in R} F(S)$ provides a statistic to quantify the evidence to reject $H_0$, whose Type I error rate can be controlled, producing an asymptotically valid $\gamma$-level hypothesis test $P_{H_0}(\text{Reject } H_0) \leq \gamma$, for any fixed $\gamma \in (0, 1)$. We note that, because we are maximizing both over subsets $S$ and thresholds $\alpha$, our result is distinct from the straightforward application of Dvoretzky-Kiefer-Wolfowitz bounds (maximizing over $\alpha$ for a given subset $S$), which would give us $\max_{\alpha} \frac{N_\alpha(S)}{N(S)} - \alpha \rightarrow 0$ and therefore $F(S) \rightarrow 0$.\[14\]
Next, we turn our attention to the alternative hypothesis, where $S^T$ represents the truly affected subset, $k = \left| \frac{|U_X(S^T)|}{|U_X(D)|} \right|$ is its proportion of covariate profiles, and $H_1 \left( S^T \right)$ implies that there exist constants $\alpha$ and $\beta(\alpha) > \alpha$ such that, for all $x \in U_X(S^T)$, $\beta_x(\alpha) = F_{Y(1)|X=x}\left( F_{Y(0)|X=x}^{-1}(\alpha) \right) = \beta(\alpha)$. We then have the following results:

**Lemma 5.** Under $H_1 \left( S^T \right)$, $F^{NA} \left( S^T \right) \xrightarrow{a.s.} \max_{x}(\beta(\alpha) - \alpha)^2 \cdot \frac{kMn}{2n(1-\alpha)}$.

**Theorem 3.** Under $H_1 \left( S^T \right)$ defined in [8], let $N(x) \geq n \forall x \in U_X(D)$, fix $\epsilon > 0$, and assume $M, n \rightarrow \infty$; then for the same critical value $h(M, \epsilon)$ as in Theorem 2,

$$P_{H_1} \left( \max_{S \in R} F(S) > h(M, \epsilon) \right) \rightarrow 1.$$  

As a direct consequence of Theorem 3, the Type II error rate can be controlled and produces a hypothesis test with full asymptotic power $P_{H_1} (\text{Reject } H_0) \rightarrow 1$. Note that in this context we consider a fixed alternative $\beta(\alpha)$, as opposed to a local alternative where $\beta_M(\alpha) \rightarrow \alpha$ as $M, n \rightarrow \infty$. Additionally, the critical value $h(M, \epsilon)$ is the same in Theorems 2 and 3, we are therefore showing that $P(\text{Type I error}) + P(\text{Type II error}) \rightarrow 0$. We note that for any given experiment (with finite $M$ and $n$), permutation testing can be used to control the Type I error rate of our scanning procedure, and conditions have been shown where permutation calibrations achieve the Type II error rates of an oracle scan test [3].

These results intuitively capture our statistic’s ability to conclude that the null hypothesis is false—i.e., there exists some subset that follows $H_1$, and therefore invalidates $H_0$. However, this does not necessarily provide a guarantee that the statistic will exactly capture the true subset. Therefore, next we will derive finite sample conditions under which our framework achieves subset correctness: $S^* = S^T$.

We begin by demonstrating that the score function of interest can be re-written as an additive function if we condition on the value of the null and alternative hypothesis parameters $\alpha$ and $\beta(\alpha)$ from the hypothesis test in [3, 5]. More specifically, the score of a subset $S$ can be decomposed into the sum of contributions (measured by a function $\omega$) from each individual covariate profile $x$ contained within the subset. For example, with respect to $F^{NA}$, $\omega^{NA}(\alpha, \beta, N_\alpha(x), N(x)) = C_{\alpha,\beta}^1 N_\alpha(x) + C_{\alpha,\beta}^2 N(x)$, where each $C$ is only a function of $\alpha$ and $\beta$, and therefore constant.

**Lemma 6.** $F(S)$ can be written as $\max_{\alpha, \beta} \sum_{x \in U_X(S)} \omega(\alpha, \beta, N_\alpha(x), N(x))$, for $\alpha, \beta \in (0, 1)$ representing quantile values of the control and treatment potential outcomes distributions respectively.

Next, we seek to demonstrate some important properties of the $\omega$ functions. More specifically we have that

**Lemma 7.** $\omega^{NA}(\alpha, \beta, N_\alpha(x), N(x))$ is concave with respect to $\beta$, maximized at $\beta_{mle}(x) = \frac{N_\alpha(x)}{N(x)}$, and has two roots $(\beta_{\min}(x), \beta_{\max}(x))$.

We show the same result in Lemma 8 for $\omega^{BJ}$. Intuitively, $(\beta_{\min}(x), \beta_{\max}(x))$ is the interval over which $\omega$ is concave and makes a positive contribution to the score of a subset, while this contribution is maximized at $\beta_{mle}(x)$; we note that in the case of $\omega^{NA}$, $\beta_{\min}(x) = \alpha$. We are now interested in the relationship between $r_{\max} = \beta_{\max}(x) - \alpha$ and $r_{\min} = \beta_{mle}(x) - \alpha$.

**Lemma 9.** With respect to $\omega^{NA}(\alpha, \beta, N_\alpha(x), N(x))$, $\frac{r_{\min}(x)}{r_{\max}(x)} = 2$.

We show a similar result in Lemma 10 for $\omega^{BJ}$. Given these two properties of the $\omega$ function, we can now provide the sufficient conditions for the detected subset to be exactly correct, i.e., $S^* = S^T$. We introduce some additional notation: $r_{mle-h}^\text{aff} = \max_{x \in U_X(S^T)} r_{mle}(x)$, $r_{mle-l}^\text{aff} = \min_{x \in U_X(S^T)} r_{mle}(x)$, $r_{mle-h}^\text{unaff} = \max_{x \in U_X(S^T)} r_{mle}(x)$, $\eta = \frac{\sum_{x \in U_X(S^T)} N(x)}{\sum_{x \in U_X(S)} N(x)}$, and invertible function $R: r_{\max}(x) \mapsto r_{\min}(x)$. From Lemma 9 we know that with respect to $\omega^{NA}$, $B^{NA}(r) = \frac{r}{2}$. We also introduce the concepts of $\nu$—homogeneous, which means that $\frac{r_{mle-l}^\text{aff}}{r_{mle-h}^\text{aff}} < \nu$, and $\delta$—strong, which means that $\frac{r_{mle-l}^\text{aff}}{r_{mle-h}^\text{aff}} > \delta$. Intuitively, the concept of homogeneity
measures how similarly the treatment affects each $F_{Y|x=x}$ for $x \in U_X(S^T)$, while strength measures how large of an effect the treatment exhibits across all $F_{Y|x=x}$ for $x \in U_X(S^T)$. More specifically, these concepts respectively imply that for any pair of the affected covariate profiles $(x_i, x_j) \in U_X(S^T)$, the anomalous signal (i.e., treatment effect) observed in $x_i$ is less than $\nu$-times that which is observed in $x_j$, and the treatment effect observed in every affected covariate profile is more than $\delta$-times that of the unaffected profiles. Using these concepts we have the following results:

**Theorem 4.** Under $H_1(S^T)$ defined in \( \mathcal{S} \), where $|U_X(S^T)| = t$, $\exists \nu > 1$ such that if the observed effect (as measured by $\omega$) across the $t$ covariate profiles in $S^T$ is $\nu$-homogeneous, and at least $1$-strong, then the highest scoring subset $S^* \supseteq S^T$.

**Theorem 5.** Under $H_1(S^T)$ defined in \( \mathcal{S} \), where $|U_X(S^T)| = t$, $\exists \delta > 1$ such that if the observed effect (as measured by $\omega$) across the $t$ covariate profiles in $S^T$ is $\frac{\delta}{t}$-strong, then the highest scoring subset $S^* \subseteq S^T$.

Together, these results demonstrate that the test statistic $\max_S F(S)$ possesses desirable statistical properties. Theorems 2 and 3 imply that the asymptotic Type I and II errors of our procedure can be controlled, with implications for maximization over subsets of empirical processes more generally. Theorems 4 and 5 indicate that for a score function there exist constants $\nu$ and $\delta$, both of which equal 2 for $F^{NA}$, that define how similar and strong the treatment effect must be in the affected subpopulation, to ensure that the highest-scoring subset corresponds exactly to the true affected subset $S^* = S^T$. To our knowledge, this is the first work on heterogeneous treatment effects that provides conditions on the exactness of subpopulation discovery.

5 Empirical Analysis

In this section we empirically demonstrate the utility of the TESS framework as a tool to identify subpopulations with significant treatment effects. We use data from the Tennessee Student/Teacher Achievement Ratio (STAR) randomized experiment \[41\] in order to provide representative performance in real-world policy analysis. We review the original STAR data \[5,1\], and describe our procedure for simulating affected subpopulations \[5,2\].

Through the simulation results described in \[5,3\] we compare the ability of TESS to detect significant subpopulations to three recently proposed statistical learning approaches: Causal Tree \[8\], Interaction Tree \[8\, 34\], and Causal Forest \[39\]. Specifically, we evaluate each method on two general metrics: detection power and subpopulation accuracy. Detection power measures $P_{H_0}(\text{Reject } H_0)$, or how well a method can detect the existence—necessarily the location—of treatment effect heterogeneity in the experiment. Subpopulation accuracy, on the other hand, is specifically designed to measure how well a method can precisely and completely capture the subpopulation(s) with significant treatment effects.

Finally, we conduct an exploratory analysis of the STAR dataset, and in \[5,4\] discuss the subpopulations identified by TESS as affected by treatments. In some cases, the identified subpopulation is consistent with the literature on the STAR experiment; in other cases, TESS uncovers previously unreported, but intuitive and believable, subpopulations. These empirical results demonstrate TESS’s potential to generate potentially useful and non-obvious hypotheses for further exploration and testing.

5.1 Tennessee STAR Experiment

The Tennessee Student/Teacher Achievement Ratio (STAR) experiment is a large-scale, four-year, longitudinal randomized experiment started in 1985 by the Tennessee legislature to measure the effect of class size on student educational outcomes, as measured by standardized test scores. The experiment started monitoring students in kindergarten (during the 1985-1986 school year) and followed students until third grade. Students and teachers were randomly assigned into conditions during the first school year, with the intention for students to continue in their class-size condition for the entirety of the experiment. The three potential experiment conditions were not based solely on class size, but also the presence of a full-time teaching aide:
small classrooms (13-17 pupils), regular-size classrooms (22-25 pupils), and regular-size classrooms with aide (still 22-25 pupils). Therefore, the difference between the former two conditions is classroom size, and the difference between the latter two conditions is the inclusion of a full-time teacher’s aide in the classroom. The experiment included approximately 350 classrooms from 80 schools, each of which had at least one classroom of each type. Each year more than 6,000 students participated in this experiment, with the final sample including approximately 11,600 unique students.

The Tennessee STAR dataset has been well studied and analyzed, both by the project’s internal research team [10,11] and by external researchers [24,25]. As indicated by [24], the investigations have primarily focused on comparing means and computing average treatment effects. [24] presents a detailed econometric analysis and draws similar conclusions to the previous research: students in small classrooms perform better than those in regular classrooms, while there is no significant effect of a full-time teacher’s aide, or moderation from teacher characteristics. Moreover, the effect accumulates each year a student spends in a small classroom [21]. Additionally, these conclusions are robust in the presence of potentially compromising experimental design challenges: imbalanced attrition, subsequent changes in original treatment assignment, and fluctuating class-sizes [24].

5.2 Experimental Simulation Setup

The goal of our experimental simulation is to replicate conditions under which a researcher would want to use an algorithm to discover subpopulations with significant treatment effects, and to observe how capable various algorithms are at identifying the correct subpopulation(s). In order to replicate realistic conditions, we use the STAR experiment as our base dataset, and inject into it subpopulations (of a given size) with a treatment effect (of a given magnitude). More specifically, we treat each student-year as a unique record and for each record capture ten covariates: student gender, student ethnicity, grade, STAR treatment condition, free-lunch indicator, school development environment, teacher degree, teacher ladder, teacher experience, and teacher ethnicity. We note that each of these variables, other than teacher experience, is discrete; we discretize experience into five-year intervals: [0, 5), [5, 10), ..., [30, 30). The number of values a covariate can take ranges from two to eight. By preserving the overall data structure of the STAR experiment—number of covariates, covariate value correlations, subpopulations, sample sizes, etc.—our simulations are more able to replicate the structure (and challenges) faced by experimenters.

The process we follow to generate a simulated treatment effect begins with selecting a subpopulation \( S_{\text{affected}} \) to affect. Recall that the dataset contains a set of discrete covariates \( X = (X^1, ..., X^d) \), where each \( X^j \) can take on a vector of values \( V^j = \{v^j_m\}_{m=1,...,|V^j|} \) and \( |V^j| \) is the arity of covariate \( X^j \). Therefore, we define a subpopulation as \( S = \{v^1 \times ... \times v^d\} \), where \( v^j \subseteq V^j \). The affected subpopulation is generated at random based on two parameters: \( \text{num\_cous} \), or the number of covariates to select, and \( \text{value\_prob} \), or probability a covariate value is selected. We select \( \text{num\_cous} \) covariates at random, and for each of these covariates we select each of their values with probability equal to \( \text{value\_prob} \), ensuring that at least one value for each of these covariates is selected. The final affected subpopulation is then \( S_{\text{affected}} = \{v^1 \times ... \times v^d\} \), where \( v^j \) is the selected values if \( X^j \) is one of the \( \text{num\_cous} \) covariates, and otherwise \( v_j = V^j \). In other words, for a random subset of covariates, \( S_{\text{affected}} \) only includes a random subset of their values, and for all other covariates \( S_{\text{affected}} \) includes all of their values. This treatment effect simulation scheme allows for variation in the size of the subpopulation that is affected: instances of \( S_{\text{affected}} \) can constitute a small subpopulation (a challenging detection task), a large subpopulation (a relatively easier detection task), or something in between. Therefore a set of simulations, with varying parameter values, captures the spectrum of conditions a researcher may face when analyzing an experiment to identify subpopulations with significant treatment effects.

The next step in the process involves partitioning the dataset into treatment and control groups, and generating outcomes for each record. Outcomes are drawn randomly from one of two distributions: the null distribution \( f_0 \) or the alternative distribution \( f_1 \). Any record in the treatment group that has a covariate profile \( x \in U_X(S_{\text{affected}}) \) has outcomes generated by \( f_1 \); all other records have outcomes drawn from \( f_0 \). Therefore only \( S_{\text{affected}} \) has a treatment effect, whose effect magnitude is the distributional difference
between $f_0$ and $f_1$, represented by the parameter $\delta$.

Each of the methods we consider in these experiments has a unique approach to identifying potential subpopulations with differential treatment effects. Furthermore, as mentioned in [21], most methods in the literature do not provide a process for identifying extreme treatment effects. Therefore, we devise intuitive post-processing steps in an attempt to represent how researchers would use each method to identify potential subpopulations that have significant treatment effects. Each method returns identified subpopulations and corresponding scores (measures) of the treatment effect. For the single tree-based methods [8, 34] we follow the suggestion of [8] to perform inference (via a two-sample Welch T-Test) in each leaf of the tree, and we then sort the leaves based on their statistical significance. The final subpopulation returned by the tree is the leaf with the most statistically significant treatment effect, and the final treatment effect measure is this leaf’s statistical significance ($p$-value). For a method that provides an individual level treatment effect (and estimate of variance) [39], we propose to perform inference for each unique covariate profile, and return those that are statistically significant. The final treatment effect measure is the smallest $p$-value of the covariate profiles. The TESS algorithm, by design, provides the subpopulation it determines to have a statistically significant distributional change (treatment effect) and a measure of this change, so no post-processing is necessary.

### 5.2.1 Detection Power

For any given combination of simulation parameter values ($\delta, num_covs, value_prob$), detection power measures $P(\text{Reject } H_0 \mid H_1(S_{\text{affected}}))$, or how well a method is able to identify the presence of $S_{\text{affected}}$. This is accomplished by comparing the treatment effect measure (score of the detected subset) found under $H_1(S_{\text{affected}})$ to the distribution of the treatment effect measure under $H_0$. More specifically, for a given set of parameter values, we generate a random dataset which only exhibits a treatment effect in the randomly selected subpopulation $S_{\text{affected}}$; each method attempts to detect this subpopulation. As described in §5.2, each method returns a final treatment effect measure for the subpopulation it detects in this affected dataset. For the same dataset, we then conduct randomization testing to determine how significant this treatment effect measure is under $H_0$. We make many copies of the dataset (1000 in our experiments) and in each copy, we generate new outcomes (drawn from $f_0$) such that no subpopulation has a treatment effect. Each method then generates a detected subpopulation and corresponding treatment effect measure for each of these null datasets. These treatment effect measures from the null datasets together provide an empirical estimate of the distribution of the treatment effect measure under $H_0$ for that method. Subsequently, a $p$-value is computed for the treatment effect measure captured under $H_1(S_{\text{affected}})$. This process is repeated many times (300 in our experiments), where each time we 1) generate a random $S_{\text{affected}}$, 2) generate a random dataset under $H_1(S_{\text{affected}})$ and compute each method’s treatment effect measure, and 3) generate 1000 copies of the dataset with effect to compute each method’s treatment effect measure distribution under $H_0$. This process creates 300 $p$-values for each method which describe how extreme each of the $S_{\text{affected}}$ appear under $H_0$. A method rejects $H_0$ for a given $p$-value if it is less than or equal to some test-level $\gamma$, corresponding to the $1 - \gamma$ quantile of the null distribution ($\gamma = 0.05$ in our experiments). Therefore, the detection power $P(\text{Reject } H_0 \mid H_1(S_{\text{affected}}))$ is captured as the proportion of $p$-values that are sufficiently extreme that they lead to the rejection of $H_0$ at level $\gamma$.

### 5.2.2 Detection Accuracy

While detection power measures how well a method identifies the presence of a subpopulation with a treatment effect $S_{\text{affected}}$, as compared to datasets with no treatment effect, detection accuracy measures how well a method can precisely and completely identify the affected subpopulation $S_{\text{affected}}$. Accurately identifying in which subpopulation(s) a treatment effect exists can be crucial, particularly when there is no prior theory to guide which subpopulations to inspect, or when the goal itself is to develop intuition for new theory. As described in §5.2, each of the methods we consider is able to return the subpopulation that it determines as having the most statistically significant treatment effect $S_{\text{detected}}$. Each method will pick out a set of covariate profiles, which could have coherent structure (as with TESS, Causal Tree, and Interaction Tree),
or be an unstructured collection of individually significant covariate profiles (as with Causal Forest). To accommodate both types of subpopulations, we therefore define detection accuracy as

\[
\text{accuracy} = \frac{|S_{\text{detected}} \cap S_{\text{affected}}|}{|S_{\text{detected}} \cup S_{\text{affected}}|} = \frac{\sum R_i \mathbb{1}\{R_i \in S_{\text{detected}} \cap S_{\text{affected}}\}}{\sum R_i \mathbb{1}\{R_i \in S_{\text{detected}} \cup S_{\text{affected}}\}}.
\]

where \(R_i\) are records in the treatment group. This definition of accuracy, commonly known as the Jaccard coefficient, is intended to balance precision (i.e., what proportion of the detected subjects truly have a treatment effect) and recall (i.e., what proportion of the subjects with a treatment effect are correctly detected). We note that \(0 \leq \text{accuracy} \leq 1\); high accuracy values correspond to a detected subset \(S_{\text{detected}}\) that captures many of the subjects with treatment effects and few or no subjects without treatment effects.

### 5.3 Simulation Results

Our first set of results involve a treatment effect that is a mean shift in a normal distribution: the null distribution \(f_0 = N(0, 1)\) and the alternative \(f_1 = N(\delta, 1)\), where \(\delta\) captures the magnitude of the signal (treatment effect). Recall from \([5,2]\) that there are three parameters that we can vary to change the size and magnitude of the signal. For our simulation, we specifically consider \(\delta \in \{0.25, 0.5, \ldots, 3.0\}\), \(\text{num}_\text{cous} \in \{1, 2, \ldots, 10\}\), and \(\text{value}_\text{prob} \in \{0.1, 0.2, \ldots, 0.9\}\); the former controls magnitude of the treatment effect, while the latter two control the concentration of the treatment effect (i.e., the expected size of the affected subpopulation). Instead of considering every combination, we select the middle value of each parameter interval as a reference point (\(\delta = 1.5, \text{num}_\text{cous} = 5, \text{value}_\text{prob} = 0.5\)) and measure performance changes for one parameter, while keeping the others fixed.

Figure 1a shows the changes in each method’s detection power performance as we vary each of the three parameters that contribute to the strength of the treatment effect. From each of the three graphs we observe that TESS consistently exhibits more power than (or equivalent to) the other methods. More importantly, TESS exhibits statistically significant improvements in power for the most challenging ranges of parameter values (i.e., more subtle signals). The top plot varies effect size (or \(\delta\)), which is positively associated with signal strength and negatively associated with detection difficulty; for values 2.0 and below TESS has significantly higher detection power than the competing methods. The middle plot varies the number of covariates selected to have only a subset of values be affected (\(\text{num}_\text{cous}\)). This parameter is negatively associated with signal strength and positively associated with detection difficulty; for values 5 and above, TESS has significantly higher detection power. The bottom plot varies the expected proportion of values, for the selected covariates, which will be affected (\(\text{value}_\text{prob}\)). This parameter is positively associated with signal strength and negatively associated with detection difficulty; for values 0.5 and below TESS exhibits significantly higher detection power. We see that, for sufficiently strong signals (based on both signal magnitude and concentration), all methods are able to distinguish between experiments with and without a subpopulation exhibiting a treatment effect, while TESS provides significant advantages in detection power for weaker signals.

Figure 1b shows the changes in each method’s detection accuracy as we vary each of the three parameters that contribute to the strength of the treatment effect. From each of the three graphs we observe that TESS consistently exhibits significantly higher accuracy than any other method. Recall that we measure subpopulation accuracy as in \([9]\), which captures both precision and recall of the subpopulation returned by a method. The single tree methods tend to have high precision but low recall, resulting in compromised overall accuracy. Intuitively, these results indicate that the truly affected subpopulation is being spread over multiple leaves of the tree, despite its goal of partitioning the data into subpopulations with similar outcomes. This phenomenon may be caused by the greedy search aspect of tree learning: if the tree splits the affected subpopulation between two branches of the tree, the recall of any leaf will be compromised, especially when this split occurs close to the root of the tree. The Causal Forest ensemble method, on the other hand, exhibits relatively higher recall than precision. These results indicate that it is difficult for Causal Forest to distinguish between the covariate profiles that do and do not make up the truly affected subpopulation, as profiles from both sets appear to have statistically significant treatment effects. This
Figure 1: Ability of each method to identify subpopulations with mean shift treatment effects. The three parameters start as fixed ($\delta = 1.5$, num_covs = 5, value_prob = 0.5) and then are varied individually to see how detection ability varies.

Figure 2: Ability of each method to identify subpopulations with an unaffected mean, but distributional treatment effect. The three parameters start as fixed ($\delta = 1.5$, num_covs = 5, value_prob = 0.5) and then are varied individually to see how detection ability varies.
inability stems from the fact that ensemble methods are designed to provide individual level predictions, therefore their conclusions regarding the statistical significance of a covariate profile are made in isolation from the other covariate profiles that also make up the affected subpopulation. Unlike single-tree methods, ensemble methods do not provide coherent and natural groupings of subpopulations. TESS, however, does provide a coherent subpopulation, which seems to balance precision and recall, maintaining a significantly higher subpopulation accuracy.

It is also important to note that the data generating process for these simulations (a treatment effect that occurs as a mean shift between treatment and control distributions) corresponds to the modeling assumptions of the current methods in the literature, which specifically attempt to detect mean shifts, while TESS is designed to detect more general distributional changes. TESS’s improved performance, as compared to the competing methods, in these adverse conditions may be due to its subset-scanning based approach, which combines information across groups of data in an attempt to find exactly and only the affected subset of data. Even if each individual covariate profile that is truly affected exhibits small evidence of a treatment effect, TESS can leverage the group structure and signal of all the affected covariate profiles, and correctly conclude that collectively the subpopulation exhibits significant evidence of a treatment effect. Additionally, the fact that TESS executes its optimization iteratively, unlike the greedy search of tree-based methods, enables it to rectify initial choices of subset that are later determined to be inferior.

Our second set of results considers treatment effects that do not align with the mean shift assumption that pervades the literature. Therefore, the null distribution is still $f_0 = N(0, 1)$; however, the alternative is a mixture distribution $f_1 = \frac{1}{2}N(-\delta, 1) + \frac{1}{2}N(\delta, 1)$. Here $\delta$ still captures the magnitude of the signal (treatment effect), and the remainder of the simulation process remains unchanged. This mixture distribution alternative, however, changes the detection task dramatically: while the average treatment effect is zero, there is still a clear difference in the outcome distribution between treated and control individuals.

Figure 2 shows how each method’s detection power and accuracy change as we vary each of three parameters that contribute to the strength of the treatment effect. If we compare these simulations to those above with a mean shift, TESS exhibits a consistent pattern of high performance, while the performance of the competing methods is dramatically lower. The detection power results indicate that, for the competing methods, it is hard to distinguish even strong distributional changes from random chance, while the accuracy results indicate that their pinpointing of the affected subpopulation is little better than random guessing. Given that there is no observable mean shift in these simulations, these results are consistent with what we expect: TESS is designed to identify more general distributional changes, while the other methods are unable to identify distributional changes without corresponding mean shifts.

### 5.4 A Case Study on Identifying Subpopulations: Tennessee STAR

There appears to be a consensus in the literature that the presence of a teaching aide in a regular-size classroom has an insignificant effect on test scores [16, 24, 33, 41]. (One significant effect was observed in first grade, but this effect was largely considered to be a false positive.) Therefore, we want to use TESS to compare regular classrooms with an aide to regular classrooms without an aide, to determine if there appears to be a subpopulation that was significantly and positively affected by the treatment. To do so, we replicated the analysis of the internal STAR team, using TESS to extend the results, with the goal of demonstrating what the STAR team could have surmised with present-day tools for uncovering heterogeneity. We replicate the original STAR analysis from [33, 41] which includes the sum of the Stanford math and reading scores as the outcome of interest. For the data provided to TESS for detection, we combine the panel data across years and include student’s grade level as a covariate. We would also like to obtain an unbiased estimate of the average treatment effect in the subpopulation identified by TESS. Therefore, we follow a cross-validation paradigm, where the entire dataset is partitioned into ten folds, and iteratively each fold is held out as a validation set (to obtain an estimate of the treatment effect) while the remaining nine folds are provided to TESS (for detection). We further partition the data into records corresponding to students observed in a regular classroom with an aide and a regular classroom without an aide, which serve as treatment and control groups respectively. In three of the ten folds, TESS identified exactly the same subset, which we will call the “detected subpopulation”. Essentially, this detected subpopulation is composed of students
in second or third grade, who attended an inner-city or urban school, receiving instruction from a teacher with 10 or more years of experience\footnote{The detected subpopulation excluded teacher experience between 25 and 30 years. Including this range yields qualitatively the same results and conclusions.}. Therefore, it appears that the presence of an aide raised the test-scores of students exhibiting the selected covariate values described above for grade, school type, and teacher experience, in addition to any values for gender, free-lunch status, teacher ethnicity, and teacher degree. The subpopulations that were returned in each of the ten folds exhibited a large amount of agreement with the detected subpopulation: the fold subpopulations exhibited 88% agreement (on average) with the detected subpopulation on the detection status of a record. The estimated average treatment effect for this detected subpopulation, averaged across all validation folds, is approximately a 34.19 point increase in total test score (36.45 and 22.28 for second and third grades respectively).

Given this consistency across folds, we use the full data to better understand the effect in the detected subpopulation generally. Table 3 shows the evaluation of the treatment effect for all second-grade students (column 1), second-grade students in the detected subpopulation (column 2), and second-grade students in the complement of the detected subpopulation (column 3). Additionally, Figure 3 shows the kernel density plots of the cumulative scores for all second-grade students and students in the detected subpopulation respectively. Figure 3a depicts a strong similarity in the distribution of all second graders’ scores with and without a full-time aide; there is a slight difference around the center of the distribution, but its magnitude is not sufficiently large to be significant, as seen by column 2 of Table 3. Conversely, Figure 3b depicts a difference in test scores for the detected subpopulation of second graders: there appears to be a clear effect of the treatment (dominated by a large mean shift), supported by column 2 of Table 3. We conduct a similar analysis with third graders, and observe similar results in Figure 4 and Table 3. However, the effect of the treatment in third grade appears to result in less of a mean shift, and is better characterized by a change in the skew (third moment) and therefore, the overall form of the distribution (Figure 4b).
Table 3: Table of estimated treatment effects on student test scores in 2nd and 3rd grade. *** indicates \( p < 0.001 \).

|                | All (2\textsuperscript{nd}) | Detected (2\textsuperscript{nd}) | Undetected (2\textsuperscript{nd}) | All (3\textsuperscript{rd}) | Detected (3\textsuperscript{rd}) | Undetected (3\textsuperscript{rd}) |
|----------------|-----------------------------|-----------------------------------|-----------------------------------|-----------------------------|-----------------------------------|-----------------------------------|
| Treatment (std. dev.) | 3.479                       | 36.066***                         | 1.309                             | -0.291                      | 18.703***                         | 0.1                               |
| P-value          | 0.172                       | <0.001                            | 0.637                             | 0.898                       | <0.001                            | 0.968                             |
| Observations     | 4263                        | 620                               | 3643                              | 4063                        | 706                               | 3357                              |

because TESS is able to identify effects that change the distribution (and therefore higher order moments) of test scores, even if the difference in mean score between treatment and control students in third grade was smaller, TESS could potentially still identify the existence of a treatment effect.

There appears to be another consensus in the literature that small classrooms have a consistent, positive, and significant effect [16, 24, 41]; therefore, we also compare small classrooms to regular classrooms, and determine whether there appears to be a subpopulation which is the main driver of this effect. We conduct an analysis as above but with STAR data records corresponding to students observed in a small classroom (treatment group) and a regular classroom (control group). For this analysis, TESS identified the entire population, which is congruent with the previous literature’s analysis of the consistent and significant average treatment effect in each grade. This result from TESS appears to indicate that the effect of small classroom size was not limited to a specific subpopulation. For both TESS analyses, we also conducted permutation testing to compensate for multiple hypothesis testing. Based on these results, we conclude that there is a less than 0.01% chance we would obtain a subpopulation with a score as extreme under the null hypothesis.

The detected subpopulation in the classrooms with aides is not only statistically significant, but may also provide domain insight into the efficacy of full-time aides. A possible explanation for the effect we observe in the detected subpopulation is the fact that 13 schools were chosen at random to have teachers participate in an in-service training session, which the literature has also deemed ineffective [41]. More specifically, 57 teachers were selected each summer from these schools to participate in a three-day in-service to help them teach more effectively in whatever class type they were assigned to; part of the instruction focused on how to work with an aide and also had the aides present. We note that the in-service only occurred during the summers prior to 2nd and 3rd grade, which are the grades identified by TESS. Therefore, it is possible that when provided proper training, the combination of an aide and an experienced teacher can provide a significantly enhanced education environment even in the challenging teaching environments that exist in inner-city and urban schools. An additional explanation is that the educational benefits may be cumulative—i.e., in each additional year a student in this subpopulation has access to the combination of an aide and experienced teacher, the treatment effect compounds—similar to what has been demonstrated in small classrooms for the overall population [24]. However, unlike in small classrooms, for this subpopulation in regular classrooms with an aide, the effects were not large enough to be distinguishable from zero (given the much smaller sample size of the affected subpopulation and smaller treatment effect) until after two years. While a more detailed follow-up analysis of these hypotheses might reveal other causal factors and mechanisms at work, we believe that these results do present evidence that a treatment previously believed to be ineffective may actually have been effective for a particularly vulnerable subpopulation. Therefore, this analysis provides a sense of how TESS can be utilized as a tool for data-driven hypothesis generation in real-world policy analysis.

6 Conclusions

This paper has presented several contributions to the literature on statistical machine learning approaches for heterogeneous treatment effects. We proposed the Distribution Average Treatment Effect (DATE) estimand, which generalizes the focal estimands used in this literature. Moreover, as a specific example of DATE, we derived the Nonparametric Average Treatment Effect (NATE) estimand, which allows detection of treatment effects that manifest as arbitrary effects on the potential outcome distributions (or specific
quantiles), rather than being limited to detection of mean shifts. Furthermore, we consider the challenge of identifying whether any subpopulation has been affected by treatment, and precisely characterizing the affected subpopulation, as opposed to the more typical problem setting of estimating individual-level treatment effects. We formalize the identification of subpopulations with significant treatment effects as an anomalous pattern detection problem, and present the Treatment Effect Subset Scan (TESS) algorithm, which serves as a computationally efficient test statistic for the maximization of our NATE estimand over all subpopulations. We demonstrate that the estimator used by TESS satisfies the linear-time subset scanning property, allowing it to be efficiently and exactly optimized over subsets of a covariate’s values, while evaluating only a linear rather than exponential number of subsets. This efficient conditional optimization step is incorporated into an iterative procedure which jointly maximizes over subsets of values for each covariate in the data: the result is a subpopulation, described as a subset of values for each covariate, which demonstrates the most evidence for a statistically significant treatment effect. In addition to its computational efficiency, we derive desirable statistical properties for the TESS estimator: bounded asymptotic probability of Type I and Type II errors, as well as providing sufficient conditions under the alternative hypothesis that will result in TESS exactly identifying the affected subpopulation. These properties apply more generally to the class of nonparametric scan statistics upon which TESS is built; therefore, this theory also serves as a contribution to the anomalous pattern detection and scan statistics literatures.

In addition to proposing a novel algorithm with desirable properties, we provide an extensive comparison between TESS and other recently proposed statistical machine learning methods for heterogeneous treatment effects (Causal Tree, Interaction Tree, and Causal Forest) through semi-synthetic simulations. Our results indicate that TESS consistently outperforms the other methods in its ability to identify and precisely characterize subpopulations which exhibit treatment effects. TESS significantly outperforms competing methods in the challenging scenarios where the treatment effect signal is weak (i.e., the signal magnitude is low or the affected subpopulation is small) because the subset scanning approach allows it to combine subtle signals across various dimensions of data in order to identify effects of interest. Moreover, TESS’s detection performance is consistent even when the treatment outcome distribution in the affected subpopulation has the same mean as the control outcome distribution, while the competing methods demonstrate essentially no ability to identify the affected subpopulation in the absence of a mean shift.

After demonstrating TESS’s performance through simulation, we explore the well-known Tennessee STAR experiment, searching for previously unidentified subpopulations with significant treatment effects. As a result of this analysis, TESS uncovered an intuitive subpopulation that seems to have experienced extremely significant improved test scores as a result of having a teacher’s aide in the classroom, a treatment that has consistently been considered ineffective (as measured by the average treatment effect) by the literature on the Tennessee STAR. This provides a sense of how TESS can be utilized as a tool for generating hypotheses to be further explored and tested. We do however caution researchers to view algorithms like TESS not as a replacement, but rather an assistive tool, for developing scientific and behavioral theory. Results discovered by these methods should be investigated further and evaluated to develop a deeper theoretical understanding of the phenomena they uncover. When used to this end, these tools fill a critical void: in many contexts it is rare to know a priori which hypotheses are relevant and supported by data, and the use of traditional methods (e.g., regression) puts the onus on the researcher to know which hypothesis to test. This process necessitates that theory comes first, and subsequent investigation is a form of confirmatory analysis. However, such a process can become an impediment to data-driven discovery: there is an increasing need for scalable methods to use (big) data to generate new hypotheses, rather than just confirming pre-existing beliefs.

In the late 1970s, John W. Tukey began to outline his vision for the future of statistics, which included a symbiotic relationship between exploratory and confirmatory data analysis. He argues these two forms of data analysis “can–and should–proceed side by side” because he believed ideas “come from previous exploration more often than from lightning strokes” To this end Tukey advocates for using data to suggest hypotheses to test, or what we now call data-driven hypothesis generation. We see our work as the natural evolution of Tukey’s vision of data analysis: we develop an approach—rigorously conducted and theoretically grounded—to conduct exploratory analysis in randomized experiments, with the hope of catalyzing “lightning strokes” of discovery and the advancement of science.
References

[1] J. U. Adams. Genetics: Big hopes for big data. Nature, 527:S108–S109, Nov 2015.

[2] A. Alyass, M. Turcotte, and D. Meyre. From big data analysis to personalized medicine for all: challenges and opportunities. BMC Medical Genomics, 8(1):613, Jun 2015.

[3] M. L. Anderson. Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry preschool, and Early Training projects. Journal of the American Statistical Association, 103:1481–1495, Dec 2008.

[4] J. D. Angrist and J.-S. Pischke. Mostly Harmless Econometrics: An Empiricist’s Companion. Princeton University Press, Dec 2008.

[5] F. J. Anscombe. Graphs in statistical analysis. The American Statistician, 27(1):17, Feb 1973.

[6] E. Arias-Castro, R. M. Castro, E. Tnczos, and M. Wang. Distribution-free detection of structured anomalies: Permutation and rank-based scans. Journal of the American Statistical Association, 0(ja):0–0, 2017.

[7] S. F. Assmann, S. J. Pocock, L. E. Enos, and L. E. Kasten. Subgroup analysis and other (mis)uses of baseline data in clinical trials. The Lancet, 355:1064–1069, 2000.

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

[9] W. S. Barnett. Benefit-cost analysis of the Perry preschool program and its policy implications. Educational evaluation and policy analysis, 7(4):333–342, Jan 1985.

[10] R. H. Berk and D. H. Jones. Goodness-of-fit test statistics that dominate the Kolmogorov statistics. Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete, 47:47–59, 1979.

[11] F. Chen and D. B. Neill. Non-parametric scan statistics for event detection and forecasting in heterogeneous social media graphs. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’14, pages 1166–1175, 2014.

[12] J. N. Cohn, D. G. Archibald, S. Ziesche, and Others. Effect of vasodilator therapy on mortality in chronic congestive heart failure. New England Journal of Medicine, 314(24):1547–1552, Jun 1986.

[13] J. N. Cohn, G. Johnson, S. Ziesche, and Others. A comparison of enalapril with hydralazine–isosorbide dinitrate in the treatment of chronic congestive heart failure. New England Journal of Medicine, 325(5):303–310, Aug 1991.

[14] E. Duflo. Field experiments in development economics. In R. Blundell, W. Newey, and T. Persson, editors, Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress, volume 2, pages 322–348. 2006.

[15] A. Dvoretzky, J. Kiefer, and J. Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. The annals of mathematical statistics, 27(3):642–669, Sep 1956.

[16] J. Folger and C. Breda. Evidence from project STAR about class size and student achievement. Peabody Journal of Education, 67(1):17–33, Sep 1989.

[17] J. H. Friedman. Multivariate adaptive regression splines. Annals of Statistics, 19(1):1–67, Mar 1991.

[18] P. Gaenssler and J. A. Wellner. Glivenko–Cantelli Theorems. John Wiley & Sons, Inc., Hoboken, NJ, USA, Jul 2004.
[19] 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.

[20] J. Grimmer, S. Messing, and S. J. Westwood. Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods. *Political Analysis*, Conditional Acceptance, 2017.

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

[22] L. Jager and J. A. Wellner. Goodness-of-fit tests via phi-divergences. *The Annals of Statistics*, 35(5):2018–2053, Oct 2007.

[23] 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*, New York, NY, USA, 2013.

[24] A. B. Krueger. Experimental estimates of education production functions. *Quarterly Journal of Economics*, 114(2):497–532, May 1999.

[25] E. McFowland III, S. D. 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.

[26] D. B. Neill. Fast subset scan for spatial pattern detection. *Journal of the Royal Statistical Society (Series B: Statistical Methodology)*, 74(2):337–360, 2012.

[27] D. B. Neill, E. McFowland III, and H. Zheng. Fast subset scan for multivariate event detection. *Statistics in medicine*, 32(13):2185–2208, 2013.

[28] B. Nye, L. V. Hedges, and S. Konstantopoulos. The effects of small classes on academic achievement: The results of the Tennessee class size experiment. *American Educational Research Journal*, 37(1):123–151, 2000.

[29] S. Saul. F.D.A. Approves a Heart Drug for African-Americans, June 2005. [https://www.nytimes.com/2005/06/24/health/fda-approves-a-heart-drug-for-african-americans.html](https://www.nytimes.com/2005/06/24/health/fda-approves-a-heart-drug-for-african-americans.html) (Accessed: 06-05-2018).

[30] L. J. Schweinhart, H. V. Barnes, and D. P. Weikart. Significant benefits: The High/Scope Perry preschool study through age 27. Technical Report 10, High/Scope Educational Research Foundation, 1993.

[31] G. R. Shorack and J. A. Wellner. *Empirical Processes with Applications to Statistics*. Society for Industrial and Applied Mathematics, May 2012.

[32] S. D. Speakman, E. McFowland III, and D. B. Neill. Scalable Detection of Anomalous Patterns With Connectivity Constraints. *Journal of Computational and Graphical Statistics*, 24(4):1014–1033, Dec 2015.

[33] J. Stock and M. Watson. *Introduction to Econometrics 2nd edition*. Pearson, 2007.

[34] X. Su, C.-L. Tsai, H. Wang, D. M. Nickerson, and B. Li. Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, 10:141–158, Dec 2009.

[35] 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.

[36] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, 58:267–288, 1996.
[37] J. W. Tukey. *Exploratory Data Analysis*. Addison-Wesley Publishing Company, 1977.

[38] J. W. Tukey. We need both exploratory and confirmatory. *The American Statistician*, 34(1), 1980.

[39] S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 2018.

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

[41] E. R. Word, J. Johnston, H. P. Bain, and Others. The state of Tennessee’s Student/Teacher Achievement Ratio (STAR) project: Technical report 19851990. *Nashville: Tennessee State Department of Education*, 1990.
A Score Functions

To begin we revisit the general form of the score function—or equivalently the treatment effect test statistic—that we refer to as the nonparametric scan statistic. Additionally, we establish equivalences, as different forms will lend themselves to various proof strategies we implement later.

\[ \max_S F(S) = \max_{S, \alpha} F_\alpha(S) = \max_{S, \alpha, \beta} F_{\alpha, \beta}(S) \]

To begin we revisit the general form of the score function—or equivalently the treatment effect test statistic—thus the substitution

\[ \max_{S, \alpha, \beta} F_{\alpha, \beta}(S) = \max_{S, \alpha, \beta} \sum_{x \in U_X(S)} \omega(\alpha, \beta, N_\alpha(x), N(x)). \] (10)

First, we demonstrate that our empirical distribution functions from the control and treatment groups are unbiased estimators under the assumption of unconfoundedness.

Lemma 1. If \( Y_i(0), Y_i(1) \perp W_i \mid X_i \), then \( \hat{F}_{Y|X=x} \) and \( \hat{F}_{Y^T|X=x} \) are unbiased estimators of \( F_{Y(0)|X=x} \) and \( F_{Y(1)|X=x} \) respectively.

Proof.

\[
\mathbb{E} \left[ \hat{F}_{Y^C|X=x}(y) \right] = \mathbb{E}_{Y|X=x} \left[ \sum_{Q_i:X_i=x} \mathbb{I} \{W_i = 0\} \mathbb{I} \{Y_i^{\text{obs}} \leq y\} \right] = \sum_{Q_i:X_i=x} \mathbb{E}_{Y|W=0} \mathbb{E}_{Y|X=x} \left[ \mathbb{E}_{Y|W=0, X_i=x} \mathbb{I} \{Y_i(0) \leq y\} \right] = \sum_{Q_i:X_i=x} \mathbb{E}_{Y|X=x} \left[ \mathbb{E}_{Y_i(0) \leq y} \right] = \sum_{Q_i:X_i=x} \mathbb{E}_{Y_i(0) \leq y} = F_{Y(0)|X=x}(y).
\]

A similar argument shows that \( \mathbb{E} \left[ \hat{F}_{Y^T|X=x}(y) \right] = F_{Y(1)|X=x}(y) \), assuming that unconfoundedness holds and thus the substitution \( \mathbb{E}_{Y|W_i=1, X_i=x} \mathbb{I} \{Y_i(1) \leq y\} = \mathbb{E}_{Y_i(0) \leq y} \) can be made. □

Corollary 2. As a direct consequence of \( \mathbb{E} \left[ \hat{F}_{Y^C|X=x}(y) \right] = F_{Y(0)|X=x}(y) \), we also have that \( \mathbb{E} \left[ \hat{F}_{Y^C|X=x}(y) \right] \) is strongly consistent, \( \| \hat{F}_{Y^C|X=x} - F_{Y(0)|X=x} \|_\infty \xrightarrow{a.s.} 0 \). The rate of convergence is exponential in sample size, \( P \left( \| \hat{F}_{Y^C|X=x} - F_{Y(0)|X=x} \|_\infty > \epsilon \right) \leq 2e^{-2n_0\epsilon^2}, \epsilon > 0 \), where \( n_0 \) is the number of control units with \( X_i = x \). Similar arguments apply for \( \hat{F}_{Y^T|X=x} \).

Given these properties of the empirical conditional distribution functions, we can now turn our attention to the score function. In [35] we introduced two score functions: Berk-Jones, \( F_{\alpha}^{BJ}(S) = N(S) \text{Div}_{KL} \left( \frac{N_\alpha(S)}{N(S)}, \alpha \right) \), where \( \text{Div}_{KL} \) is the Kullback-Leibler divergence, and the Normal Approximation, \( F_{\alpha}^{NA}(S) = N(S) \text{Div}_{\chi^2} \left( \frac{N_\alpha(S)}{N(S)}, \alpha \right) = \frac{(N_\alpha(S) - N(S)\alpha)^2}{2N(S)\alpha(1-\alpha)} \). There are a collection of well-known supremum goodness-of-fit statistics used in the literature, all of which are described in [22], that are each a transformation of \( F_{\alpha}^{NA}(S) \):
the Kolmogorov-Smirnov statistic
\[ F_{KS}(S) = \max_{\alpha} F_{\alpha}^{KS}(S) = \max_{\alpha} \frac{(N_\alpha(S) - N(S)\alpha)}{\sqrt{N(S)}} = \max_{\alpha} \sqrt{2\alpha(1-\alpha)}F_{\alpha}^{NA}(S), \]

the Cramer-von Mises statistic
\[ F_{CV}(S) = \max_{\alpha} F_{\alpha}^{CV}(S) = \max_{\alpha} \frac{(N_\alpha(S) - N(S)\alpha)^2}{N(S)} = \max_{\alpha} 2\alpha(1-\alpha)F_{\alpha}^{NA}(S), \]

the Higher-Criticism statistic
\[ F_{HC}(S) = \max_{\alpha} F_{\alpha}^{HC}(S) = \max_{\alpha} \frac{(N_\alpha(S) - N(S)\alpha)}{\sqrt{N(S)\alpha(1-\alpha)}} = \max_{\alpha} \sqrt{2F_{\alpha}^{NA}(S)}, \]

and the Anderson-Darling statistic
\[ F_{AD}(S) = \max_{\alpha} F_{\alpha}^{AD}(S) = \max_{\alpha} \frac{(N_\alpha(S) - N(S)\alpha)^2}{N(S)\alpha(1-\alpha)} = \max_{\alpha} 2F_{\alpha}^{NA}(S). \]

As a result of this connection between \( F_{NA}(S) \) and these other statistics, we have the following:

**Lemma 2.** If \( S \) maximizes \( F_{\alpha}^{NA}(S) \), then it maximizes \( F_{\alpha}^{KS}(S), F_{\alpha}^{CV}(S), F_{\alpha}^{HC}(S) \) and \( F_{\alpha}^{AD}(S) \).

**Proof.** First, we note that \( T(F_{\alpha}^{NA}(S)) \), where \( T(x) = (b\alpha)^a \), for \( b \in \{1, 2, 2\alpha(1-\alpha)\} \) and \( a \in \{1, \frac{1}{2}\} \) is a monotonically increasing transformation. Therefore, \( \arg\max_S F_{\alpha}^{NA}(S) = \arg\max_T T(F_{\alpha}^{NA}(S)) \), because \( \arg\max \) is invariant to monotone transformations. 

### B Supplementary Materials: Proofs of Lemmas and Theorems

In this section, we provide detailed proofs of the Lemmas and Theorems stated in the main text. There are additional Lemmas presented here that are not stated in the main text, but are still beneficial in support of our Theorems. Before presenting the proofs, we (re-)introduce notation that will be used throughout the proofs.

#### B.1 Notation

\( S_T \): the truly affected (rectangular) subset.

\( S^* \): the highest scoring (rectangular) subset, \( \arg\max_{S \in R} F(S) \), where \( R \) is the space of all rectangular subsets.
\( \alpha^* \): the \( \alpha \) at which \( S^* \) is highest scoring, \( \arg \max_{\alpha} F_\alpha(S^*) \).

\( S^*_u \): the highest scoring unconstrained subset, \( \arg \max_S F(S) \).

\( \alpha^*_u \): the \( \alpha \) at which \( S^*_u \) is highest scoring, \( \arg \max_{\alpha} F_\alpha(S^*_u) \).

\( U_X \): a function which returns the unique covariate profiles (non-empty tensor cells) in a set.

\( M = |U_X(D)| \), the number of unique covariate profiles in our data, or equivalently the number of non-empty cells in our data tensor.

\( k = \frac{|U_X(S^T)|}{|U_X(D)|} \), the proportion of non-empty cells that are affected under \( H_1(S^T) \).

\( \beta(\alpha) : P(p(y) \leq \alpha | H_1(S^T)) \), for all the p-values of covariate profiles in \( S^T \).

\( h(M, \epsilon) \): the critical value for the test statistic, \( \max_{S \in R} F(S) \), for a given \( M \).

\( \Phi \): Cumulative distribution function of standard normal.

\( \phi \): Probability density function of standard normal.

### B.2 Statistical Properties

We now demonstrate desirable statistical properties of \( F_{BJ}(S) \) and \( F_{NA}(S) \); these properties will also extend to the other statistics described in Appendix A because of their close relationship to \( F_{NA}(S) \). More specifically, we demonstrate that using \( F(S) \) we can appropriately (fail to) reject \( H_0 \) with high probability.

The results derived in this section assume \( N(x) \geq n \) for all \( x \in U_X(D) \), i.e., each unique covariate profile in the data has at least \( n \) data points, and we consider the case where \( M, n \rightarrow \infty \). We would like to show that for the same critical value \( h(M, \epsilon) \) we have the following:

\[
P_{H_0} \left( \max_{S \in R} F(S) > h(M, \epsilon) \right) \rightarrow 0,
\]

\[
P_{H_1} \left( \max_{S \in R} F(S) > h(M, \epsilon) \right) \rightarrow 1.
\]

Toward this pursuit, the first result we show is that in the limit \( F_{BJ} \) is well approximated by \( F_{NA} \), which will then allow us to focus the remainder of our results on \( F_{NA} \) specifically.

**Lemma 3.** \( F_{BJ}(S) \simeq F_{NA}(S) \) as \( N(S) \rightarrow \infty \).

**Proof.** Recall that \( K(x, y) = \text{DivKL}(x, y) = x \log \frac{x}{y} + (1 - x) \log \frac{1 - x}{1 - y} \). By expanding \( K(x, y) \) through a Taylor series, we have

\[
K(x, y) = K(y, y) + \frac{\partial K(x, y)}{\partial x} \bigg|_{x = y} (x - y) + \frac{\partial^2 K(x, y)}{\partial^2 x} \bigg|_{x = y'} \frac{(x - y)^2}{2}.
\]

\[
= 0 + 0 + \frac{(x - y)^2}{2y'(1 - y')}
\]
for some $y'$ such that $|y' - x| \leq |y - x|$. Therefore,

\[
F^{BJ}(S) = \max_\alpha N(S) K \left( \frac{N_a(S)}{N(S)}, \alpha \right)
\]

\[
= \max_\alpha N(S) \frac{\left( \frac{N_a(S)}{N(S)} - \alpha \right)^2}{2\alpha(1 - \alpha)} \quad \left( \text{where } \left| \alpha' - \frac{N_a(S)}{N(S)} \right| \leq \left| \alpha - \frac{N_a(S)}{N(S)} \right| \right)
\]

\[
\leq \max_\alpha N(S) \left[ \frac{\left( \frac{N_a(S)}{N(S)} - \alpha \right)^2}{2\alpha(1 - \alpha)} \right.
\quad \left. \vee \quad \frac{\left( \frac{N_a(S)}{N(S)} - \alpha \right)^2}{2\alpha(1 - \alpha)} \right]
\]

\[
\geq \max_\alpha N(S) \left[ \frac{\left( \frac{N_a(S)}{N(S)} - \alpha \right)^2}{2\alpha(1 - \alpha)} \right]
\]

Furthermore, under $H_0$, \(\frac{N_a(S)}{N(S)} \xrightarrow{a.s.} \alpha \Rightarrow \alpha' \xrightarrow{a.s.} \alpha\), which by the continuous mapping theorem results in

\[
F^{BJ}(S) = \max_\alpha N(S) \frac{\left( \frac{N_a(S)}{N(S)} - \alpha \right)^2}{2\alpha(1 - \alpha)} = F^{NA}(S).
\]

However, under $H_1$ (\(S^T\)), \(\frac{N_a(S)}{N(S)} \xrightarrow{a.s.} \beta(\alpha)\), therefore asymptotically for $F^{BJ}(S)$ we have,

\[
\max_\alpha N(S) \frac{\left( \frac{N_a(S)}{N(S)} - \alpha \right)^2}{2\alpha(1 - \alpha)} \quad \left( 1 \wedge \frac{\alpha(1 - \alpha)}{\beta(\alpha)(1 - \beta(\alpha))} \right) \leq F^{BJ}(S)
\]

\[
\leq \max_\alpha N(S) \frac{\left( \frac{N_a(S)}{N(S)} - \alpha \right)^2}{2\alpha(1 - \alpha)} \quad \left( 1 \wedge \frac{\alpha(1 - \alpha)}{\beta(\alpha)(1 - \beta(\alpha))} \right).
\]

We can see that $F^{BJ}(S)$ is bounded above and below by either $F^{NA}(S)$ or a constant times $F^{NA}(S)$. □

Now, we will show that when the null hypothesis is true—i.e., the treatment does not have an effect—in the limit of large $M$ and $n$, the score of the most anomalous subset is linear in $M$ and constant in $n$.

**Lemma 4.** Under $H_0$ defined in [8], \(F^{NA}(S^*_u) \xrightarrow{P} \max_x \frac{M\phi(z)^2}{2(1 - \Phi(z))} \approx 0.202 M\), where $\phi$ and $\Phi$ are the Gaussian pdf and cdf respectively.

**Proof.** From Theorem 1, we know that if \(\{x_{(1)}, \ldots, x_{(M)}\}\) are data elements—and specifically in this context are the $M$ unique covariate profiles in \(U_X(D)\)—sorted according to their priority function, which for covariate profiles is \(\frac{N_a(x)}{N(x)}\), where \(x_{(t)}\) has the $t^{th}$ highest priority, then

\[
S^*_u = \{x_{(1)}, \ldots, x_{(t)}\}_{t \in [1,M]} = \left\{ x \left| \frac{N_a(x)}{N(x)} > t(\alpha) \right. \right\}.
\]

With Bin representing the Binomial distribution, for each of the unique covariate profiles $x \in U_X(D)$, $N_a(x) \sim \text{Bin}(N(x), \alpha)$. Given that $N(x) \geq n$, we have that asymptotically $P(\text{Bin}(N(x), \alpha) > t(\alpha)N(x))$ is upper bounded by $P(\text{Bin}(n, \alpha) > t(\alpha)n)$ for fixed $\alpha$ and $t(\alpha) > \alpha$, so we can focus on the simple case $N(x) = n$ for all $x$. Therefore, asymptotically, $|S^*_u| \sim \text{Bin}\left(M, P(\text{Bin}(n, \alpha) > t(\alpha)n)\right)$, $\mathbb{E}_{H_0}[N(S^*_u)] = \frac{M}{2}$.
Furthermore, this implies, 

$$\frac{N_\alpha(S_n^*)}{N(S_n^*)} \xrightarrow{a.s.} \mathbb{E} \left[ \frac{N_\alpha(x)}{n}, \frac{N_\alpha(n > t(\alpha))}{n} \right] = \mathbb{E} \left[ \frac{X \sim \text{Bin}(n, \alpha)}{n}, \frac{X > t(\alpha)}{n} \right]$$

$$\approx \sqrt{n} \left( \frac{X - \alpha}{\alpha(1 - \alpha)} \right) > Z'(\alpha) \quad \left( \text{with } Z'(\alpha) = \frac{\sqrt{n}(t(\alpha) - \alpha)}{\sqrt{\alpha(1 - \alpha)}} \right)$$

Using the definition of $F^{NA}(S_n^*)$ we have

$$F^{NA}(S_n^*) = \max_{\alpha} F^{NA}_\alpha(S_n^*)$$

$$= \max_{\alpha} \left( \frac{N_\alpha(S_n^*)}{N(S_n^*)} - \alpha \right)^2 \frac{N(S_n^*)}{2\alpha(1 - \alpha)}$$

$$\Rightarrow \max_{\alpha} \left( \frac{\sqrt{n}\alpha(1 - \alpha)}{n} \frac{\phi(Z'(\alpha))}{(1 - \Phi(Z'(\alpha)))} \right)^2 \frac{Mn(1 - \Phi(Z'(\alpha)))}{2\alpha(1 - \alpha)}$$

$$= \max_{\alpha} \frac{M\phi(Z'(\alpha))^2}{2(1 - \Phi(Z'(\alpha)))} \approx 0.202 M,$$

where [12] is a result of the continuous mapping theorem and [11]. Furthermore, the convergence in probability of [12] is a result of the convergence in distribution to a constant.

However, these asymptotic results fail when $\alpha$ is allowed to become arbitrarily small, decreasing to zero as $n$ increases. A simple solution is to fix constants $\alpha_{\min} > 0$ and $\alpha_{\max} < 1$ and define $F^{NA}(S_n^*)$ as the maximum over $F^{NA}_{\alpha}(S_n^*)$ for $\alpha \in [\alpha_{\min}, \alpha_{\max}]$. Restricting the range of $\alpha$ values solves the asymptotic convergence issues for the $F^{NA}$, $F^{HC}$, $F^{AD}$, and $F^{BJ}$ score functions, while the $F^{KS}$ and $F^{CV}$ statistics converge for unrestricted $\alpha$.

Now we can use these asymptotic results to bound the probability that the highest scoring rectangular subset $F^{NA}(S^*)$ exceeds a threshold under the null hypothesis, again maximizing $F^{NA}$ over a range of $\alpha$ values from $\alpha_{\min} > 0$ to $\alpha_{\max} < 1$.

**Theorem 2.** Under $H_0$ defined in [8], let $N(x) \geq n, \forall x \in U_X(D)$, fix $\epsilon > 0$, and assume $M, n \rightarrow \infty$; then there exist a constant $C \leq \max Z \frac{\phi(Z)^2}{2(1 - \Phi(Z))} \approx 0.202$ and critical value $h(M, \epsilon) = CM + \epsilon$ such that

$$P_{H_0} \left( \max_{S \in R} F(S) > h(M, \epsilon) \right) \rightarrow 0.$$

**Proof.** First, we note that under $H_0$, $F(S^*) \leq F(S_n^*)$, because the detected subset $S^*$ is the arg max over all rectangular subsets, while $S_n^*$ is the arg max over all subsets. We now consider the score function $F^{NA}$
and the critical value \( h(M, \epsilon) = \max_\alpha \frac{M \phi(Z'(\alpha))^2}{2(1 - \Phi(Z'(\alpha)))} + \epsilon \), for any \( \epsilon > 0 \).

\[
P_{H_0}(\text{Reject } H_0) = P_{H_0}\left( F^{N^A}(S^*) > h(M, \epsilon) \right)
= P_{H_0}\left( F^{N^A}(S^*) > \max_\alpha \frac{M \phi(Z'(\alpha))^2}{2(1 - \Phi(Z'(\alpha)))} + \epsilon \right)
\leq P_{H_0}\left( F^{N^A}(S^*_u) > \max_\alpha \frac{M \phi(Z'(\alpha))^2}{2(1 - \Phi(Z'(\alpha)))} + \epsilon \right)
= P_{H_0}\left( F^{N^A}(S^*_u) - \max_\alpha \frac{M \phi(Z'(\alpha))^2}{2(1 - \Phi(Z'(\alpha)))} > \epsilon \right)
\leq P_{H_0}\left( \left| F^{N^A}(S^*_u) - \max_\alpha \frac{M \phi(Z'(\alpha))^2}{2(1 - \Phi(Z'(\alpha)))} \right| > \epsilon \right)
\xrightarrow{P} 0,
\]

where the final line follows from Lemma 4. Furthermore, from Lemma 3 we have \( F^{BJ}(S^*_u) \xrightarrow{a.s.} F^{N^A}(S^*_u) \) under \( H_0 \), which implies that this result holds for \( F^{BJ} \). Finally, by Lemma 2 all other score functions under consideration are maximizations over a monotonic transformation \((T(F)) \) of the continuous function \( F^{N^A}(S) \); therefore, the result for \( \max_\alpha F^{N^A}(S) \) will have a direct analogy for \( \max_\alpha T(F^{N^A}(S)) \).

Next, we analyze the score of the truly affected subset, when the null hypothesis is false.

**Lemma 5.** Under \( H_1(S^T) \), \( F^{N^A}(S^T) \xrightarrow{a.s.} \max_\alpha (\beta(\alpha) - \alpha)^2 \frac{kmn}{2a(1 - \alpha)} \).

**Proof.** First, recognize that \( N(S^T) \xrightarrow{P} kmn \) and recall that \( \mathbb{E}_{H_1(S^T)}[N_\alpha(S^T)] = N(S^T) \beta(\alpha) \). Therefore, we have the following:

\[
F^{N^A}(S^T) = \max_\alpha F^{N^A}_\alpha(S^T)
= \max_\alpha \left( \frac{N_\alpha(S^T) - N(S^T) \alpha}{2N(S^T) \alpha(1 - \alpha)} \right)^2
= \max_\alpha \left( \frac{N_\alpha(S^T)}{N(S^T)} - \alpha \right)^2
\xrightarrow{a.s.} \beta(\alpha).
\]

Finally, by the continuous mapping theorem we have

\[
F(S^T) \xrightarrow{a.s.} \max_\alpha (\beta(\alpha) - \alpha)^2 \frac{kmn}{2a(1 - \alpha)}.
\]

**Theorem 3.** Under \( H_1(S^T) \) defined in [3], let \( N(x) \geq n \forall x \in U_X(D) \), fix \( \epsilon > 0 \), and assume \( M, n \to \infty \); then for the same critical value \( h(M, \epsilon) \) as in Theorem 3,

\[
P_{H_1}\left( \max_{S \in R} F(S) > h(M, \epsilon) \right) \to 1.
\]

33
In this section, we are still interested in studying the properties of our framework under $B.3$ Subset Correctness. Now that we have a lower bound on $F(S)$ under $H_1(S^T)$, we consider the critical value $h(M, \epsilon) = \max_\alpha \frac{M\phi(Z'(\alpha))}{2(1-\Phi(Z'(\alpha)))} + \epsilon$, for any $\epsilon > 0$, and the score function $F^{NA}$.

$$P_{H_1} (\text{Reject } H_0) = P_{H_1} (F^{NA}(S^*) > h(M, \epsilon))$$

because the detected subset $S^*$ is a maximization over all rectangular subsets while $S^T$ is one such subset. Now that we have a lower bound on $F(S)$ under $H_1(S^T)$, we consider the critical value $h(M, \epsilon) = \max_\alpha \frac{M\phi(Z'(\alpha))}{2(1-\Phi(Z'(\alpha)))} + \epsilon$, for any $\epsilon > 0$, and the score function $F^{NA}$.

$$P_{H_1} (\text{Reject } H_0) = P_{H_1} (F^{NA}(S^*) > h(M, \epsilon))$$

First, we note that under $BJ$ the collection of these data elements, i.e., $F$ subset.

Proof. First, we note that from the derivations of $F^{BJ}(S)$ and $F^{NA}(S)$ in §3.5 that if we do not set $\beta =$

B.3 Subset Correctness

In this section, we are still interested in studying the properties of our framework under $H_1(S^T)$, however we are now concerned about the correctness of our detected subset $S^*$: as the objective is for the detected subset $S^* = S^T$. If $x$ is a data element, i.e., one of the $M$ unique covariate profiles in the data; $U_X(D)$ is the collection of these data elements, i.e., $U_X(D) = \{x_1, \ldots, x_M\}$; and both $U_X(S^T), U_X(S^T) \subseteq U_X(D)$. The results in this section are general, and are therefore applicable to an unconstrained (or constrained) $S^T$; therefore the result for $\max_\alpha F^{NA}(S)$ will have a direct analogy for $\max_\alpha T(F^{NA}(S))$. 

where (13) follows from Lemma 3. Furthermore, from Lemma 3 we have $F^{BJ}(S^*) \overset{a.s.}{\rightarrow} F^{NA}(S^*)$ under $H_0$, which implies that comparing $F^{BJ}$ to the same critical value $h(M, \epsilon)$ will also yield the above result. Finally, by Lemma 2 all other score functions under consideration are maximizations over a monotonic transformation $(T(F_\alpha))$ of the continuous function $F^{NA}(S)$; therefore, the result for $\max_\alpha F^{NA}(S)$ will have a direct analogy for $\max_\alpha T(F^{NA}(S))$. 

Lemma 6. $F(S)$ can be written as $\max_\alpha, \beta \sum_{x \in U_X(S)} \omega(\alpha, \beta, N_\alpha(x), N(x))$, for $\alpha, \beta \in (0, 1)$ representing quantile values of the control and treatment potential outcomes distributions respectively.

Proof. First we note that from the derivations of $F^{BJ}(S)$ and $F^{NA}(S)$ in §3.5 that if we do not set $\beta =$
Lemma 7. \( \omega_{\text{mle}}(S) \) but instead treat \( \beta \in (0, 1) \) as a given quantity, then

\[
F^{BJ}(S) = \max_{\alpha} F^{BJ}_{\alpha}(S)
= \max_{\alpha, \beta} F^{BJ}_{\alpha, \beta}(S)
= \max_{\alpha, \beta} N_\alpha(S) \log \left( \frac{\beta}{\alpha} \right) + (N(S) - N_\alpha(S)) \log \left( \frac{1 - \beta}{1 - \alpha} \right)
= \max_{\alpha, \beta} N_\alpha(S) \log \left( \frac{\beta}{\alpha} \right) - N_\alpha(S) \log \left( \frac{1 - \beta}{1 - \alpha} \right) + N(S) \log \left( \frac{1 - \beta}{1 - \alpha} \right)
= \max_{\alpha, \beta} N_\alpha(S) \log \left( \frac{\beta(1 - \alpha)}{\alpha(1 - \beta)} \right) + N(S) \log \left( \frac{1 - \beta}{1 - \alpha} \right)
= \max_{\alpha, \beta} \sum_{x \in U_X(S)} \left( \frac{\beta(1 - \alpha)}{\alpha(1 - \beta)} \right) N_\alpha(x) + \log \left( \frac{1 - \beta}{1 - \alpha} \right) N(x)
= \max_{\alpha, \beta} \sum_{x \in U_X(S)} C^{BJ}_\alpha \beta N_\alpha(x) + C^{BJ}_\alpha N(x)
= \max_{\alpha, \beta} \sum_{x \in U_X(S)} \omega^{BJ}(\alpha, \beta, N_\alpha(x), N(x))
\]

\[
F^{NA}(S) = \max_{\alpha} F^{NA}_{\alpha}(S)
= \max_{\alpha, \beta} F^{NA}_{\alpha, \beta}(S)
= \max_{\alpha, \beta} N_\alpha(S) \frac{(\beta - \alpha)}{\alpha(1 - \alpha)} + \frac{N(S)(\alpha^2 - \beta^2)}{2\alpha(1 - \alpha)}
= \max_{\alpha, \beta} \frac{(\beta - \alpha)}{\alpha(1 - \alpha)} \left( \sum_{x \in U_X(S)} N_\alpha(x) \right) + \frac{\alpha^2 - \beta^2}{2\alpha(1 - \alpha)} \left( \sum_{x \in U_X(S)} N(x) \right)
= \max_{\alpha, \beta} \sum_{x \in U_X(S)} \frac{(\beta - \alpha)}{\alpha(1 - \alpha)} N_\alpha(x) + \frac{\alpha^2 - \beta^2}{2\alpha(1 - \alpha)} N(x)
= \max_{\alpha, \beta} \sum_{x \in U_X(S)} C^{NA}_\alpha \beta N_\alpha(x) + C^{NA}_\alpha N(x)
= \max_{\alpha, \beta} \sum_{x \in U_X(S)} \omega^{NA}(\alpha, \beta, N_\alpha(x), N(x))
\]

where all the \( C_{\alpha, \beta} \)'s are constants with respect to given values of \( \alpha, \beta \).

We now have that the score of a subset \( S \) can be decomposed into the sum of contributions (measured by a function \( \omega \)) from each individual element contained within the subset. Next, we seek to demonstrate some important properties of the \( \omega \) functions. More specifically, \( \omega \) is a concave function with respect to \( \beta \), which has two roots and a unique maximum.

**Lemma 7.** \( \omega^{NA}(\alpha, \beta, N_\alpha(x), N(x)) \) is concave with respect to \( \beta \), maximized at \( \beta_{\text{mle}}(x) = \frac{N_\alpha(x)}{N(x)} \), and has two roots \( (\beta_{\text{min}}(x), \beta_{\text{max}}(x)) \).
Proof. Firstly,

\[
\frac{\partial \omega^{NA}(\alpha, \beta, N_\alpha(x), N(x))}{\partial \beta} = \frac{N_\alpha(x) - N(x)\beta}{\alpha(1 - \alpha)} \\
= -\frac{N(x)}{\alpha(1 - \alpha)}\beta + \frac{N_\alpha(x)}{\alpha(1 - \alpha)} \\
\text{(set) } 0 = -\frac{N(x)}{\alpha(1 - \alpha)}\beta + \frac{N_\alpha(x)}{\alpha(1 - \alpha)} \\
0 = -N(x)\beta + N_\alpha(x) \\
\beta = \frac{N_\alpha(x)}{N(x)},
\]  

(14)

shows that the first derivative is the equation of a line, with a negative slope, and (15) shows that this line has one root at \(\frac{N_\alpha(x)}{N(x)}\). This implies \(\omega^{NA}\) is concave with respect to \(\beta\) (with at most two roots which we will refer to as \(\beta_{\min}(x)\) and \(\beta_{\max}(x)\)) and is maximized at \(\frac{N_\alpha(x)}{N(x)}\).

Lemma 8. \(\omega^{BJ}(\alpha, \beta, N_\alpha(x), N(x))\) is concave with respect to \(\beta\), maximized at \(\beta_{\text{mle}}(x) = \frac{N_\alpha(x)}{N(x)}\), and has two roots.

Proof.

\[
\frac{\partial \omega^{BJ}(\alpha, \beta, N_\alpha(x), N(x))}{\partial \beta} = \frac{N_\alpha(x) - N(x)\beta}{\beta(1 - \beta)} \\
\text{(set) } 0 = \frac{N_\alpha(x) - N(x)\beta}{\beta(1 - \beta)} \\
0 = N_\alpha(x) - N(x)\beta \\
\beta = \frac{N_\alpha(x)}{N(x)}
\]

shows that \(\omega^{BJ}\) is maximized (if it is concave) at \(\frac{N_\alpha(x)}{N(x)}\) and has at most two roots (which we will refer to as \(\beta_{\min}(x)\) and \(\beta_{\max}(x)\)). Additionally,

\[
\left.\frac{\partial^2 \omega^{BJ}(\alpha, \beta, N_\alpha(x), N(x))}{\partial^2 \beta}\right|_{\beta = \frac{N_\alpha(x)}{N(x)}} = -\frac{\beta^2 N_\alpha(x) + (1 - 2\beta)N(x)}{(\beta - 1)^2 \beta^2} \bigg|_{\beta = \frac{N_\alpha(x)}{N(x)}} < 0
\]

shows that \(\omega^{BJ}\) is concave with respect to \(\beta\).

Now that we have demonstrated that \(\omega\) is concave, we now demonstrate a key insight about the difference between \(\alpha\) and \(\beta_{\max}(x)\) (i.e., \(r_{\max}\)) relative to the difference between \(\alpha\) and \(\beta_{\text{mle}}(x)\) (i.e., \(r_{\text{mle}}\)).

Lemma 9. With respect to \(\omega^{NA}(\alpha, \beta, N_\alpha(x), N(x))\), \(r_{\max}(x) / r_{\text{mle}(x)} = 2\).

Proof. First, by Lemma 7 we know that, with respect to \(\beta\), \(\omega^{NA}\) is concave and has at most 2 roots
\( \omega^{NA}(\alpha, \beta, N\alpha(x), N(x)) = \frac{N\alpha(x) (\beta - \alpha)}{\alpha(1 - \alpha)} + \frac{N(x)(\alpha^2 - \beta^2)}{2\alpha(1 - \alpha)} \)

(set) \( 0 = \frac{N\alpha(x)(\beta - \alpha)}{\alpha(1 - \alpha)} + \frac{N(x)(\alpha^2 - \beta^2)}{2\alpha(1 - \alpha)} \)

\( = 2N\alpha(x)(\beta - \alpha) + N(x)(\alpha^2 - \beta^2) \)

\( = (-N(x)) \beta^2 + (2N\alpha(x)) \beta + (-2\alpha N\alpha(x) + N(x)\alpha^2) \)

\( \{\beta_{\min}(x), \beta_{\max}(x)\} = \frac{-2N\alpha(x) \pm \sqrt{(-2N\alpha(x)) - 4(-N(x))(-2N\alpha(x) + (N(x)\alpha)^2)}}{-2N(x)} \)

\( = \frac{N\alpha(x) \pm \sqrt{N\alpha(x)^2 - 2N\alpha(x)N(x)\alpha + (N(x)\alpha)^2}}{N(x)} \)

\( = \frac{N\alpha(x) \pm \sqrt{(N\alpha(x) - 2N(S)\alpha)^2}}{N(x)} \)

\( = \frac{N\alpha(x) \pm (N\alpha(x) - 2N(x)\alpha)}{N(x)} \)

\( = \{\alpha, 2 \beta_{\text{mle}}(x) - \alpha\} \).

This implies that \( \beta_{\max}(x) - \alpha = 2(\beta_{\text{mle}}(x) - \alpha) \) and \( r_{\max}(x) = 2r_{\text{mle}}(x) \), with respect to \( \omega^{NA} \).

**Lemma 10.** With respect to \( \omega^{BJ}(\alpha, \beta, N\alpha(x), N(x)) \),

\[
\frac{r_{\max}(x)}{r_{\text{mle}}(x)} \begin{cases} < 2 & \text{if } \beta_{\text{mle}}(x) > \frac{1}{2} \\ = 2 & \text{if } \beta_{\text{mle}}(x) = \frac{1}{2} \\ > 2 & \text{otherwise}. \end{cases}
\]

*Proof.* First, by Lemma 8, we know that, with respect to \( \beta \), \( \omega^{BJ} \) is concave and has at most 2 roots \((\beta_{\min}(x), \beta_{\max}(x))\). One of the solutions of \( \omega^{BJ} \) must be \( \alpha \), so let us assume that \( \beta_{\min}(x) = \alpha \); this will be true when \( \beta > \alpha \), which intuitively corresponds to our case of interest: when the covariate profile contains more significant (extreme) \( p \)-values than expected. Furthermore, we know that \( \omega^{BJ} \) achieves a maximum at \( \beta_{\text{mle}} = \frac{N\alpha(x)}{N(x)} \). With these properties we can show the first case \((1 \leq \frac{r_{\max}(x)}{r_{\text{mle}}(x)} < 2)\) by first recognizing that trivially \( \beta_{\text{mle}} \leq \beta_{\max} \), and \( \beta_{\text{mle}} - \alpha \leq \beta_{\max} - \alpha \). To show the upper bound of the first case, it suffices to show that \( \omega^{BJ}(\alpha, \beta_{\text{mle}} - \epsilon, N\alpha(x), N(x)) \geq \omega^{BJ}(\alpha, \beta_{\text{mle}} + \epsilon, N\alpha(x), N(x)) \) for some \( \epsilon > 0 \). The essential implication is that the concave function \( \omega^{BJ} \) increases at a slower rate (until it reaches its maximum) than it decreases. This further implies that the distance between \( \beta_{\text{mle}} \) and \( \alpha \) is higher than \( \beta_{\text{mle}} \) and \( \beta_{\max} \), and therefore the desired result.

Recall from Lemma 8 that

\[
\frac{\partial \omega^{BJ}(\alpha, \beta, N\alpha(x), N(x))}{\partial \beta} = \frac{N\alpha(x) - N(x)\beta}{\beta(1 - \beta)}
\]

\[
= N(x) \left[ \frac{\beta_{\text{mle}}(x) - \beta}{\beta(1 - \beta)} \right],
\]

which means the slope of \( \omega^{BJ} \) is proportional to \( \frac{\beta_{\text{mle}}(x) - \beta}{\beta(1 - \beta)} \). We now compare the slope around the inflection point \( \beta_{\text{mle}}(x) \), and recognize that at \( \beta = \beta_{\text{mle}}(x) + \epsilon \) the slope is negative with absolute value proportional
to \(\frac{\beta_{\text{mle}}(x) + \epsilon}{\beta_{\text{mle}}(x) - \epsilon}\). At \(\beta = \beta_{\text{mle}}(x) - \epsilon\) the slope is positive with absolute value proportional to \(\frac{1}{\epsilon}\). Therefore,

\[
\beta_{\text{mle}}(x) > \frac{1}{2} \iff (\beta_{\text{mle}}(x) + \epsilon)(1 - \beta_{\text{mle}}(x) - \epsilon) < (\beta_{\text{mle}}(x) - \epsilon)(1 - \beta_{\text{mle}}(x) + \epsilon)
\]

\[
\iff (\beta_{\text{mle}}(x) + \epsilon)(1 - \beta_{\text{mle}}(x) - \epsilon) > (\beta_{\text{mle}}(x) - \epsilon)(1 - \beta_{\text{mle}}(x) + \epsilon)
\]

\[
\iff \frac{r_{\text{max}}(x)}{r_{\text{mle}}(x)} < 2.
\]

The demonstration of the remaining two conditions follow precisely the same approach above, mutatis mutandis.

Now that we have built up the necessary properties of the \(\omega\) functions, we now will discuss the sufficient conditions for the detected subset to be exactly correct \(S^* = S^T\). To begin we re-introduce some additional notation:

\[
r_{\text{mle}-h} = \max_{x \in U_X(S^T)} r_{\text{mle}}(x),
\]

\[
r_{\text{mle}-l} = \min_{x \in U_X(S^T)} r_{\text{mle}}(x),
\]

\[
r_{\text{unaff}} = \max_{x \notin U_X(S^T)} r_{\text{mle}}(x),
\]

\[
\eta = \left(\frac{\sum_{x \in U_X(S^T)} N(x)}{\sum_{x \in U_X(S)} N(x)}\right),
\]

\[
\nu - \text{homogeneous}: r_{\text{mle}-h} < \nu,
\]

\[
\text{\(\delta\)-strong}: \frac{r_{\text{mle}-l}}{r_{\text{unaff}}} > \delta,
\]

\[
R: (0, 1) \mapsto (0, 1).
\]

More specifically, \(R\) is an invertible function such that \(R: r_{\text{max}}(x) \mapsto r_{\text{mle}}(x)\) i.e., if \(R\) is applied to \(r_{\text{max}}(x)\) it would produce the corresponding \(r_{\text{mle}}(x)\). From Lemma 9 we know that with respect to \(\omega^{NA}\), \(R^{NA}(r) = \frac{r}{2}\), while from Lemma 10 we know that with respect to \(\omega^{BJ}\), \(R^{BJ}(r) \leq \frac{r}{2}\) under certain conditions.

The first result we provide is a sufficient condition for guaranteeing that the detected subset includes all the elements from the true subset \((S^* \supseteq S^T)\). More specifically, we show that such a condition is sufficient homogeneity of the affected data elements: for a given value \(\nu\), and any pair of affected covariate profiles \((x_i, x_j) \in U_X(S^T))\), the anomalous signal (i.e., treatment effect) observed in \(x_i\) is no more than \(\nu\)-times that which is observed in \(x_j\).

**Theorem 4.** Under \(H_1(S^T)\) defined in [8], where \(|U_X(S^T)| = t\), \(\exists \nu > 1\) such that if the observed effect (as measured by \(\omega\)) across the \(t\) covariate profiles in \(S^T\) is \(\nu\)-homogeneous, and at least \(1\)-strong, then the highest scoring subset \(S^* \supseteq S^T\).

**Proof.** First, let \(\{x_{(1)}, \ldots, x_{(i)}\}\) be the data elements in \(S^T\) sorted by the priority function (Theorem 1) \(G(x) = \frac{N(x)}{N(x)} = \beta_{\text{mle}}(x)\). By the assumption of an observed signal that is at least \(1\)-strong, these data
elements are the $t$ highest priority data elements. Additionally, let $\nu = \frac{\frac{\alpha_{\text{mle}}}{\beta_{\text{mle}}}}{R(\frac{\alpha_{\text{mle}}}{\beta})}$. Therefore,

$$\nu - \text{homogeneous} \implies \nu > \frac{\alpha_{\text{mle}}}{\beta_{\text{mle}}}, \frac{\beta_{\text{mle}}}{\alpha_{\text{mle}}}$$

$$\therefore \quad \frac{\alpha_{\text{mle}}}{\beta_{\text{mle}}} > \frac{\beta_{\text{mle}}}{\alpha_{\text{mle}}}$$

$$\implies \quad R(\frac{\alpha_{\text{mle}}}{\beta_{\text{mle}}}) > R(\frac{\beta_{\text{mle}}}{\alpha_{\text{mle}}})$$

$$\implies \quad R^{-1}(\frac{\alpha_{\text{mle}}}{\beta_{\text{mle}}}) > \frac{\beta_{\text{mle}}}{\alpha_{\text{mle}}}$$

$$\implies \quad \beta_{\text{max}}(x(t)) - \alpha > \beta_{\text{mle}}(x(1)) - \alpha$$

$$\implies \quad \beta_{\text{max}}(x(t)) > \beta_{\text{mle}}(x(k)) \quad (\forall k)$$

$$\implies \quad \beta_{\text{max}}(x(t)) > \beta_{\text{mle}}(S^*)$$

$$\therefore \quad \omega(\alpha, \beta_{\text{mle}}(S^*), N\alpha(x(t)), N(x(t))) > 0$$

$$\therefore \quad |S^*| \geq t$$

$$\therefore \quad S^* \supseteq S^T.$$  

Intuitively, $\beta_{\text{mle}}(x(t))$ and $\beta_{\text{mle}}(x(1))$ are respectively the smallest and largest $\beta_{\text{mle}}$ of all the $x \in U_X(S^T)$. Furthermore, $\beta_{\text{mle}}(x(t)) \leq \beta_{\text{mle}}(x(k)) \leq \beta_{\text{max}}(x(t)) \forall k \in [1, t]$, which means $\beta_{\text{mle}}(S^*) \leq \beta_{\text{max}}(x(t))$ for the optimal subset $S^*$. Moreover, the $S^*$ that maximizes $F_{\alpha, \beta}$ will include any covariate profile $x$ that would make a positive contribution to the score $F_{\alpha, \beta}$ at the given value of $\beta$. Such a positive contribution occurs when the concave $\omega$ function of $x$ is positive. At the optimal $\alpha$ and $\beta = \beta_{\text{mle}}(S^*)$ the $\omega$ function for each of the $\{x(1), \ldots, x(t)\}$ is positive because $\beta_{\text{max}}$ (the larger root of the $\omega$ functions) for each of these elements is greater than $\beta_{\text{mle}}(S^*)$.

**Corollary 3.** From Lemma 9 we know that with respect to $\omega^{NA}$, $\frac{\alpha}{\beta} = 2$. Additionally, from Lemma 10 we know that with respect to $\omega^{BJ}$, $\frac{\alpha}{\beta} \leq 2$ under certain conditions. Therefore, we can conclude that at $\alpha^*$, 2-homogeneity (and 1-strength) is sufficient for $S^* \supseteq S^T$ with respect to $F^{NA}$; to $F^{BJ}$, under some conditions; and to the other score functions described above, by Lemma 3. Essentially, if the observed proportions of $p$-values significant at $\alpha^*$ vary by no more than a factor of 2 across all of the $x \in U_X(S^T)$, then the detected subset will include all of the affected data elements.

The next result we provide is a sufficient condition for guaranteeing that the detected subset will only include elements from the true subset ($S^* \subseteq S^T$). More specifically, we show that such a condition is sufficient strength of the affected data elements; or intuitively, for a given value $\delta$, the anomalous signals observed in every affected data elements is more than $\delta$-times that of the unaffected data elements.

**Theorem 5.** Under $H_1(S^T)$ defined in $S$, where $|U_X(S^T)| = t$, $\exists \delta > 1$ such that if the observed effect (as measured by $\omega$) across the $t$ covariate profiles in $S^T$ is $\frac{\delta}{\eta}$-strong, then the highest scoring subset $S^* \subseteq S^T$.

**Proof.** First, let $D = \{x(1), \ldots, x(t), x(t+1), \ldots, x(M)\}$ be the data elements sorted by the priority function (Theorem 1) $G(x) = \frac{N\alpha(x)}{N(x)} = \beta_{\text{mle}}(x)$. By the assumption of $\delta > 1$ (an observed signal that is at least
$1$-strong, $S^T = \{x_1, \ldots, x_t\}$. Additionally, let $\delta = \frac{R^{-1}(r_{\text{unaff}}_{\text{mle}})}{r_{\text{mle}} - h}$. Therefore,

$$\frac{\delta - \eta}{\eta} \implies R^{-1}(r_{\text{unaff}}_{\text{mle}}) < \frac{\delta_{\text{mle}} - l}{r_{\text{mle}} - h}$$

$$\therefore R^{-1}(r_{\text{unaff}}_{\text{mle}}) < \frac{\delta_{\text{mle}} - l}{r_{\text{mle}} - h}$$

$$\implies R^{-1}(r_{\text{mle}} - h) < \frac{\sum_{x \in U_X(S^T)} N(x)}{\sum_{x \in U_X(D)} N(x)} r_{\text{mle}}$$

$$= \frac{\sum_{x \in U_X(S^T)} N(x)}{\sum_{x \in U_X(D)} N(x)} \left( \text{since } r_{\text{mle}} \geq r_{\text{mle}} - l \quad \forall x \in U_X(S^T) \right)$$

$$\leq \frac{\sum_{x \in U_X(S^T)} N(x)}{\sum_{x \in U_X(D)} N(x)} + \frac{\sum_{x \in U_X(S^T)} N(x)}{\sum_{x \in U_X(D)} N(x)}$$

$$= \frac{\sum_{x \in U_X(D)} N(x)}{\sum_{x \in U_X(D)} N(x)} \left( \frac{N(x)}{N(x)} - \alpha \right) N(x)$$

$$= \frac{\sum_{x \in U_X(D)} N(x)}{\sum_{x \in U_X(D)} N(x)} N(x) - N(x) \alpha$$

$$= \frac{\sum_{x \in U_X(D)} N(x)}{\sum_{x \in U_X(D)} N(x)} N(x)$$

$$= \frac{\sum_{x \in U_X(D)} N(x)}{\sum_{x \in U_X(D)} N(x)} N(x) - \alpha$$

$$\therefore \beta_{\text{mle}}(x_{t+1}) - \alpha < \beta_{\text{mle}}(D) - \alpha$$

$$\implies \beta_{\text{mle}}(x_{t+1}) < \beta_{\text{mle}}(x_t)$$

$$\implies \beta_{\text{max}}(x_{t+1}) < \beta_{\text{mle}}(S^*)$$

$$\therefore \omega(\alpha, \beta_{\text{mle}}(S^*), N_\alpha(x_{t+1}), N(x_{t+1})) < 0$$

$$\implies |S^*| \leq t$$

$$\therefore S^* \subseteq S^T$$

Intuitively, $\beta_{\text{mle}}(x_t)$ and $\beta_{\text{mle}}(x_{t+1})$ are respectively the smallest affected and largest unaffected $\beta_{\text{mle}}$ values. Furthermore, $\beta_{\text{max}}(x_{t+1}) \leq \beta_{\text{mle}}(x_t)$, which means $\beta_{\text{max}}(x_{t+1}) \leq \beta_{\text{mle}}(S^*)$ for the optimal subset $S^*$. Moreover, the $S^*$ that maximizes $F_{\alpha, \beta}$ will not include any data element $x$ that makes a non-positive contribution to the score $F_{\alpha, \beta}$ at the given value of $\beta$. Such a non-positive contribution occurs when the concave $\omega$ function of $s$ is non-positive. At the optimal $\alpha$ and $\beta = \beta_{\text{mle}}(S^*)$ the $\omega$ function for each of the $\{x_{t+1}, \ldots, x_M\}$ are non-positive because $\beta_{\text{max}}$ (the larger root of the $\omega$ functions) for each of these elements is less than $\beta_{\text{mle}}(S^*)$.

**Corollary 4.** From Lemma 10 we know that with respect to $\omega^{\text{NA}}$, $R^{-1}(r) = 2$. Additionally, from Lemma 10 we know that with respect to $\omega^{\text{BJ}}$, $R^{-1}(r) \geq 2$ under certain conditions. Therefore, we can conclude that at $\alpha^*$, $\frac{2}{\eta}$-strength is sufficient for $S^* \subseteq S^T$ with respect to $F^{\text{NA}}$; to $F^{\text{BJ}}$, under some conditions; and to the other score functions described above, by Lemma 3. Essentially, if the observed proportions of $p$-values
significant at $\alpha^*$ across all of the $x \in U_X(S^T)$ are at least $\frac{2}{\eta}$ times larger than the observed proportions for $x \notin U_X(S^T)$, then the detected subset will only include affected data elements.

**Theorem 6.** Under $H_1(S^T)$, where $S^T$ is a $t$-element subset of covariate profiles and for each element the true potential outcome distributions are unequal, $\exists \nu, \delta > 1$ such that if the observed effect (as measured by $\omega$) across these covariate profiles is $\nu$-homogeneous and $\frac{\delta}{\eta}$-strong, then $S^* = S^T$.

**Proof.**

\[ \because \ \nu - \text{homogeneous} \implies S^* \supseteq S^T \quad (\text{by Theorem 4}) \]
\[ \because \ \frac{\delta}{\eta} - \text{strong} \implies S^* \subseteq S^T \quad (\text{by Theorem 5}) \]
\[ \therefore \quad S^* = S^T \]

It follows from the above corollaries that 2-homogeneity and $\frac{2}{\eta}$-strength are sufficient for $S^* = S^T$ with respect to $F^{NA}$; to $F^{BJ}$, under some conditions; and to the other score functions described above, by Lemma 2.