Causal Inference With Bundled Variables*

Zachary Markovich†

February 22, 2023

Abstract

Bundled variables are ubiquitous in scientific research. Traits as wide ranging as DNA, complex medical interventions, and personality can all be represented as high dimensional vectors. However, the conventional tools of causal inference have been developed with univariate treatments or low dimensional conditioning sets in mind. This article proposes a framework for conducting causal inference with such bundled variables. Specifically, I propose focusing on the maximum difference in causal effects associated with two sets of bundles. This estimand is intuitive and will answer questions like "What is the difference in average potential outcomes/ treatment effects between units that received one of the most rather than one of the least effective treatment/moderator bundles?" I term this the *Maximum Causal Set Effect* (MCSE) and develop a bounding approach to estimation. The upper bound reduces the bias of the maximum of a set of random variables while the lower bound uses split sample methods to conservatively estimate the MCSE. The lower bound is also asymptotically normal, facilitating classical inference about this novel estimand.

*The author thanks Adam Berinsky, Drew Dimmery, Justin Grimmer, John Marshall, Jessica Sun, Brandon Stewart, Chloe Wittenberg, Teppei Yamamoto, and members of the Kim Research Group for the helpful comments.

†Ph.D. Student, Department of Political Science, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139. Email: zmarko@mit.edu

1 Introduction

Consider the plight of an analyst attempting to measure the magnitude of treatment effect heterogeneity present in a fully randomized experiment. Recently developed non-parametric estimation techniques can provide an estimate for the conditional average treatment effect for every combination of conditioning covariates [Green and Kern, 2012, Athey and Imbens, 2016, Grimmer et al., 2017, Wager and Athey, 2018, Künzel et al., 2019]. However, scientific theories rarely make predictions that can be easily tested using such a large array of estimates [Markovich, 2021]. Instead, they typically make precise predictions about the role of one or two conditioning variables, which can be tested using conventional regression methods, or they make a general prediction about the overall magnitude of causal effect heterogeneity, which is not easily assessed with such a large number of estimates.

A similar problem emerges in the many causes setting [Imbens, 2000, Wang and Blei, 2019, Li et al., 2019, Wang et al., 2018, Zheng et al., Forthcoming]. For example, Wang and Blei [2019] create a toolkit for estimating the causal effect on a film's box office associated with every individual actor present in the Internet Movie Databse (IMDB), but an economist trying to understand the general determinants of box office performance is more likely to want to understand the magnitude of the effect that actors have on a film's financial success in general rather than thousands of imprecisely estimated treatment effects associated with every individual actor.

Such scenarios represent examples of causal inference involving *bundled variables* because they require modeling the combined effect of many traits. In contrast, the standard tools of causal inference are stated in terms of a small number of variables. Scientific research often requires analyzing such bundled variables. Indeed, modeling the influence of variables as disparate as an organism's genome [Stephens and Balding, 2009, Visscher et al., 2017], an individual's personality [Pervin, 2003], or the contents of a body of text [Fong and Grimmer, 2016, Egami et al., 2018, Fong and Grimmer, forthcoming] can all be understood as examples of bundled variables. Bundled variables principally pose an interpretability problem because

the researcher will not know which of the many distinct causal effects they should use to parameterize their theory.

In this article, I propose a framework for quantifying the causal effect of such bundled variables. Specifically, I build on the informal heuristic approach of reporting just the 5 or 10 variables associated with the largest or smallest causal effects. For example, a researcher arguing in favor of the large causal effect of actors on a movie's box office performance might point to the difference between a small number of actors that appear to make a large contribution to a movie's financial performance and a different subset of actors which actually harmed a movie's box office. Similarly, a researcher arguing there is a great deal of treatment effect heterogeneity might highlight a handful of covariate strata associated with very large or very small treatment effects.

I formalize this heuristic approach to create a more general strategy for summarizing the distribution of many causal effects. Specifically, I assume that the researcher has identified a set of *treatment types* denoted by $\mathbf{t}_i$ for $i = 1 \dots N$ with support $\mathbb{T}$ and has specified some causal estimand that is defined in terms of different sets of treatment types, $\tau(\mathcal{T}', \mathcal{T}'')$, for any $\mathcal{T}', \mathcal{T}'' \subseteq \mathbb{T}$. In the many causes setting, $\mathbf{t}_i$ will typically be a vector where each element indicates whether respondent i received the corresponding treatment and $\tau(\mathcal{T}', \mathcal{T}'')$ will be defined as the difference in expected potential outcomes associated with being treated with an element of $\mathcal{T}'$ rather than $\mathcal{T}''$. In the movies example, each element j of vector $\mathbf{t}_i$ might be a dummy variable where a 1 indicates that actor j appeared in film i while a zero indicates that they did not. $\tau(\mathcal{T}', \mathcal{T}'')$ would then be defined as the difference in counterfactual box office performances between movies that featured a combination of actors in $\mathcal{T}'$ and movies that featured a combination of actors in $\mathcal{T}''$. In the treatment effect heterogeneity setting on the other hand, $\mathbf{t}_i$ will denote different combinations of the treatment and conditioning covariates and $\tau(\mathcal{T}', \mathcal{T}'')$ will denote the difference in conditional treatment effects given that those conditioning covariates are in $\mathcal{T}'$ rather than $\mathcal{T}''$. I propose then reporting the maximum value of $\tau(\mathcal{T}', \mathcal{T}'')$ subject to the constraint that $\mathcal{T}'$ and $\mathcal{T}''$ have at least a researcher specified

probability q of occurring, an estimand I term the *Maximum Causal Set Effect* (MCSE).

I provide a bounding approach to estimation. By Jensen's inequality, the maximum of a set of random variables will be greater than the maximum of the expectations of those random variables. The upper bounding estimator builds on this strategy by using monte carlo methods to reduce this bias. The lower bounding estimator instead uses a split sample approach to generate a donwardly biased estimate of the MCSE. I also show that the lower bound is asymptotically normal and amenable to the non-parametric bootstrap, facilitating a valid test of the null hypothesis that the MCSE is equal to zero.

After briefly reviewing the relevant literature to this problem, the article proceeds as follows. The next section outlines two motivating examples of bundled treatments. Section 3 introduces some basic notation and formally defines the MCSE. Section 4 discusses the properties of the upper and lower bounding estimatators for the MCSE. Section 5 presents simulation results, while Section 6 revisits my motivating examples.

Previous Literature Although a large literature considers strategies for reducing confounding when there are multiple treatments[e.g. Imbens, 2000, McCaffrey et al., 2013, Rassen et al., 2013, Yang et al., 2016, Lopez et al., 2017, Li et al., 2019, Wang et al., 2018] and identifying treatment effect heterogeneity when there is a large set of conditioning covariates [e.g. Athey and Imbens, 2016, Chernozhukov et al., 2018, Wager and Athey, 2018]; however, the problem of interpretability that this paper focuses on has been largely neglected. The only existent proposal for addressing this challenge is to dimension reduce the relevant causal variables and then focus on a simple causal query defined in terms of that latent trait.[Fong and Grimmer, 2016, Nabi et al., 2017, Zou et al., 2020, Fong and Grimmer, forthcoming] In some cases, this dimension reduction is theoretically motivated and can be understood as targeting a different causal estimand than what is considered in this paper. For example Fong and Grimmer [2016] focus on identifying the effect of topics that appear in a body of text. In this case, the dimension reduction allows the analyst to estimate the

causal effect of a specific latent trait. However, the dimension reduction is most frequently invoked as a way of simplifying the causal analysis without a theoretical motivation for focusing a low dimensional latent trait. In these cases, such dimension reduction techniques will discard much of the variation in the full bundle and risk understating the magnitude of causal effects.

2 Motivating Examples

This section outlines two motivating examples of bundled variables. The first focuses on the role of race in statistical models while the second examines the influence of campaign expenditures on election outcomes. Both highlight examples where dimension reduction techniques are inappropriate for quantifying the causal effects of the bundled treatment, but the MCSE may be helpful.

2.1 Race

Modern theories of race emphasize its multi-faceted nature. Many individuals' racial identities are fluid and vary with "time, place, and social context" [Saperstein, 2006]. In particular, conceptions of race may depend on age, immigration status, and educational attainment [Saperstein and Penner, 2012]. For example, Freeman et al. [2011] find that changing the clothing shown in pictures resulted in changes to the perceived racial identity of the individuals displayed. Such theories have led Sen and Wasow [2016] to argue that "race is rarely if ever a single, uniform entity... Racial categories are the product of a complex fusion of factors including societal values, skin color, cultural traits, physical attributes, diet, region of ancestry, institutional power relationships, and education. In other words, race is an aggregate of many components; metaphorically, it is a bundle."

In spite of the predominance of this constructivist view within the sociological and anthropological literature, race most frequently enters statistical models as a binary trait. This

approach is inherently reductive and risks understating the magnitude of effects because it ignores much of the variation in the treatment bundles that compose race. Quintana [2021] makes some progress on this challenge by using a machine learning approach to identify a set of variables that form the "bundle of sticks" that composes race in the context of early childhood education. Specifically, they identify a set of 18 variables that are sufficient to block the relationship between race (measured as a binary trait) and scores on standardized tests in the Early Childhood Longitudinal Study (ECLS). Quintana [2021] refers to this set as the "markov blanket" of race. While Quintana [2021] presents results showing that predictions about test results made using race's markov blanket have a higher out of sample accuracy than those using just the binary label, he lacks an estimand to quantify the combined causal effect of the markov blanket.

2.2 Campaign Expenditures

The ability of elected officials to manipulate the views of their constituents is central to the functioning of democratic governance. Political campaigns represent the most overt form of this influence; however, the empirical evidence supporting their impact on public opinion is quite weak. Observational studies show a high degree of model dependence [Levitt, 1994, Gerber, 1998, Schuster, 2020] while recent experimental results suggest the effects are typically quite small [Kalla and Broockman, 2018, Coppock et al., 2020]. Kalla and Broockman [2018] summarize this body of evidence in a recent meta-analysis writing, "the direct persuasive effects of [campaigns'] voter contact and advertising in general elections are essentially zero."

Most attempts to study the effect of campaigns have looked at just a single dimension of campaigning. For example, Sides et al. focus on just television advertising, Kalla and Broockman [2018] consider only direct contact from campaigns, and Schuster [2020] examines just aggregate campaign expenditures. In reality though, campaigns are multi-faceted and engage in many activities meant to influence the outcome of an election simultaneously.

In this sense, political campaigns represent bundled variables because they each engage in a bundle of different activities in the run up to an election. However, previous statistical approaches have not been effective in quantifying the causal effect of this complete bundle.

3 Notation and Causal Estimands

3.1 Notation

I assume that the researcher observes a set of N independent $(\mathbf{t}_i, Y_i, \mathbf{x}_i)$ triplets where Y_i is the outcome, $\mathbf{t}_i$ is a length J vector indicating the *treatment type* received by unit i, and $\mathbf{x}_i$ is a length J vector representing a set of background covariates that causal effects should be adjusted for. Additionally, let $\mathbb{T}$ denote the support of the distribution of $\mathbf{t}_i$.

I also assume that the researcher has specified some causal contrast, $\tau(\mathcal{T}', \mathcal{T}'')$, that is defined in terms of two sets, $\mathcal{T}', \mathcal{T}'' \subseteq \mathbb{T}$. There is a nearly unlimited number of possible interesting choices for $\tau(\mathcal{T}', \mathcal{T}'')$. In this paper, I will focus on just two primary examples: the causal effect of *many causes* and treatment effect heterogeneity when there are *many moderators*. In the many causes case, I denote the set of potential outcomes for unit i as $\{Y_i(\mathbf{t}) : \mathbf{t} \in \mathbb{T}\}$, then:

$$\tau(\mathcal{T}', \mathcal{T}'') \equiv \mathbb{E}_{\mathbf{t}_i}\left(\mathbb{E}_{\mathrm{Y_i}}\left(Y_i(\mathbf{t}_i)|\mathbf{t}_i \in \mathcal{T}'\right) - \mathbb{E}_{\mathrm{Y_i}}\left(Y_i(\mathbf{t}_i)|\mathbf{t}_i \in \mathcal{T}''\right)\right)$$

where subscripting on the expectation operator indicates the variable that the expectation is being taken over. In this case, $\tau(\mathcal{T}', \mathcal{T}'')$ represents the average effect of receiving a set of treatments contained in $\mathcal{T}'$ rather than $\mathcal{T}''$. In the *many moderators* case on the other hand let the first element of $\mathbf{t}_i$ be arbitrarily denoted $T_i \in \{0, 1\}$, which represents the value of some binary treatment assigned to unit i. Analogously, let the remaining elements of $\mathbf{t}_i$ represent the set of conditioning covariates recieved by unit i and be denoted x_i. Then define the set of potential outcomes for each unit i as $\{Y_i(1), Y_i(0)\}$ and let:

$$\tau(\mathcal{T}', \mathcal{T}'') \equiv \mathbb{E}_{\mathbf{t}_i}\left(\mathbb{E}_{\mathrm{Y_i}}\left(Y_i(1) - Y_i(0) | \mathbf{t}_i \in \mathcal{T}'\right)\right) - \mathbb{E}_{\mathbf{t}_i}\left(\mathbb{E}_{\mathrm{Y_i}}\left(Y_i(1) - Y_i(0) | \mathbf{t}_i \in \mathcal{T}''\right)\right)$$

where subscripting on the expectation operator again indicates the variable that the expectation is being taken over. While I limit my simulations and applications to these two examples, $\tau(\mathcal{T}', \mathcal{T}'')$ could easily be defined in terms of different quantiles instead of expectations, ratios of different causal effects, etc.

3.2 Defining the Maximum Causal Set Effect

The main challenge for defining causal effects under this framework is that a different value of $\tau(\mathcal{T}', \mathcal{T}'')$ can be defined for every distinct pair of sets $\mathcal{T}', \mathcal{T}'' \subseteq \mathbb{T}$, leaving the analyst without a single unambiguous quantity of interest to report to summarize their findings. In this section, I define a causal estimand which overcomes this challenge by focusing on the contrast between two sets $\mathcal{T}_q^{\text{Max}}$ and $\mathcal{T}_q^{\text{Min}}$ which maximize $\tau(\mathcal{T}', \mathcal{T}'')$. To avoid choosing sets $\mathcal{T}_q^{\text{Max}}$ and $\mathcal{T}_q^{\text{Min}}$ which correspond to unrepresentative edge cases, I require that the sets have at least a researcher specified probability of occurring: q. Formally, let the set of subsets of $\mathbb{T}$ such that the probability that $\mathbf{t}_i$ is in $\mathbb{T}$ is at least $q \in (0, 1)$ be defined as: $\mathcal{T}_q \equiv \{\mathcal{T}' \subseteq \mathbb{T} : P(\mathbf{t}_i \in \mathcal{T}') \geq q\}$ where,

$$P(\mathbf{t}_i \in \mathcal{T}') = \int_{\mathbb{T}} g(\mathbf{t}) \mathbb{1}\{\mathbf{t} \in \mathbb{T}\} d\mathbf{t}$$

I then define the *Maximum Causal Set Effect* as:

$$\text{MCSE}_q = \max_{\mathcal{T}', \mathcal{T}'' \in \mathcal{T}_q} \tau(\mathcal{T}', \mathcal{T}'') = \tau\left(\mathcal{T}_q^{\text{Max}}, \mathcal{T}_q^{\text{Min}}\right)$$

I refer to $\mathcal{T}_q^{\text{Max}}$ as the *maximum causal set* and $\mathcal{T}_q^{\text{Min}}$ as the *minimum causal set.*[1]

[1]To be explicit, $\mathcal{T}_q^{\text{Max}}, \mathcal{T}_q^{\text{Min}} = \arg\max_{\mathcal{T}', \mathcal{T}'' \in \mathcal{T}_q} \tau(\mathcal{T}', \mathcal{T}'')$

4 Estimation

This section outlines the two estimators which compose the bounds. The first subsection presents the upper bound while the second subsection develops the lower bound.

4.1 Upper Bound

My estimation approach is oriented towards a setting where the researcher has identified an estimator for $\tau(\mathcal{T}', \mathcal{T}'')$ such that:

Assumption 1. $\forall, \mathcal{T}', \mathcal{T}'' \in \mathcal{T}_q$,

$$\mathbb{E}\left(\hat{\tau}(\mathcal{T}', \mathcal{T}'')\right) = \tau(\mathcal{T}', \mathcal{T}'')$$

This assumption will not typically be verifiable by the analyst and instead will rest on additional untestable assumptions. However, such assumptions are routinely made by researchers conducting causal inference and are generally not considered overly onerous. For example, in the context of many experimentally randomized causes, SUTVA [Rubin, 1974], random assignment to treatment, and positivity will be sufficient to unbiasedly estimate $\tau(\mathcal{T}', \mathcal{T}'')$ using the difference in means observed between units treated with an element of $\mathcal{T}'$ and an element $\mathcal{T}''$. In an observational setting on the other hand, regression, matching, or weighting techniques could be used to estimate the same quantity of interest while adjusting for covariates and weakening the assumption of random assignment of treatment to one of conditional ignorability.

Additionally, I assume that the researcher has specified the distribution of $\mathbf{t}_i$ in the target population:

Assumption 2. *The probability density function $g(\mathbf{t})$ in the target population is known.*

The choice of $g(\mathbf{t})$ should be interpreted as defining the target estimand. If the empirical distribution of $\mathbf{t}_i$ is used, the corresponding MCSE_q can be interpreted as a sample estimand.

On the other hand, if the analyst has knowledge about the frequency of different values of $\mathbf{t}_i$ in the population, they can instead use that distribution to define $g(\mathbf{t})$. For example, a medical researcher might rely on data from clinical trials to measure the efficacy of different bundles of medical treatments, but might wish to define the MCSE in terms of frequency of those treatments observed in medical records data.

With these two assumptions in hand, a straightforward approach to estimating the MCSE_q would be to calculate the MCSE as as the maximum of all the estimates $\hat{\tau}(\mathcal{T}', \mathcal{T}'')$:

$$\widehat{\text{MCSE}}_q^{\text{Max}} = \max_{\mathcal{T}', \mathcal{T}'' \in \mathcal{T}_q} \hat{\tau}(\mathcal{T}', \mathcal{T}'') \tag{1}$$

Since the max is a convex function, a single application of Jensen's inequality shows that this estimator will suffer from an upwards bias.

In the remainder of this subsection, I develop a monte carlo approach for reducing this bias. Specifically. I begin with the additional assumption that the estimators, $\hat{\tau}(\mathcal{T}', \mathcal{T}'')$, are normally distributed with an unbiased estimator for their variance-covariance matrix:

Assumption 3. *$\hat{\tau}(\mathcal{T}', \mathcal{T}'')$ are jointly normally distributed and there is an unbiased estimator for their variance-covariance matrix.*

Assumption 3 embeds two conditions, the first of which is quite strong. It will hold asymptotically for many estimators, but it is unlikely to be satisfied in finite samples. In Section 5, I present some simulation results suggesting that this estimator is fairly robust in the face of mild deviations from normality and performs well in finite samples. The second condition on the other hand is more mild, and simply requires there be an unbiased estimator for the variance-covariance matrix.[2]

The intuition behind the upper bounding estimator is to use this asymptotic distribution in order to generate a correction for the bias of the estimator defined in equation (1). Specifically, I propose taking random draws from distribution of $\hat{\tau}(\mathcal{T}', \mathcal{T}'')$ and using the

[2]Note, this assumption is easily satisified by estimators that take the form of an ordinary least squares regression or a simple difference in means. For example, those based on matching, fixed effects models, etc.

difference between the maximum of each draw and the known max of expections of those draws (i.e. $\widehat{\text{MCSE}}_q^{\text{Max}}$) as an estimate for the bias of $\widehat{\text{MCSE}}_q^{\text{Max}}$. Formally, let $\hat{\tau}(\mathcal{T}', \mathcal{T}'')^{(b)}$ for $b = 1 \dots B$ denote these draws. Then for each b, let:

$$\widehat{\text{MCSE}}_q^{(b)} = \max_{\mathcal{T}', \mathcal{T}'' \in \mathcal{T}_q} \hat{\tau}(\mathcal{T}', \mathcal{T}'')^{(b)}$$

The estimator then takes the form:

$$\widehat{\text{MCSE}}_q^{\text{Upper}} = \widehat{\text{MCSE}}_q^{\text{Max}} - \left(\frac{1}{B} \sum_{b=1}^{B} \widehat{\text{MCSE}}_q^{(b)} - \widehat{\text{MCSE}}_q^{Max} \right)$$

These assumptions lead to the following proposition, showing the upwards bias of $\widehat{\text{MCSE}}_q^{\text{Upper}}$:

Proposition 1. *Under Assumptions 1-3,*

$$\mathbb{E}\left(\widehat{MCSE}_q^{Max}\right) \geq \mathbb{E}\left(\widehat{MCSE}_q^{Upper}\right) \geq MCSE_q$$

Proof in Appendix C.1

Intuitively, the positive bias in $\widehat{\text{MCSE}}_q^{\text{Upper}}$ emerges from the observation that the difference between the expectation of the maximum of a set of random variables and the maximum of their expectations is decreasing with the difference between the maximum of those expected values and expectations of the remaining random variables. For example, in the extreme case where it is known that the random variable with the largest expectation will be larger than all the other random variables with probability 1, the expectation of the maximum of those random variables will be equal to the expectation of the random variable with the largest expectation.

The principal limitation of this approach is the dependence on the normality of $\hat{\tau}(\mathcal{T}', \mathcal{T}'')$ specified in Assumption 3. This assumption is only likely to be plausible in large samples, and such convergence will likely depend on other assumptions, like that $\mathbf{t}_i$ be low dimensional or that $\tau(\mathcal{T}', \mathcal{T}'')$ satisfy some functional form assumptions like linearity. Although it is

difficult to analyze the finite sample performance of the upper bound analytically, I examine its performance in simulated data in Section 5 and find that it compares favorably both to $\widehat{\text{MCSE}}_q^{\text{Max}}$ and to the lower bound that I introduce in the next section.

In very large samples, the difference between each monte carlo estimate, $\widehat{\text{MCSE}}_q^{(b)}$ and $\widehat{\text{MCSE}}_q^{Max}$ will be minimal. Consequently, so long as the analyst has specified a consistent estimator for $\tau(\mathcal{T}', \mathcal{T}'')$, the resulting upper bound will also be consistent, as formalized in the following proposition:

Proposition 2. *If for all* $\mathcal{T}', \mathcal{T}''' \in \mathcal{T}_q$, $\hat{\tau}(\mathcal{T}', \mathcal{T}'') \xrightarrow[n\to\infty]{p} \tau(\mathcal{T}', \mathcal{T}'')$ *then* $\widehat{MCSE}_q^{Upper} \xrightarrow[n\to\infty]{p} MCSE_q$. *Similarly,* $\mathcal{T}', \mathcal{T}''' \in \mathcal{T}_q$, $\hat{\tau}(\mathcal{T}', \mathcal{T}'') \xrightarrow[n\to\infty]{a.s.} \tau(\mathcal{T}', \mathcal{T}'')$ *then* $\widehat{MCSE}_q^{Upper} \xrightarrow[n\to\infty]{a.s.} MCSE_q$.

Proof in Appendix C.2.

Proposition 2 simply requires that $\hat{\tau}(\mathcal{T}', \mathcal{T}'')$ be a consistent estimator for $\tau(\mathcal{T}', \mathcal{T}'')$, which is distinct from the conditions provided in Assumptions 1-3. For example, a regularized estimator for $\tau(\mathcal{T}', \mathcal{T}'')$ could easily be biased, but consistent.

4.2 Lower Bound

This subsection lays out the lower bounding estimator for the MCSE_q. It begins by describing the estimator and establishing its conservatism and consistency in Section 4.2.1. Section 4.2.2 provides an interval estimator and discusses a framework for conducting classical inference about the lower bound, and Section 4.2.3 provides some brief results useful for implementing the lower bounding estimator.

4.2.1 Estimator Definition and Basic Properties

The lower bounding estimator relies on split sample methods to avoid an upward bias when estimating the MCSE_q. Specifically, I begin by assuming the researcher has randomly split the data into two subsets. One subset will be used to estimate $\hat{\tau}(\mathcal{T}', \mathcal{T}'')$ while the other is used to estimate the probability that any two subsets of $\mathcal{T}_q$ correspond to the true maximum

and minimum causal sets: $\hat{P}(\mathcal{T}' = \mathcal{T}_q^{\text{Max}} \cap \mathcal{T}'' = \mathcal{T}_q^{\text{Min}})$. I retain Assumptions 1 and 2 when developing the lower bound and also assume that $\hat{P}(\mathcal{T}' = \mathcal{T}_q^{\text{Max}} \cap \mathcal{T}'' = \mathcal{T}_q^{\text{Min}})$ satisfy the following basic conditions:

Assumption 4. *$\hat{P}(\mathcal{T}' = \mathcal{T}_q^{Max} \cap \mathcal{T}'' = \mathcal{T}_q^{Min})$ satisfies the following conditions:*

- $\sum_{\mathcal{T}', \mathcal{T}'' \in \mathcal{T}_q} \hat{P}(\mathcal{T}' = \mathcal{T}_q^{Max} \cap \mathcal{T}'' = \mathcal{T}_q^{Min}) = 1$
- $\forall \mathcal{T}', \mathcal{T}'' \in \mathcal{T}_q$,
$$0 \leq \hat{P}(\mathcal{T}' = \mathcal{T}_q^{Max} \cap \mathcal{T}'' = \mathcal{T}_q^{Min}) \leq 1$$
- $\hat{P}(\mathcal{T}' = \mathcal{T}_q^{Max} \cap \mathcal{T}'' = \mathcal{T}_q^{Min}) = 0$ *for all* $\mathcal{T}', \mathcal{T}'' \notin \mathcal{T}_q$

These conditions are extremely mild and simply require that $\hat{P}(\mathcal{T}' = \mathcal{T}_q^{\text{Max}} \cap \mathcal{T}'' = \mathcal{T}_q^{\text{Min}})$ obey the first two Kolmogorov probability axioms and assign zero probability to sets that contain too few treatment types using the analyst specified distribution $g(\mathbf{t})$.

Finally, my proposed lower bound takes the form of an average of the estimates $\hat{\tau}(\mathcal{T}', \mathcal{T}'')$ weighted by $\hat{P}(\mathcal{T}' = \mathcal{T}_q^{\text{Max}} \cap \mathcal{T}'' = \mathcal{T}_q^{\text{Min}})$:

$$\widehat{\text{MCSE}}_q^{\text{Lower}} = \sum_{\mathcal{T}', \mathcal{T}'' \in \mathcal{T}_q} \hat{P}(\mathcal{T}' = \mathcal{T}_q^{\text{Max}} \cap \mathcal{T}'' = \mathcal{T}_q^{\text{Min}}) \hat{\tau}(\mathcal{T}', \mathcal{T}'')$$

Under assumptions 1 and 4, $\widehat{\text{MCSE}}_q^{\text{Lower}}$ can be interpreted as a weighted average of unbiased estimators for the causal effect of being treated with a treatment type in one set rather than another. Because MCSE_q is defined as the maximum of such causal effects for any two subsets of $\mathbb{T}$ with the required probability of occuring, it will always be greater than the expectation of this average, leading to the following proposition:

Proposition 3. *Under assumptions 1,2, and 4*

$$\mathbb{E}\left(\widehat{MCSE}_q^{Lower}\right) \leq MCSE_q$$

Proof in appendix C.3.

It is worth emphasizing this finite sample conservatism requires no assumptions about the good performance of $\hat{P}(\mathcal{T}' = \mathcal{T}_q^{\text{Max}} \cap \mathcal{T}'' = \mathcal{T}_q^{\text{Min}})$. For example, $\hat{P}(\mathcal{T}' = \mathcal{T}_q^{\text{Max}} \cap \mathcal{T}'' = \mathcal{T}_q^{\text{Min}})$ could be based on a misspecified or overregularized model and the corresponding MCSE_q estimator would still be conservative. Additionally, the expectation of $\mathbb{E}\left(\widehat{\text{MCSE}}_q^{\text{Lower}}\right) \geq 0$ as long as $\hat{P}(\mathcal{T}' = \mathcal{T}_q^{\text{Max}} \cap \mathcal{T}'' = \mathcal{T}_q^{\text{Min}})$ is not negatively correlated with $\tau(\mathcal{T}', \mathcal{T}'')$. Consequently, $\widehat{\text{MCSE}}_q^{\text{Lower}}$ will be unbiased when $\text{MCSE}_q = 0$.

However if the analyst is willing to assume consistency for both $\hat{P}(\mathcal{T}' = \mathcal{T}_q^{\text{Max}} \cap \mathcal{T}'' = \mathcal{T}_q^{\text{Min}})$ and $\hat{\tau}(\mathcal{T}', \mathcal{T}'')$ then the corresponding estimator for MCSE_q will also be consistent, leading to the following proposition:

Proposition 4. *If* $\forall \mathcal{T}', \mathcal{T}'' \in \mathcal{T}_q$, $\hat{P}(\mathcal{T}' = \mathcal{T}_q^{Max} \cap \mathcal{T}'' = \mathcal{T}_q^{Min} | \mathcal{S}^{Prob}) \xrightarrow[n\to\infty]{p} \mathbb{1}\{\mathcal{T}' = \mathcal{T}_q^{Max}\}\mathbb{1}\{\mathcal{T}'' = \mathcal{T}_q^{Min}\}$ *and* $\hat{\tau}(\mathcal{T}', \mathcal{T}'') \xrightarrow[n\to\infty]{p} \tau(\mathcal{T}', \mathcal{T}'')$ *then*

$$\widehat{MCSE}_q^{Lower} \xrightarrow[n\to\infty]{p} MCSE_q$$

and if $\forall\, \mathcal{T}', \mathcal{T}'' \in \mathcal{T}_q$, $\hat{P}(\mathcal{T}' = \mathcal{T}_q^{Max} \cap \mathcal{T}'' = \mathcal{T}_q^{Min} | \mathcal{S}^{Prob}) \xrightarrow[n\to\infty]{a.s.} \mathbb{1}\{\mathcal{T}' = \mathcal{T}_q^{Max}\}\mathbb{1}\{\mathcal{T}'' = \mathcal{T}_q^{Min}\}$ *and* $\hat{\tau}(\mathcal{T}', \mathcal{T}'') \xrightarrow[n\to\infty]{a.s.} \tau(\mathcal{T}', \mathcal{T}'')$ *then*

$$\widehat{MCSE}_q^{Lower} \xrightarrow[n\to\infty]{a.s.} MCSE_q$$

Proof in Appendix C.4.

Together with Proposition 2, these results suggest that the upper and lower bounds should be close to each other in a large sample.

4.2.2 Asymptotic Normality and Classical Inference

A useful assumption when demonstrating the asymptotic normality of the lower bounding estimator is that $\hat{\tau}(\mathcal{T}', \mathcal{T}'')$ can be represented as a linear function of the data:

Assumption 5. *Let* $Z = \{\mathbf{t}_i, X_i : i \in \mathcal{S}^{Est}\}$. *For any* $\mathcal{T}', \mathcal{T}'' \in \mathcal{T}_q$ *there exists a set of*

transformations $\{f_i(Z, \mathcal{T}', \mathcal{T}'') : i \in \mathcal{S}^{Est}\}$ *such that:*

$$\hat{\tau}(\mathcal{T}', \mathcal{T}'') = \sum_{i \in \mathcal{S}^{Est}} f_i(\mathbf{t}_i, \mathcal{T}', \mathcal{T}'') Y_i$$

Although at first glance restrictive, nearly all common estimators for causal inference can fit this framework. For example, if $\hat{\tau}(\mathcal{T}', \mathcal{T}'')$ is defined as the difference in outcome means between between units treated with an element of $\mathcal{T}'$ and units treated with an element of $\mathcal{T}''$, then $f_i()$ will be $\frac{1}{\sum_{i \in \mathcal{S}^{\text{Est}}} \mathbb{1}\{\mathbf{t}_i \in \mathcal{T}'\}}$ for units treated with an element of $\mathcal{T}'$, $-\frac{1}{\sum_{i \in \mathcal{S}^{\text{Est}}} \mathbb{1}\{\mathbf{t}_i \in \mathcal{T}''\}}$ for units treated with an element of $\mathcal{T}''$, or zero for units treated with neither. Matching based approaches will work similarly and estimators based on weighting by the inverse probability of treatment match this exact form from the outset. Regression based estimators can also fit this framework, as they are explicitly defined as a linear transformation of the estimation set outcomes.

This assumption eases the derivation of asymptotic normality because it suggests that, under some basic regularity conditions, $\hat{\tau}(\mathcal{T}', \mathcal{T}'')$, will be asymptotically normal. The following proposition uses the central limit theorem derived by Neumann [2013] to show that multiplication by $\hat{P}(\mathcal{T}' = \mathcal{T}_q^{\text{Max}} \cap \mathcal{T}'' = \mathcal{T}_q^{\text{Min}})$ will not impact this convergence so that the asymptotic normality of $\hat{\mathcal{T}}_q^{\text{Max}}$ can be preserved under some mild regularity conditions:

Proposition 5. *If* $\hat{P}(\mathcal{T}' = \mathcal{T}_q^{Max} \cap \mathcal{T}'' = \mathcal{T}_q^{Min})$ *satisfies assumption 4,* $\forall i, \mathbb{E}\left(Y_i^2\right) < \infty$*, and* $\forall \epsilon > 0$*,*

$$\sum_{i \in \mathcal{S}^{Est}} \frac{1}{|\mathcal{S}^{Est}|} \mathbb{E}\left(f_i(Z, \mathcal{T}', \mathcal{T}'')^2 Y_i^2 \mathbb{1}\{|f_i(Z, \mathcal{T}', \mathcal{T}'')| > \epsilon\}\right) \xrightarrow[|\mathcal{S}^{Est}| \to \infty]{} 0$$

Then, conditional on observing the estimation set values of $\mathbf{t}_i$ *and* $\mathbf{x}_i$*,*

$$\frac{\left(\widehat{MCSE_q}^{Lower} - \mathbb{E}\left(\widehat{MCSE_q}^{Lower}\right)\right)}{\sqrt{Var(\widehat{MCSE_q}^{Lower})}} \xrightarrow{D} \mathcal{N}(0, 1)$$

Proof in Appendix C.5

The final result necessary for conducting classical statistical inference is a corresponding variance estimator. This can be most easily accomplished via the non-parametric bootstrap. Specifically, Mammen [1992] shows that the non-parametric bootstrap is consistent for an asymptotically normal estimator that can be represented as a linear transformation of some set of iid data. The following lemma uses assumption 5 to provide just such a result:

Lemma 1.

$$\widehat{MCSE_q}^{Lower} = \sum_{i \in \mathcal{S}^{Est}} Y_i w_i$$

where $w_i = \sum_{\mathcal{T}', \mathcal{T}'' \in \mathcal{T}_q} f_i(Z, \mathcal{T}', \mathcal{T}'')$

Proof. The proof follows trivially by using assumption 5 to substitute $\sum_{i \in \mathcal{S}^{\text{Est}}} f_i(Z, \mathcal{T}', \mathcal{T}'')$ for $\hat{\tau}(\mathcal{T}', \mathcal{T}'')$ in the definition of $\widehat{\text{MCSE}}_q$ and then changing the order of summation. □

So the variance and confidence intervals of $\widehat{\text{MCSE}}_q$ can be consistently estimated by bootstrap resampling from the set $\{Y_i w_i : i \in \mathcal{S}^{\text{Est}}\}$.[3]

4.2.3 Choosing $\hat{P}(\mathcal{T}' = \mathcal{T}_q^{\mathbf{Max}} \cap \mathcal{T}'' = \mathcal{T}_q^{\mathbf{Min}})$ and $\hat{\tau}(\mathcal{T}', \mathcal{T}'')$

A major choice faced by the analyst in implementing this estimator for MCSE_q is specifying estimators for $\hat{P}(\mathcal{T}' = \mathcal{T}_q^{\text{Max}} \cap \mathcal{T}'' = \mathcal{T}_q^{\text{Min}})$ and $\hat{\tau}(\mathcal{T}', \mathcal{T}'')$. While the results in Sections 4.2.1 - 4.2.2 provide minimal conditions under which $\widehat{\text{MCSE}}_q^{\text{Lower}}$ can be used for valid statistical inference, they do not provide much guidance about how to choose such functions. Some intuition can be built by examining the expansion of the mean squared error (MSE) of $\widehat{\text{MCSE}}_q^{\text{Lower}}$ in the following lemma:

[3]Note, clustered standard errors can also be easily generated using the block bootstrap.

Lemma 2.

$$
\begin{aligned}
MSE &= \mathbb{E}\left(\left(\widehat{MCSE}_q^{\,Lower} - MCSE_q\right)^2\right) \\
&= \sum_{\mathcal{T}', \mathcal{T}'' \in \mathcal{T}_q} \mathbb{E}\left(\hat{P}(\mathcal{T}' = \mathcal{T}_q^{Max} \cap \mathcal{T}'' = \mathcal{T}_q^{Min})^2\right) MSE(\hat{\tau}(\mathcal{T}', \mathcal{T}'')) \\
&\quad + \sum_{\substack{\mathcal{T}^*, \mathcal{T}^{**}, \mathcal{T}', \mathcal{T}''' \in \mathcal{T}_q: \\ \mathcal{T}' \neq \mathcal{T}^*, \mathcal{T}'' \neq \mathcal{T}'''}} \mathbb{E}\left(\hat{P}(\mathcal{T}' = \mathcal{T}_q^{Max} \cap \mathcal{T}'' = \mathcal{T}_q^{Min}) \hat{P}(\mathcal{T}^* = \mathcal{T}_q^{Max} \cap \mathcal{T}^{**} = \mathcal{T}_q^{Min})\right) Cov(\hat{\tau}(\mathcal{T}', \mathcal{T}''), \hat{\tau}(\\
&\quad + \sum_{\substack{\mathcal{T}^*, \mathcal{T}^{**}, \mathcal{T}', \mathcal{T}''' \in \mathcal{T}_q: \\ \mathcal{T}' \neq \mathcal{T}^*, \mathcal{T}'' \neq \mathcal{T}'''}} \mathbb{E}\left(\hat{P}(\mathcal{T}' = \mathcal{T}_q^{Max} \cap \mathcal{T}'' = \mathcal{T}_q^{Min}) \hat{P}(\mathcal{T}^* = \mathcal{T}_q^{Max} \cap \mathcal{T}^{**} = \mathcal{T}_q^{Min})\right) \\
&\qquad \times \left(\mathbb{E}\left(\hat{\tau}(\mathcal{T}', \mathcal{T}'')\right) - \tau(\mathcal{T}_q^{Max}, \mathcal{T}_q^{Min})\right) \left(\mathbb{E}\left(\hat{\tau}(\mathcal{T}^*, \mathcal{T}^{**})\right) - \tau(\mathcal{T}_q^{Max}, \mathcal{T}_q^{Min})\right)
\end{aligned}
$$

where the first term represents a scaled sum of the MSE of all functions $\hat{\tau}(\mathcal{T}', \mathcal{T}'')$ used to estimate $\widehat{\mathrm{MCSE}}_q^{\,\mathrm{Lower}}$, the second represents a scaled sun of the covariances of all such $\hat{\tau}(\mathcal{T}', \mathcal{T}'')$, and the third represents the cross product of the biases of those functions. This decomposition suggests that, while the greedy strategy of simply choosing estimators which minimize the MSE of each $\hat{\tau}(\mathcal{T}', \mathcal{T}'')$ is largely sensible, it may not result in the optimal solution. Second, while an estimator for $\hat{P}(\mathcal{T}' = \mathcal{T}_q^{\mathrm{Max}} \cap \mathcal{T}'' = \mathcal{T}_q^{\mathrm{Min}})$ which converges to a binary indicator of whether $\mathcal{T}' = \mathcal{T}_q^{\mathrm{Max}}$ and $\mathcal{T}'' = \mathcal{T}_q^{\mathrm{Min}}$ is necessary for consistency, a binary predictor may not be optimal in finite samples, as averaging estimates of $\hat{\tau}(\mathcal{T}', \mathcal{T}'')$ for different choices of $\mathcal{T}'$ and $\mathcal{T}''$ can improve the efficiency of the $\widehat{\mathrm{MCSE}}_q$. Indeed, Proposition 6 establishes that the performance of $\widehat{\mathrm{MCSE}}_q$ will be maximized when $\hat{P}(\mathcal{T}' = \mathcal{T}_q^{\mathrm{Max}} \cap \mathcal{T}'' = \mathcal{T}_q^{\mathrm{Min}})$ takes on values that are between zero and one:

Proposition 6. $\mathbb{E}\left(\left(\widehat{MCSE}_q^{\,Lower} - MCSE_q\right)^2\right)$ *is minimized when,*

$$
\hat{P}(\mathcal{T}' = \mathcal{T}_q^{Max} \cap \mathcal{T}'' = \mathcal{T}_q^{Min}) = \mathbb{1}\{\mathcal{T}' = \mathcal{T}_q^{Max}\}\mathbb{1}\{\mathcal{T}' = \mathcal{T}_q^{Max}\}
$$

if and only if,

$$P\left(\hat{\tau}\left(\mathcal{T}_q^{Max}, \mathcal{T}_q^{Min}\right) = \tau\left(\mathcal{T}_q^{Max}, \mathcal{T}_q^{Min}\right)\right) = 0$$

Proof in Section C.6

A direct implication of this result is that bootstrap aggregation can be used to improve the performance of any binary predictor for $\hat{P}(\mathcal{T}' = \mathcal{T}_q^{\text{Max}} \cap \mathcal{T}'' = \mathcal{T}_q^{\text{Min}})$ to create a probabilistic estimator without changing the expected value of the predictions.

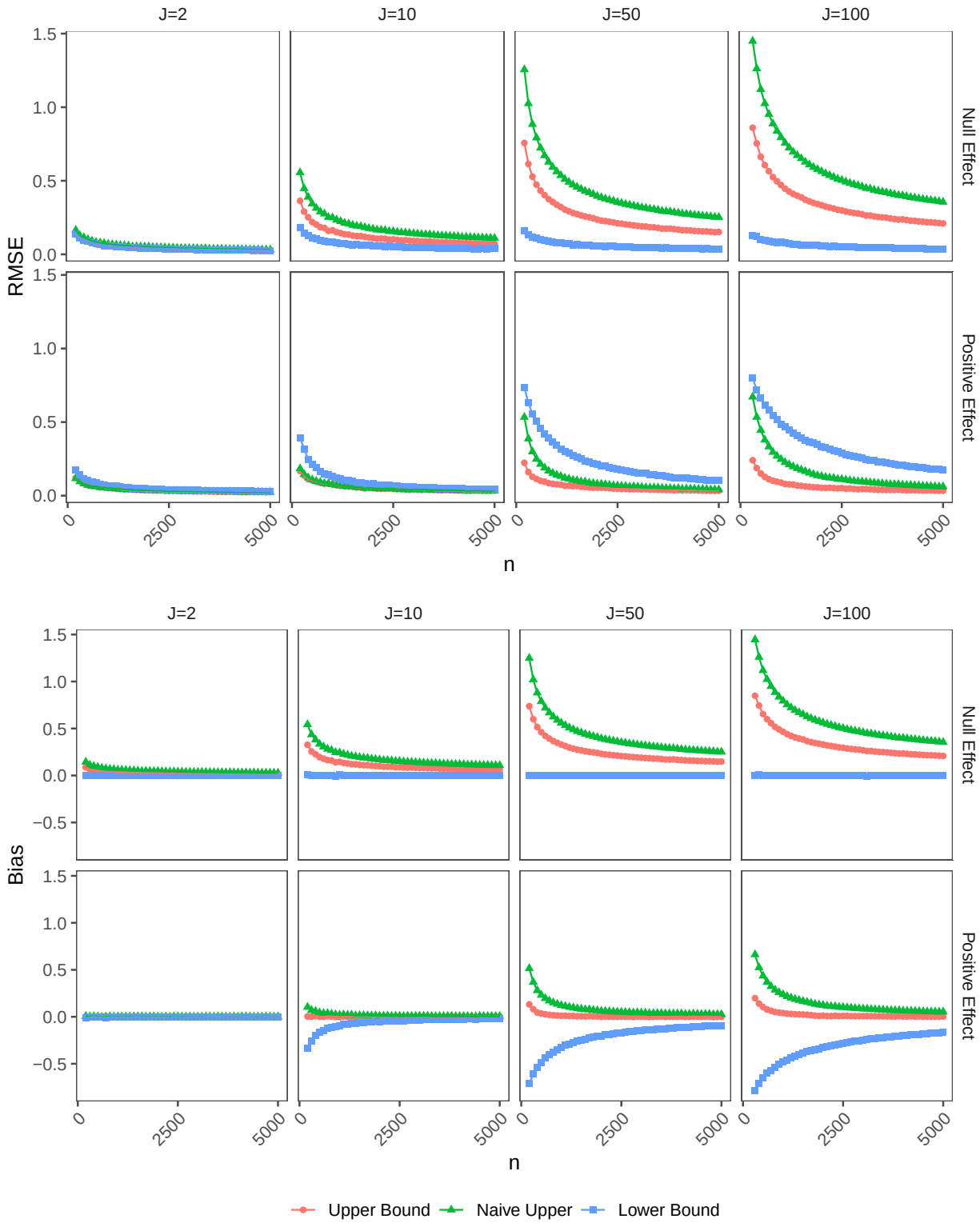
5 Simulations

To explore the finite sample properties of these estimators, I implemented them on simulated data. For each simulation, I first generated n length J vectors $\mathbf{t}_1 \dots \mathbf{t}_n$ to represent the treatment bundles as independent Bernoulli draws with a .5 probability of success. I then generated a length J vector β again as a set of independent draws from the standard normal distribution. Next, I generated the set of conditional means, $\mu_1 \dots \mu_n$ in one of two ways. In the *Positive Effect* case, $\mu_i = \mathbf{t}_i'\beta$ In the *null* case, $\mu_i = 0$. I then generated the outcome as $Y_i = \mu_i + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, 1)$. Finally, in the positive effect case, I normalized the values of Y_i so that the true MCSE_q is equal to 1. I repeated this procedure with values of n between 100 and 5000 and k between 2 and 100.

For each simulation, I compared the performance of the upper and lower bounding estimators, as well as the "naive" upper bound which is simply the maximum of all estimates, $\hat{\tau}(\mathcal{T}', \mathcal{T}'')$, (eq. (1)). In all cases, I used a correctly specified linear model for $\hat{\tau}(\mathcal{T}', \mathcal{T}'')$ and monte carlo sampling from its asymptotic distribution to estimate $\hat{P}(\mathcal{T}' = \mathcal{T}_q^{\text{Max}} \cap \mathcal{T}'' = \mathcal{T}_q^{\text{Min}})$, and I focused on the MCSE when $q = .3$. Figure 1 visualizes the results of this analysis. The top plots show the root mean squared error (RMSE) of both estimators, while the lower ones show their bias. In Appendix B, I present simulations showing similar results for the many moderators case.

In general, the estimators perform as expected. All three estimators converge as the

Figure 1: Proposed Estimators are Consistent and Biased in the Expected Directions in Simulated Data



Note: Blue and red dots identify the performance of my proposed and upper and lower bounding estimators, respectively. The green dots show the performance of the max of all estimates of causal effects, without the monte carlo bias correction used by my upper bound.

sample size increases and the bias is in the expected direction. As expected, the lower bound is always unbiased in the null case and usually delivers the lowest RMSE. On the other hand, the bias is the lowest for my proposed upper bound in the positive effect case. Reassuringly, my proposed upper bound has a lower bias than the naive upper bound in 82% of the simulations and in no case was the difference between the naive upper bound and my proposed upper bound significantly in favor of the naive upper bound. In contrast, my proposed upper bound significantly outperformed the naive upper bound in 26% of simulation parameters tested. These results suggest that the bias correction applied to my proposed upper bound is able to improve on the performance of the uncorrected, "naive" upper bound.

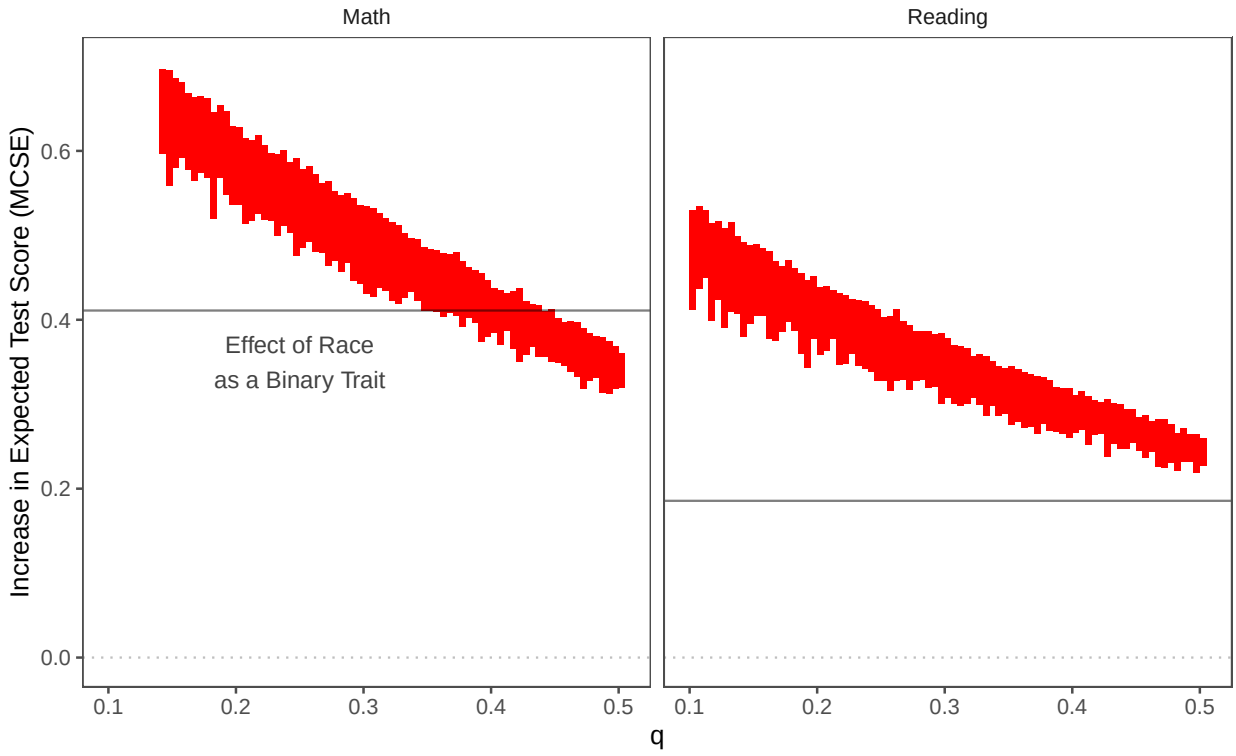
These simulation results can also provide some insight as to the tightness of the bounds when there are many distinct treatment bundles. For example, in the positive effect case when $J = 100$ and $n = 1,000$, the upper bound is still reasonably tight, with an RMSE of just .08 and bias of .034. The performance of the lower bound, on the other hand, is less impressive, with an rmse of .496 and bias of -0.487. This behavior is reversed in the null effect case though, when the lower bound is unbiased, but the upper bound has an rmse of 0.469 and bias of .487.

6 Empirical Examples Revisited

6.1 Race

My first empirical application returns to the setting of Quintana [2021] study, which focused on measuring the effect of race on standardized test scores in the ECLS. Specifically, Quintana [2021] identify 18 variables that they argue represent race in the ECLS (what they term the "markov blanket" of race) and show that they more accurately predict children's performance on standardized test scores than the binary labels do. While Quintana [2021] provide a principled approach for identifying the bundle of variables that composes race in this dataset,

Figure 2: Effect of Race's Markov Blanket on Standardized Test Scores



Note: The y-axis identifies the difference in average test scores between students with one of the q most rather than q least score increasing bundles of racial traits (the MCSE). The red shaded area represents the range between the upper and lower bounds estimated for the MCSE. The lower bound is statistically distinguishable from zero for all values of q.

they lack an estimand that can quantify the full effect of that bundle.

The MCSE provides an approach to measuring the causal effect of race's markov blanket on test scores. Specifically, it quantifies the difference between children with values in the top q rather than bottom q least score promoting values on variables contained within the markov blanket of race. I used a linear regression with no controls to model $\hat{\tau}(\mathcal{T}', \mathcal{T}'')$ and monte carlo sampling from the asymptotic distribution of $\hat{\tau}(\mathcal{T}', \mathcal{T}'')$ for $\hat{P}(\mathcal{T}' = \mathcal{T}_q^{\text{Max}} \cap \mathcal{T}'' = \mathcal{T}_q^{\text{Min}})$.

Figure 2 visualizes the results of this analysis. In the figure, the error bars visualize the range of the upper and lower bounds estimated for the MCSE. These estimates contrast with an ATE associated with race conceptualized as a binary trait of roughly .4 for the Math scores and .2 for the reading outcome. In particular, the estimate for the lower bound on the

MCSE is larger than the estimate for the ATE of the binary race label for values of q greater than .3 on the Math score outcome and .21 on the reading outcome. The lower bound for the MCSE is statistically distinguishable from zero for all values of q presented in Figure 2. Note both outcome measures were generated using an Item Response Theory model to measure student ability on standardized tests. The scores were normalized to range from between -4 and 4.

Examining the regression used to identify the most and least effective treatment bundles can provide some insight about the contents of the sets of most and least effective treatment bundles. The results of this regression are presented in Table A.1 in the supplementary information and show that the bundles correspond well with broad markers of socio-economic status. For example, family income and parental education are the traits most strongly associated with inclusion in the sets of most effective treatment bundles while BMI and school district poverty are most strongly associated with inclusion in the sets of least effective treatment bundles. This squares well with the framework outlined by Sen and Wasow [2016], which suggests that these sorts of more general markers of socio-economic status should be conceptualized as part of the bundle of traits the compose race.

While these estimates suggest that the MCSE can be a useful tool for quantifying the effect of race on important outcomes, some caution should be taken when directly equating the estimates for the MCSE to those for race as a more general concept. In particular, the variables included in the markov blanket were chosen by Quintana [2021] based on their importance in this dataset and the magnitude of the difference in causal effects provided by the MCSE relative to modeling race as a binary trait may not generalize to other settings. Additionally, these estimates are limited by the set of variables available in the ECLS. Indeed, important components of race such as skin tone, manner of speaking, etc are not measured in this dataset, and their absence may limit the magnitude of causal effects estimated. Additionally, the results presented here are premised on a linear model without controls, and there may be unmeasured confounders as well.

Finally, care must always be taken when applying causal reasoning to race. Race is often thought of as an immutable trait, and these results should be instead interpreted in the context of manipulating sets of variables within the markov blanket of race, rather than the latent concept of race itself [Holland, 1986, Sen and Wasow, 2016]. Additionally, it is important to recognize that the variables used in this analysis reflect many avenues of societal discrimination against black people – not an innate inferiority. For example, one variable Quintana [2021] selects is "Perceived interest/competence in peer relationships" which could clearly be affected by the racism of black student's white peers.

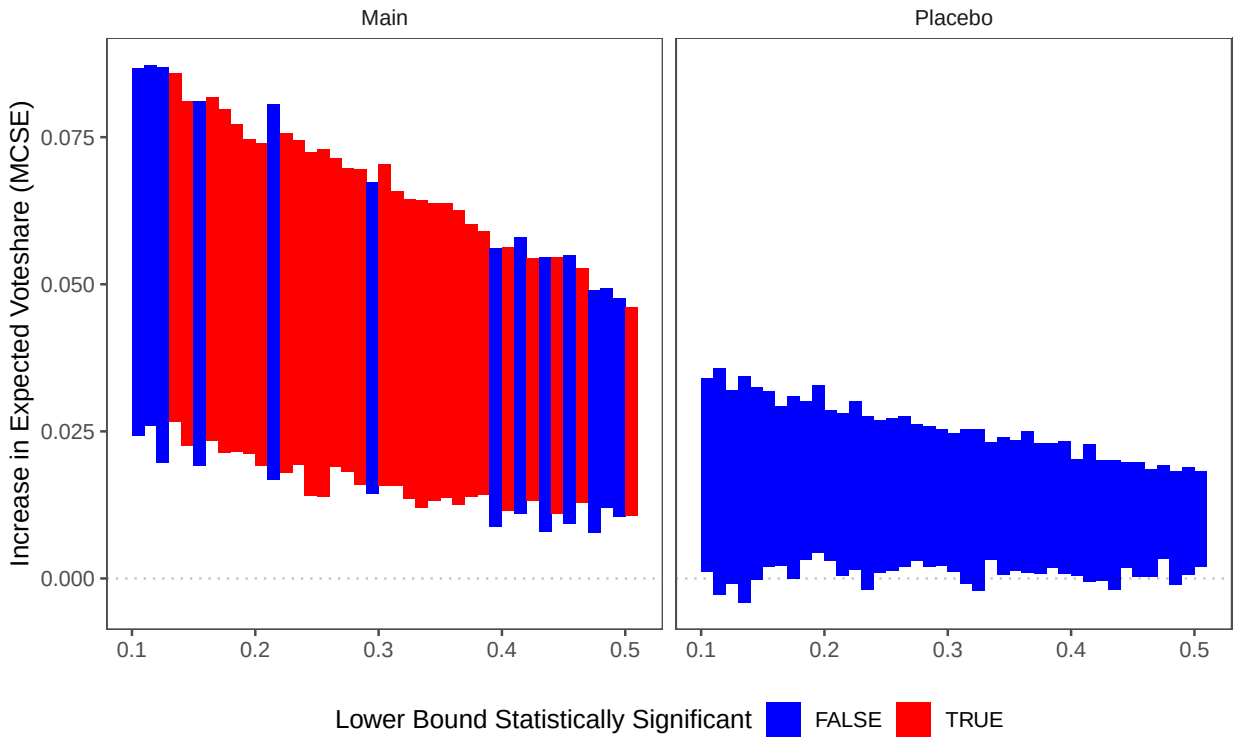
6.2 Campaign Expenditures

The MCSE can also be used to quantify the effect of campaign resource allocation on election outcomes. For this analysis, I used data collected from mandatory campaign expenditures disclosures by the Center for Responsive Politics (CRP). Specifically, I defined the treatment bundles as the fraction of a campaign's expenditures that were allocated into each of 17 categories coded by the Center for Responsive Politics (CRP).[4] I used a linear fixed effects model for $\hat{\tau}(\mathcal{T}', \mathcal{T}'')$ in both upper and lower bound and monte carlo samples from the asymptotic distribution of that model for $\hat{P}(\mathcal{T}' = \mathcal{T}_q^{\text{Max}} \cap \mathcal{T}'' = \mathcal{T}_q^{\text{Min}})$. The fixed effects model included fixed effects for the interaction of a candidate's party and the year; the interaction of the candidate's party and district; the interaction of the state, year, and candidate's party; and for each candidate. It also controls for total spending by the campaign with both a linear and quadratic terms, and the lagged same party voteshare in a district. This is similar to the specification used by Levitt [1994], but adds the state-year-party fixed effect to eliminate confounders that do not vary within a particular state in a particular year. Simpler specifications generally yield larger estimates for the MCSE, so these results may be taken as conservative.

I present the results of the regression used to identify the sets of most and least effective

[4]Note, I excluded categories that did not directly relate to campaign activities, such as transfers to party organizations or other campaigns.

Figure 3: Campaigns With the Most Strategic Expenditures Get More Support



Note: The y-axis identifies the difference in expected voteshare between campaigns with one of the q most and q least effective bundles of campaign expenditures (the MCSE). The shaded area identifies the range between the upper and lower bounds estimated for the MCSE. The color of the shading indicates whether the estimated lower bound was statistically distinguishable from zero for that value of q.

treatment bundles in Table A.2 in the supplementary information. Specifically, it shows the marginal effect of associated with an increase in the proportion of campaign funds allocated to a particular spending category. Because these proportion of spending in all categories is constrained to sum to one, I arbitrarily designated "broadcast spending" as the baseline category. These results suggest that the most effective bundles of campaign activities are disproportionately composed of campaigns that focused on events and technology. In contrast, expenditures on advertising (especially on print ads) appears to be less effective at generating voteshare for a candidate. These results suggest that the most effective campaign expenditures allocation was deployed Democrat Frankie Robbins in Oklahoma 3 in 2010 and the least effective allocation by Democrat Jerry Hilliard in Michigan 4 in 2020.

Figure 3 visualizes the results of this analysis. The left hand panel shows my main results, while the right hand panel presents the results of a placebo test that uses the results from the previous election as they outcome. These estimates suggest that choices about the allocation of campaign expenditures has a statistically significant effect on the outcome of an election. Campaigns with the 10% most effective allocations received between 3 and 9 percentage points more voteshare than campaigns with the 10% least effective allocations. The estimates for the lower bound of the MCSE are statistically significant for values of q between .15 and .4. In contrast, the estimate for the lower bound is consistently near zero and statistically insignificant for all values of q in the right hand panel.

These results suggest that there is more heterogeneity in the benefits that political campaigns have received from their activities than is typically recognized. In particular they suggest that estimates showing that the average effect of campaign spending on election outcomes is quite small (e.g.Kalla and Broockman [2018], Schuster [2020]) might mask considerable variation in the efficacy that certain campaigns receive from their expenditures. Similarly, accounts of election outcomes which emphasize the ways in which two campaigns effectively "cancel out" may need to be updated to consider this sort of heterogeneity [Sides and Vavreck, 2014].

7 Conclusion

Non-parametric estimation techniques and high dimensional datasets increasingly confront researchers with estimates for a huge number of distinct causal estimands. While the capacity to fit such models represents tremendous progress for the estimation and computational techniques that support them, scientific theories rarely make predictions about such a large number of distinct parameters. In this article, I propose a framework for making sense of such model outputs by focusing on the maximum causal contrast between two sets of a researcher specified size q, the MCSE. I develop a bounding approach to estimation about this novel estimand. Both the upper and lower bounds are consistent, and the lower bound is also asymptotically normal, facilitating classical inference about the MCSE. While this estimation approach is developed with the many causes and treatment effect heterogeneity settings in mind, the framework is extremely flexible and could be extended to a myriad of other causal qauntities of interest, speaking to its wide applicability and utility for applied researchers. While a single causal estimand will never replace the kind of careful synthesis and analysis of individual causal variables that should accompany the study of any complex phenomenon, the MCSE will still be a useful tool in a wide array of scientific disciplines.

References

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

Victor Chernozhukov, Mert Demirer, Esther Duflo, and Ivan Fernandez-Val. Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india. Technical report, National Bureau of Economic Research, 2018.

Alexander Coppock, Seth J Hill, and Lynn Vavreck. The small effects of political advertising

are small regardless of context, message, sender, or receiver: Evidence from 59 real-time randomized experiments. *Science Advances*, 6(36):eabc4046, 2020.

Naoki Egami, Christian J Fong, Justin Grimmer, Margaret E Roberts, and Brandon M Stewart. How to make causal inferences using texts. *arXiv preprint arXiv:1802.02163*, 2018.

Christian Fong and Justin Grimmer. Discovery of treatments from text corpora. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 1600–1609, 2016.

Christian Fong and Justin Grimmer. Causal inference with latent treatments. *American Journal of Political Science*, forthcoming. URL `https://www.dropbox.com/s/hxxyn5vtpjuyrw4/dexp_rev.pdf?dl=0`.

Jonathan B Freeman, Andrew M Penner, Aliya Saperstein, Matthias Scheutz, and Nalini Ambady. Looking the part: Social status cues shape race perception. *PloS one*, 6(9): e25107, 2011.

Alan Gerber. Estimating the effect of campaign spending on senate election outcomes using instrumental variables. *American Political science review*, pages 401–411, 1998.

Donald P Green and Holger L Kern. Modeling heterogeneous treatment effects in survey experiments with bayesian additive regression trees. *Public opinion quarterly*, 76(3):491–511, 2012.

Justin Grimmer, Solomon Messing, and Sean J Westwood. Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods. *Political Analysis*, 25(4):413–434, 2017.

Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

Guido W Imbens. The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3):706–710, 2000.

Joshua L Kalla and David E Broockman. The minimal persuasive effects of campaign contact in general elections: Evidence from 49 field experiments. *American Political Science Review*, 112(1):148–166, 2018.

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019.

Steven D Levitt. Using repeat challengers to estimate the effect of campaign spending on election outcomes in the us house. *Journal of Political Economy*, 102(4):777–798, 1994.

Fan Li et al. Propensity score weighting for causal inference with multiple treatments. *The Annals of Applied Statistics*, 13(4):2389–2415, 2019.

Michael J Lopez, Roee Gutman, et al. Estimation of causal effects with multiple treatments: a review and new ideas. *Statistical Science*, 32(3):432–454, 2017.

Enno Mammen. Bootstrap, wild bootstrap, and asymptotic normality. *Probability Theory and Related Fields*, 93(4):439–455, 1992.

Zachary Markovich. Answering complex causal queries with the maximum causal set effect. *Advances in Neural Information Processing Systems*, 34, 2021.

Daniel F McCaffrey, Beth Ann Griffin, Daniel Almirall, Mary Ellen Slaughter, Rajeev Ramchand, and Lane F Burgette. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in medicine*, 32(19):3388–3414, 2013.

Razieh Nabi, Todd McNutt, and Ilya Shpitser. Semiparametric causal sufficient dimension reduction of high dimensional treatments. *arXiv preprint arXiv:1710.06727*, 2017.

Michael H Neumann. A central limit theorem for triangular arrays of weakly dependent random variables, with applications in statistics. *ESAIM: Probability and Statistics*, 17: 120–134, 2013.

Lawrence A Pervin. *The science of personality*. Oxford university press, 2003.

Rafael Quintana. What race and gender stand for: using markov blankets to identify constitutive and mediating relationships. *Journal of Computational Social Science*, pages 1–29, 2021.

Jeremy A Rassen, Abhi A Shelat, Jessica M Franklin, Robert J Glynn, Daniel H Solomon, and Sebastian Schneeweiss. Matching by propensity score in cohort studies with three treatment groups. *Epidemiology*, pages 401–409, 2013.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Aliya Saperstein. Double-checking the race box: Examining inconsistency between survey measures of observed and self-reported race. *Social Forces*, 85(1):57–74, 2006.

Aliya Saperstein and Andrew M Penner. Racial fluidity and inequality in the united states. *American journal of sociology*, 118(3):676–727, 2012.

Steven Sprick Schuster. Does campaign spending affect election outcomes? new evidence from transaction-level disbursement data. *The Journal of Politics*, 82(4):1502–1515, 2020.

Maya Sen and Omar Wasow. Race as a bundle of sticks: Designs that estimate effects of seemingly immutable characteristics. *Annual Review of Political Science*, 19, 2016.

John Sides and Lynn Vavreck. *The gamble*. Princeton University Press, 2014.

John Sides, Lynn Vavreck, and Christopher Warshaw. The effect of television advertising in united states elections.

Matthew Stephens and David J Balding. Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics*, 10(10):681–690, 2009.

Peter M Visscher, Naomi R Wray, Qian Zhang, Pamela Sklar, Mark I McCarthy, Matthew A Brown, and Jian Yang. 10 years of gwas discovery: biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22, 2017.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Yixin Wang and David M Blei. The blessings of multiple causes. *Journal of the American Statistical Association*, pages 1–71, 2019.

Yixin Wang, Dawen Liang, Laurent Charlin, and David M Blei. The deconfounded recommender: A causal inference approach to recommendation. *arXiv preprint arXiv:1808.06581*, 2018.

Shu Yang, Guido W Imbens, Zhanglin Cui, Douglas E Faries, and Zbigniew Kadziola. Propensity score matching and subclassification in observational studies with multi-level treatments. *Biometrics*, 72(4):1055–1065, 2016.

Jiajing Zheng, Alexander D'Amour, and Alexander Franks. Copula-based sensitivity analysis for multi-treatment causal inference with unobserved confounding. *Journal of the American Statistical Association*, Forthcoming.

Hao Zou, Peng Cui, Bo Li, Zheyan Shen, Jianxin Ma, Hongxia Yang, and Yue He. Counterfactual prediction for bundle treatment. *Advances in Neural Information Processing Systems*, 33, 2020.

Supplementary Materials for

"Causal Inference With Bundled Variables"

A Additional Regression Results

This section presents the results of the regressions used to estimate the sets of most and least effective treatment bundles in Section 6.

Table A.1: Effect of Race's Markov Blanket on Standardized Test Scores

	Dependent variable:	
	Math	Reading
	(1)	(2)
City	0.250*	0.112
	(0.150)	(0.128)
Suburb	0.197	0.088
	(0.149)	(0.127)
Town	0.190	0.039
	(0.160)	(0.137)
Rural	0.206	0.032
	(0.150)	(0.128)
School district poverty	−0.003**	−0.003**
	(0.001)	(0.001)
BMI	−0.012***	−0.007*
	(0.005)	(0.004)
Number of books the child has	0.0001	0.0001
	(0.0001)	(0.0001)
Inhibitory control and attention	0.143***	0.077***
	(0.013)	(0.011)
Teacher reported interpersonal skills	0.073**	0.027
	(0.034)	(0.029)
Teacher reported self control	0.046	0.076**
	(0.034)	(0.029)
Time spent playing videogames	0.001	0.001
	(0.001)	(0.001)
Family income	0.016***	0.012***
	(0.003)	(0.003)
Parent educational expectations for Child	0.0003	0.009
	(0.011)	(0.010)
Parental education	0.044***	0.030***
	(0.008)	(0.007)
Constant	−1.217***	−0.502**
	(0.231)	(0.197)
Observations	1,514	1,514
R^2	0.237	0.170
Adjusted R^2	0.230	0.162
Residual Std. Error (df = 1499)	0.506	0.433
F Statistic (df = 14; 1499)	33.255***	21.891***
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table A.2: Effect of Campaign Spending on Electoral Support (Relative to Broadcast ads)

	Dependent variable:
	Voteshare
Accounting & legal	0.016
	(0.020)
Admin event expenses	0.138***
	(0.044)
Admin travel	0.094**
	(0.037)
Campaign materials	−0.006
	(0.016)
Misc admin	0.163
	(0.106)
Misc media	−0.015
	(0.014)
Print ads	−0.083
	(0.064)
Rent & utilities	0.115**
	(0.045)
Salaries	0.049**
	(0.020)
Admin data & tech	0.171*
	(0.090)
Campaign events	0.100**
	(0.048)
Media buys	−0.022
	(0.014)
Media production	0.040
	(0.041)
Media consulting	0.016
	(0.020)
Web ads	−0.035
	(0.034)
Admin consulting	0.048
	(0.058)
Observations	3,208
R^2	0.901
Adjusted R^2	0.759
Residual Std. Error	0.072 (df = 1315)
Note:	*p≤0.1; **p<0.05; ***p<0.01

B Simulation Results for Many Moderators

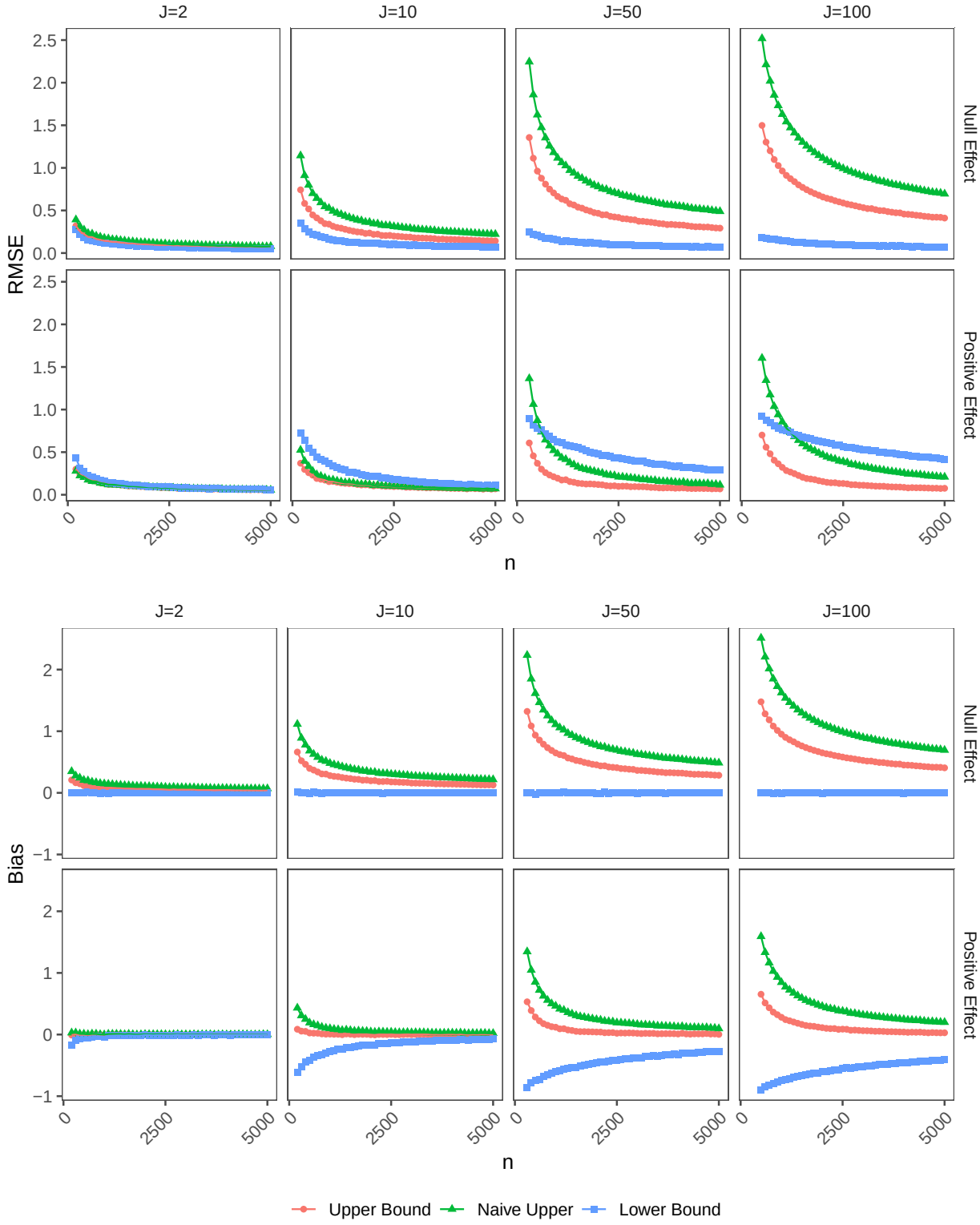
Figure B.1 presents results for the performance of all estimators when $\tau(\mathcal{T}', \mathcal{T}'')$ is defined as the difference in average treatment effects between units that received moderator bundles in $\mathcal{T}'$ instead of $\mathcal{T}''$. The data is constructed in the same as before except for,

$$Y_i = \mu_i T_i + \epsilon$$

where T_i is a bernouli random variable with a probabilty of success of .5.

The results in figure B.1 are very similar to those observed in the many causes case, although the overall level error is higher. This is unsurprsing, as the conditional average treatment effects for each treatment type are not as precisely estimated as the conditional means are.

Figure B.1: Simulation Results for Many Moderators Case



Note: This figure visualizes the performance of the proposed MCSE estimators on simulated data. J identifies the number of elements in the vector representing the treatment bundles, while n identifies the number of units. The positive effect label refers to simulations where the true MCSE is 1, while the null effect one refers to simulations where the true MCSE is 0. Top plots visualize the root mean squared error (RMSE) of the estimators, while the lower plots show their bias.

C Proofs

C.1 Proof of Prop 1

Proof. First, note that for any two vectors, v and w,

$$\max v + w \leq \max v + \max w$$

And that for any normally distributed random vectors, A, B, C such that $\mathbb{E}(A) = \mathbb{E}(B) = \mathbb{E}(C)$ and $\text{Cov}(A) = \text{Cov}(B)$, and $\text{Cov}(C) = \gamma \text{Cov}(A)$ for some constant $\gamma \in (0, 1)$:

$$\mathbb{E}(\max A + \max B - \max(A + B)) \geq \mathbb{E}(\max C + \max B - \max(C + B))$$

Because C will have the same distribution as A, but with a lower level of dispersion.

Now linearly decompose $\hat{\tau}(\mathcal{T}', \mathcal{T}'')$ as $\tau(\mathcal{T}', \mathcal{T}'') + \epsilon(\mathcal{T}', \mathcal{T}')$ where the set of normal random variables $\{\epsilon(\mathcal{T}', \mathcal{T}'') : \mathcal{T}', \mathcal{T}'' \in \mathcal{T}_q\}$ are mean zero with the same variance covariance matrix as $\{\hat{\tau}(\mathcal{T}', \mathcal{T}'') : \mathcal{T}', \mathcal{T}'' \in \mathcal{T}_q\}$, which I denote Σ. Analogously, decompose $\hat{\tau}(\mathcal{T}', \mathcal{T}'')^{(b)}$ as $\tau(\mathcal{T}', \mathcal{T}'') + \epsilon(\mathcal{T}', \mathcal{T}') + \xi(\mathcal{T}', \mathcal{T}')$, and note that $\{\epsilon(\mathcal{T}', \mathcal{T}'') : \mathcal{T}', \mathcal{T}'' \in \mathcal{T}_q\}$ is again normally distributed and mean zero, but with the estimated variance-covariance matrix, $\hat{\Sigma}$. Recall that by assumption, $\mathbb{E}\left(\hat{\Sigma}\right) = \Sigma$. So:

$$\begin{aligned}
&\mathbb{E}(\max \tau(\mathcal{T}', \mathcal{T}'') + \epsilon(\mathcal{T}', \mathcal{T}')) + \mathbb{E}(\max \xi(\mathcal{T}', \mathcal{T}')) - \mathbb{E}(\max \tau(\mathcal{T}', \mathcal{T}'') + \epsilon(\mathcal{T}', \mathcal{T}') + \xi(\mathcal{T}', \mathcal{T}')) \\
&\quad \geq \mathbb{E}(\max \tau(\mathcal{T}', \mathcal{T}'') + \max \epsilon(\mathcal{T}', \mathcal{T}')) - \mathbb{E}(\max \tau(\mathcal{T}', \mathcal{T}'') + \epsilon(\mathcal{T}', \mathcal{T}'))
\end{aligned}$$

Therefore,

$$
\begin{aligned}
&= \mathbb{E}\left(\widehat{\mathrm{MCSE}}_q^{(b)}\right) - \mathbb{E}\left(\widehat{\mathrm{MCSE}}_q\right) \\
&= \mathbb{E}\left(\max \hat{\tau}(\mathcal{T}', \mathcal{T}'')^{(b)}\right) - \mathbb{E}\left(\max \hat{\tau}(\mathcal{T}', \mathcal{T}'')\right) \\
&= \mathbb{E}\left(\max \tau(\mathcal{T}', \mathcal{T}'') + \epsilon\left(\mathcal{T}', \mathcal{T}'\right) + \xi\left(\mathcal{T}', \mathcal{T}'\right)\right) - \mathbb{E}\left(\max \tau(\mathcal{T}', \mathcal{T}'') + \epsilon\left(\mathcal{T}', \mathcal{T}'\right)\right) \\
&\leq \mathbb{E}\left(\max \tau(\mathcal{T}', \mathcal{T}'') + \epsilon\left(\mathcal{T}', \mathcal{T}'\right)\right) + \mathbb{E}\left(\max \xi\left(\mathcal{T}', \mathcal{T}'\right)\right) - \mathbb{E}\left(\max \tau(\mathcal{T}', \mathcal{T}'') - \max \epsilon\left(\mathcal{T}', \mathcal{T}'\right)\right) \\
&= \mathbb{E}\left(\max \tau(\mathcal{T}', \mathcal{T}'') + \epsilon\left(\mathcal{T}', \mathcal{T}'\right)\right) - \max \tau(\mathcal{T}', \mathcal{T}'') \\
&= \mathbb{E}\left(\max \hat{\tau}(\mathcal{T}', \mathcal{T}'')(b)\right) - \max \tau(\mathcal{T}', \mathcal{T}'') \\
&= \mathbb{E}\left(\widehat{\mathrm{MCSE}}_q\right) - \mathrm{MCSE}_q
\end{aligned}
$$

□

C.2 Proof of Proposition 2

Proof. Throughout this proof, super-scripting or sub-scripting with (n) will be used to indicate a random variable that depends on the sample size. $\widehat{\mathrm{MCSE}}_q^{\text{Upper-n}}$ and $\widehat{\mathrm{MCSE}}_q^{\text{Max-n}}$ will do the same for $\widehat{\mathrm{MCSE}}_q^{\text{Upper}}$ and $\widehat{\mathrm{MCSE}}_q^{\text{Max}}$, respectively.

Note, $\hat{\tau}(\mathcal{T}', \mathcal{T}'')_{(n)} \xrightarrow[n\to\infty]{p} \tau(\mathcal{T}', \mathcal{T}'')$ implies that the distribution of $\hat{\tau}(\mathcal{T}', \mathcal{T}'')$ converges to a point mass at 0. Therefore, for all b,

$$
\hat{\tau}(\mathcal{T}', \mathcal{T}'')_{(n)}^{(b)} \xrightarrow[n\to\infty]{p} \tau(\mathcal{T}', \mathcal{T}'')
$$

Because max is a continous operator, by the continuous mapping theorem:

$$\widehat{\mathrm{MCSE}}_q{}^{\text{Upper-n}} = \max_{t' \in \mathcal{T}_q} \hat{\tau}(\mathcal{T}', \mathcal{T}'') \tag{2}$$

$$\xrightarrow[n\to\infty]{p} \max_{t' \in \mathcal{T}_q} \tau(\mathcal{T}', \mathcal{T}'') \tag{3}$$

$$= \mathrm{MCSE}_q \tag{4}$$

Therefore,

$$\widehat{\mathrm{MCSE}}_q{}^{\text{Upper-n}} = \widehat{\mathrm{MCSE}}_q{}^{\text{Max-n}} - \left(\frac{1}{B} \sum_{b=1}^{B} \widehat{\mathrm{MCSE}}_{q(n)}{}^{(b)} - \widehat{\mathrm{MCSE}}_q{}^{\text{Max-n}} \right) \tag{5}$$

$$\xrightarrow[n\to\infty]{p} \mathrm{MCSE}_q - \left(\frac{1}{B} \sum_{b=1}^{B} \mathrm{MCSE}_q - \mathrm{MCSE}_q \right) \tag{6}$$

$$= \mathrm{MCSE}_q \tag{7}$$

$$\tag{8}$$

□

C.3 Proof for Proposition 3

Proof. Note, by construction, for any $\mathcal{T}', \mathcal{T}'' \in \mathcal{T}_q$:

$$\mathrm{MCSE}_q = \max_{\mathcal{T}', \mathcal{T}'' \in \mathcal{T}_q} \tau(\mathcal{T}', \mathcal{T}'') = \tau\left(\mathcal{T}_q^{\text{Max}}, \mathcal{T}_q^{\text{Min}}\right) \geq \tau(\mathcal{T}', \mathcal{T}'')$$

Therefore, from assumptions 4 and 1,

$$
\begin{aligned}
\text{MCSE}_q &\geq \sum_{\mathcal{T}',\mathcal{T}''\in\mathcal{T}_q} \mathbb{E}\left(\hat{P}(\mathcal{T}' = \mathcal{T}_q^{\text{Max}} \cap \mathcal{T}'' = \mathcal{T}_q^{\text{Min}})\right) \tau(\mathcal{T}', \mathcal{T}'') \\
&= \sum_{\mathcal{T}',\mathcal{T}''\in\mathcal{T}_q} \mathbb{E}\left(\hat{P}(\mathcal{T}' = \mathcal{T}_q^{\text{Max}} \cap \mathcal{T}'' = \mathcal{T}_q^{\text{Min}})\right) \mathbb{E}\left(\hat{\tau}(\mathcal{T}', \mathcal{T}'')\right) \\
&= \mathbb{E}\left(\widehat{\text{MCSE}}_q^{\text{Lower}}\right)
\end{aligned}
$$

□

C.4 Proof of Theorem 4

Proof. In the following proof, I use super-scripting by (n) to denote a quantity that is dependent on the sample size and has been estimated using a sample of size n. In particular, $\widehat{\text{MCSE}}_q^{Lower-n}$ will be used to emphasize the dependence of $\widehat{\text{MCSE}}_q^{Lower}$ on n.

Turning to the proof, first consider the case when $\forall \mathcal{T}', \mathcal{T}'' \subseteq \mathbb{T}$:,

$$
\hat{P}^{(n)}(\mathbb{T}' = \mathcal{T}_q^{\text{Max}} \cap \mathbb{T}'' = \mathcal{T}_q^{\text{Min}} | \mathcal{S}^{\text{Prob}}) \xrightarrow[n\to\infty]{p} \mathbb{1}\{\mathcal{T}' = \mathcal{T}_q^{\text{Max}}\}\mathbb{1}\{\mathcal{T}' = \mathcal{T}_q^{\text{Max}}\}
$$

and,

$$
\hat{\tau}^{(n)}(\mathcal{T}', \mathcal{T}'') \xrightarrow[n\to\infty]{p} \tau(\mathcal{T}', \mathcal{T}'')
$$

Then,

$$
\begin{aligned}
\widehat{\mathrm{MCSE}}_q^{\,Lower-n} &= \sum_{\mathcal{T}',\mathcal{T}''\in\mathcal{T}_q} \hat{P}^{(n)}(\mathbb{T}' = \mathcal{T}_q^{\mathrm{Max}} \cap \mathbb{T}'' = \mathcal{T}_q^{\mathrm{Min}} | \mathcal{S}^{\mathrm{Prob}}) \hat{\tau}^{(n)}(\mathcal{T}', \mathcal{T}'') \\
&\xrightarrow[n\to\infty]{p} \sum_{\mathcal{T}',\mathcal{T}''\in\mathcal{T}_q} \mathbb{1}\{\mathcal{T}' = \mathcal{T}_q^{\mathrm{Max}}\}\mathbb{1}\{\mathcal{T}' = \mathcal{T}_q^{\mathrm{Max}}\} \hat{\tau}^{(n)}(\mathcal{T}', \mathcal{T}'') \\
&\xrightarrow[n\to\infty]{p} \tau(\mathcal{T}_q^{\mathrm{Max}}, \mathcal{T}_q^{\mathrm{Min}}) \\
&= \mathrm{MCSE}_q
\end{aligned}
$$

The proof for almost sure convergence is identical to that for convergence in probability, however, convergence in probability is replaced with almost sure convergence in all cases.

□

C.5 Proof for Proposition 5

Proof. As is the case in Appendix C.4, I use super-scripting by (n) to denote a quantity that is dependent on the sample size and has been estimated using a sample of size n. In particular, $\widehat{\mathrm{MCSE}}_q^{\,Lower-n}$ will be used to emphasize the dependence of $\widehat{\mathrm{MCSE}}_q^{\,Lower}$ on n.

For any $\mathcal{T}', \mathcal{T}'' \in \mathcal{T}_q$ let $v_i^{(n)} = f_i^{(n)}(Z, \mathcal{T}', \mathcal{T}'')Y_i - \mathbb{E}\left(f_i^{(n)}(Z, \mathcal{T}', \mathcal{T}'')Y_i\right)$ where $f_i^{(n)}()$ and Z have the same definition as in assumption 5. Also, to simplify notation, let $p^{(n)} = \hat{P}(\mathcal{T}' = \mathcal{T}_q^{\mathrm{Max}} \cap \mathcal{T}'' = \mathcal{T}_q^{\mathrm{Min}})$.

Neumann [2013] shows that a sufficient condition for $\sum_{i\in\mathcal{S}^{\mathrm{Est}}} p^{(n)} v_i^{(n)}$ to be asymptotically normal is that along with satisfying the lindberg-feller condition and having finite second moments (which are both assumed in the statement of the proposition), for any i, j, and any measurable function l such that $||l||_\infty \leq 1$,

$$
\mathrm{Cov}(l(p^{(n)} v_{-j}^{(n)}) p^{(n)} v_i^{(n)}, p^{(n)} v_j^{(n)} | Z) = 0
$$

and

$$\text{Cov}(l(p^{(n)}v^{(n)}_{-i,j}), p^{(n)}v^{(n)}_i p^{(n)}v^{(n)}_j) = 0$$

where $p^{(n)}v^{(n)}_{-j}$ denotes the vector of outcomes except for $v^{(n)}_i$ and $p^{(n)}v^{(n)}_{-i,j}$ denotes the same, but exempting both units i and j. To simplify the notation in the following derivations, I omit the explicit conditioning on Z, but all expectations and covariance should be understood as being conditional on Z. However, note the assumption that $Y_i \perp\!\!\!\perp Y_j$ for all $i \neq j$ implies that conditional on Z, $v^{(n)}_i \perp\!\!\!\perp v^{(n)}_j$

Now, considering the first condition,

$$\begin{aligned}
&\text{Cov}(l(p^{(n)}v^{(n)}_{-j})p^{(n)}v^{(n)}_i, p^{(n)}v^{(n)}_j) \\
&= \mathbb{E}\left(l(p^{(n)}v^{(n)}_{-j})p^{(n)}v^{(n)}_i p^{(n)}v^{(n)}_j\right) - \mathbb{E}\left(l(p^{(n)}v^{(n)}_{-j})p^{(n)}v^{(n)}_i\right)\mathbb{E}\left(p^{(n)}v^{(n)}_j\right) \\
&= \mathbb{E}\left(l(p^{(n)}v^{(n)}_{-j})p^{(n)}v^{(n)}_i p^{(n)}\right)\mathbb{E}\left(v^{(n)}_j\right) - \mathbb{E}\left(l(p^{(n)}v^{(n)}_{-j})p^{(n)}v^{(n)}_i\right)\mathbb{E}\left(p^{(n)}\right)\mathbb{E}\left(v^{(n)}_j\right) \\
&= \mathbb{E}\left(v^{(n)}_j\right)\left(\mathbb{E}\left(l(p^{(n)}v^{(n)}_{-j})p^{(n)}v^{(n)}_i p^{(n)}\right) - \mathbb{E}\left(l(p^{(n)}v^{(n)}_{-j})p^{(n)}v^{(n)}_i\right)\mathbb{E}\left(p^{(n)}\right)\right) \\
&= 0\left(\mathbb{E}\left(l(p^{(n)}v^{(n)}_{-j})p^{(n)}v^{(n)}_i p^{(n)}\right) - \mathbb{E}\left(l(p^{(n)}v^{(n)}_{-j})p^{(n)}v^{(n)}_i\right)\mathbb{E}\left(p^{(n)}\right)\right) \\
&= 0
\end{aligned}$$

Next, the second condition,

$$
\begin{aligned}
&\mathrm{Cov}(l(p^{(n)}v_{-i,j}), p^{(n)}v_i p^{(n)}v_j) \\
&= \mathbb{E}\left(l(p^{(n)}v_{-i,j})p^{(n)}v_i p^{(n)}v_j\right) - \mathbb{E}\left(l(p^{(n)}v_{-i,j})\right)\mathbb{E}\left(p^{(n)}v_i p^{(n)}v_j\right) \\
&= \mathbb{E}\left(l(p^{(n)}v_{-i,j})\left(p^2\right)^{(n)}\right)\mathbb{E}\left(v_i\right)\mathbb{E}\left(v_j\right) - \mathbb{E}\left((p^{(n)}v_{-i,j})\right)\mathbb{E}\left(\left(p^{(n)}\right)\right)^2\mathbb{E}\left(v_i^{(n)}\right)\mathbb{E}\left(v_j^{(n)}\right) \\
&= \mathbb{E}\left(v_i^{(n)}\right)\mathbb{E}\left(v_j^{(n)}\right)\left(\mathbb{E}\left((p^{(n)}v_{-i,j})\left(p^{(n)}\right)^2\right) - \mathbb{E}\left((p^{(n)}v_{-i,j}^{(n)})\right)\mathbb{E}\left(\left(p^{(n)}\right)^2\right)\right) \\
&= 0\left(\mathbb{E}\left((p^{(n)}v_{-i,j}^{(n)})\left(p^{(n)}\right)^2\right) - \mathbb{E}\left((p^{(n)}v_{-i,j}^{(n)})\right)\mathbb{E}\left(\left(p^{(n)}\right)^2\right)\right) \\
&= 0
\end{aligned}
$$

So, we conclude that $\sum_{i\in\mathcal{S}^{\mathrm{Est}}} p^{(n)}v_i^{(n)} \xrightarrow[n\to\infty]{D} \mathcal{N}\left(0, \mathrm{Var}\left(\sum_{i\in\mathcal{S}^{\mathrm{Est}}} v_i^{(n)}\right)\right)$. Since the sum of normal random variables is also normally distributed, we then conclude that, $\widehat{\mathrm{MCSE}}_q^{\,Lower-n} = \sum_{\mathcal{T}',\mathcal{T}''\in\mathcal{T}_q} \hat{P}^{(n)}(\mathcal{T}'' = \mathcal{T}_q^{\mathrm{Max}} \cap \mathcal{T}'' = \mathcal{T}_q^{\mathrm{Min}}|\mathcal{S}^{\mathrm{Prob}})\hat{\tau}^{(n)}(\mathcal{T}', \mathcal{T}'')$ will also be asymptotically normal, completing the proof. □

C.6 Proof for Proposition 6

Arbitrarily index every ordered pair $(\mathcal{T}', \mathcal{T}'') \in \mathcal{T}_q$ with $1 \dots J$ and let $\hat{p}_i$ and $\hat{\tau}_i$ denote the corresponding random variables $\hat{P}(\mathcal{T}' = \mathcal{T}_q^{\mathrm{Max}} \cap \mathcal{T}'' = \mathcal{T}_q^{\mathrm{Min}})$ and $\hat{\tau}(\mathcal{T}', \mathcal{T}'')$ and let d_i represent the binary prediction for whether $\mathcal{T}_i' = \mathcal{T}_q^{\mathrm{Max}}$ and $\mathcal{T}_i' = \mathcal{T}_q^{\mathrm{Min}}$ from the binary predictor. Similarly, let $\epsilon_i = \hat{\tau}_i - \tau$. First note that by the tower property,

$$\begin{aligned}
\mathbb{E}\left(\left(\sum_{i=1}^{J} d_i\epsilon_i\right)^2\right) &= \mathbb{E}\left(\mathbb{E}\left(\left(\sum_{i=1}^{J} d_i\epsilon_i\right)^2 \middle| d\right)\right) \\
&= \mathbb{E}\left(\mathbb{E}\left(\left(\sum_{i=1}^{J} \mathbb{1}\{d_i = 1\}\epsilon_i\right)^2 \middle| d\right)\right) \\
&= \mathbb{E}\left(\left(\sum_{i=1}^{J} \mathbb{E}\left(\mathbb{1}\{d_i = 1\}\epsilon_i^2\right) \middle| d\right)\right) \\
&= \sum_{i=1}^{J} \mathbb{E}\left(\mathbb{1}\{d_i = 1\}\right)\mathbb{E}\left(\epsilon_i^2\right) \\
&= \sum_{i=1}^{J} \mathbb{E}\left(\hat{p}_i\right)\mathbb{E}\left(\epsilon_i^2\right)
\end{aligned}$$

And by the Cauchy Schwarz inequality,

$$\begin{aligned}
\mathbb{E}\left(\left(\sum_{i=1}^{J} \hat{p}_i\epsilon_i\right)^2\right) &= \mathbb{E}\left(\sum_{i,j=1}^{J} \hat{p}_i\epsilon_i\hat{p}_j\epsilon_j\right) \\
&\leq \mathbb{E}\left(\sum_{i=1}^{J} \hat{p}_i^2\epsilon_i^2\right) \\
&\leq \sum_{i=1}^{J} \mathbb{E}\left(\hat{p}_i^2\right)\mathbb{E}\left(\epsilon_i^2\right) \\
&\leq \sum_{i=1}^{J} \mathbb{E}\left(\hat{p}_i\right)\mathbb{E}\left(\epsilon_i^2\right)
\end{aligned}$$

where the final inequality holds because $\hat{p}_i$ is between 0 and 1.