Where to Experiment?
A Best Subsets Ensemble for Purposive Site Selection

Adam Bouyamourn*

August 29, 2024

Abstract

Choosing where to conduct an experiment when a researcher is faced with a universe of possible sites is an important practical problem for applied researchers. I propose a new method to select experimental sites optimal in terms of Mean Squared Error for the Population Average Treatment Effect (PATE) and the Conditional Average Treatment Effect (CATE). I show that finding sites optimal for the PATE requires selected sites that minimize distance between subset covariate means and population covariate means. For the CATE, the optimal site selection minimizes the difference in distributions between the selected subset and the population. I develop a novel ensemble method that uses Best Subsets as a base learner to implement this. I show how the user can incorporate prior information from pilot studies, information about site variances and sample sizes, and user-specified welfare weights when distributional impacts of a policy are of interest. Naturalistic simulation studies demonstrate the performance of the proposed method.

*PhD Candidate, Charles and Louise Travers Department of Political Science, University of California, Berkeley. I am grateful to Kirk Bansak, Eli Ben-Michael, Thad Dunning, Naoki Egami, Avi Feller, Erin Hartman, Tara Slough, participants at EITM 2024, and the Berkeley Methods Workshop for helpful feedback. An R package to implement the proposed method, `BSESS`, is forthcoming.

1 Introduction

Multi-site experiments have become increasingly popular across the social sciences, as researchers seek to improve the generalizability and representativeness of conclusions learned from experiments (Dunning et al., 2019; Slough and Tyson, 2023). Recent attention has focused on the statistical properties of an procedure in which a researcher must decide *where* to experiment, before *actually conducting* the experiment.

Randomization has well-studied finite sample properties (Li and Ding, 2016), and is minmax regret optimal without structural assumptions (Kallus, 2018). However, the site selection problem arguably highlights the worst-case performance of randomization: the number of chosen sites is small; there may be significant between-site variation, and the regret of a given randomization procedure can be high (Bertsimas et al., 2015a; Kasy, 2016). Further, we may have access to good information about sites that allow us to make informed decisions about what a 'good' or representative allocation of sites would look like.

1.1 Contributions

First, I consider the problem of selecting sites that minimize the Mean Square Error of the experimental estimate of the PATE and the CATE. When covariates are observed and informative, and under an assumption of no unobserved confounding, I show that we can represent the former problem one of minimizing the distance of mean vectors between subset and population, and that a constrained solution to this problem chooses the set of sites that minimize the ℓ^2 distance between covariate means in the population and selected subsample. This result is similar to that of Olea et al. (2024), who show that k-medians minimizes the worst-case welfare-regret of a site selection, and to that of Tipton and Mamakos (2023), who show a similar result in the context of covariate balancing vectors.

For the CATE, I show that an optimal set of sites in a Mean Squared Error sense minimizes the ℓ^2 *discrepancy* between covariates in the sample and the selected subset. I prove that this can be approximated by minimizing the ℓ^2 *Quantile Discrepancy* between selected covariates in the subset and population. I show a connection between this problem and minimizing the Wasserstein Distance between sample and population: as the quantile mesh over which discrepancy is minimizing becomes increasingly fine-grained, the two are equivalent, and any solution also solves the optimal transport problem. To the best of my knowledge, the problem of selecting experimental *sites* for the CATE has not previously been studied (Arbour et al., 2021).

To solve these problems in practice, I highlight the connection between the above results and the classic Best Subsets problem in machine learning. Best Subsets finds K-sparse solutions to estimation problems, which makes it especially appropriate for site selection problems in which a Researcher faces a budget constraint, such that they can experiment in exactly K sites. Further, K-sparse solutions to ℓ_0 minimization problems can be shown to have better worst-case performance than approximate solutions based on ℓ_1 minimization (for instance, the LASSO) (Zhang et al., 2014). Recent developments in Mixed Integer Programming mean that it is now possible to implement Best Subsets with optimality guarantees with practical-length runtimes (Nesterov, 2004, 2012; Bertsimas et al., 2015b). Further, this improved performance makes it possible to develop *ensemble methods* based on Best Subsets, in which the base algorithm is implemented many times with updating based on empirical losses (Zhou, 2012; Littlestone and Warmuth, 1994).

I develop an ensemble method that uses Best Subsets as a base learner to the mean discrepancy and distributional discrepancy problems. This involves resampling covariates, and using Monte Carlo Cross-Validation with Weighted Majority voting to update site selections across iterations.

As an extension, I show how the researcher can incorporate prior information about site-level variances and site sizes into the decision, and provide options for the Researcher to include welfare weights, and prior information about treatment effects.

I show that the method has favourable empirical performance to alternative methods, in terms of both in-sample performance and out-of-sample performance, across several naturalistic simulation studies. In the main text, I present results based on Auerbach and Thachil (2018). Because this study involves a large universe of sites, it is possible to assess both the internal and external validity of our approach by evaluating the sites chosen by my and alternative methods on both an 'observed' and 'unobserved' set of sites. This gives us insight into the finite sample performance of our method, and its performance in a context in which realistic distribution shift is induced by differences between unobserved and observed sites.

My method is applicable to a number of other practical problems in statistics and Political Science. First, it can be used to select units to enroll in an experiment. Second, it can be used to assign treatment assignments to a fixed set of individuals. Finally, we can also use this method for cluster randomization, when a Researcher seeks to assign treatment at the cluster-level, once sites have already been selected.

One moral for applied researchers is that there is a trade-off between optimizing for the PATE and the CATE: in general, one can choose sites that are more representative of the average, or more representative of the full distribution, but not necessarily both. I describe this as a trade-off between representativeness and generalizability: and note that this is likely to be of practical significance for applied researchers. This is a 'no free lunch' result in the style of (Wolpert and Macready, 1997).

1.2 Related Approaches

Several approaches to developing site selection methods have been proposed in the literature in econometrics, statistics, computer science, and political science.

Olea et al. (2024) is closet to my approach. They prove that k-medians is minmax optimal for reducing the regret of a site selection, and similarly use a mixed integer program to select sites that solve this problem.

Egami and Lee (2024) develop a method for site selection based on the Synthetic Control Method (Alberto Abadie and Hainmueller, 2010; Abadie et al., 2015; Xu, 2017; Sun et al., 2023). They choose sites that, up to weights, well-approximate sites that are not included in a given selection. This approach emphasizes choosing sites that define a convex hull, within which excluded sites are contained. I discuss the differences between my proposal in more detail and this paper in Appendix A.3.

Addanki et al. (2022) use the Gram-Schmidt Walk to find treatment assignments that balance covariates subject to a constraint on the sample size. This approach benefits from the statistical guarantees of the Gram-Schmidt process for solving discrepancy minimization problems, but does not produce vectors that select an exact number of sites for inclusion in an experiment. It was developed with application in the technology industry in mind, where the sample size may still be large.

Tipton (2013) proposes a method that uses cluster analysis to identify strata of sites. Given those strata, sites are then randomly sampled from each stratum. Gechter et al. (2024) develops a Bayesian method, in which a prior distribution over treatment effects based on pilot data, and a prior welfare function for aggregating estimated treatment effects, are assumed. This enables the Researcher to use Monte Carlo simulations to evaluate the expected welfare under different site selections: this simulated posterior then generates estimates of the best (from the perspective of ex post welfare) subset of sites in which to experiment.

My paper is relevant to an important literature in External Validity, generalization, and distribution shift. External Validity has emerged as a key concern in empirical research in the social sciences following important an critique from (Deaton and Cartwright, 2018). Shadish et al. (2002) describe representation (of a particular population) and generalization (to a particular target population) as the key goals of external validity. Slough and Tyson (2023) develop a formal framework for external validity based on Blackwell experiments, which offers an information-geometric formulation of the notion of a space of experiments

(Murray and Rice, 1993). Within their framework, I consider the problem of minimizing *target* discrepancy between selected settings S and a universe of settings P. Findley et al. (2016) and Hartman (2021) provide formal frameworks for thinking about external validity in applied research contexts. Our goal is to consider the set of sites that are C-valid based on observables. Like Hartman (2021), we require a contextual exclusion restriction, or the assumption of no unobserved moderators, in order successfully choose sites.

I also contribute to a literature on ensemble methods in political science. Montgomery et al. (2012) introduced ensemble methods to political science by motivating Ensemble Bayesian Model Averaging. Grimmer et al. (2017, 2021) provide overview of ensemble methods in political science, and Samii et al. (2016) applies the procedure to analysing the results of an experiment to reduce recidivism in Colombia More recently, ensemble methods have been used across causal inference to learn heterogeneous treatment effects, with Random Forests, the SuperLearner approach, and the X-Learner being three recent examples of popular methodologies used by social scientists to learn heterogeneous causal effects (Wager and Athey, 2018; Athey et al., 2019; van der Laan et al., 2007; van der Laan and Petersen, 2007; van der Laan and Rose, 2011; Künzel et al., 2019).

1.3 Outline of Paper

In Section 2, I describe the problem of purposive site selection, and motivate our optimization approach for minimizing MSE with respect to the PATE and CATE. In Section 3, I derive the objective function that minimizes the MSE with respect to the PATE, and note that it is a support-constrained k-means problem. In Section 4, derive an approximation to the objective function that minimizes the MSE with respect to the CATE, and note that it is a support-constrained k-quantiles problem, or Quantile Discrepancy Minimization problem.

In Section 5, I describe how to use recent advances theory and implementation of Best Subsets, and show how we can apply the method to the problem of site selection. In Section 6, I introduce an ensemble method based on Best Subsets, in which we aggregate site selections by weighted majority voting. In Section 7, we develop a naturalistic simulation study based on Auerbach and Thachil (2018). In Section 8, I extend the method to two cases where we have prior information is available: the first, we have additional information about pretreatment outcomes from a pilot study; the second, where we have information about site-level variances. I also show how the user can incorporate welfare weights into the analysis. Section 9 concludes.

2 An Optimization-based Approach to Site Selection

2.1 Setup

We consider a Researcher choosing a subset of sites S from a well-defined finite population of sites $\mathcal{P}$. The Researcher faces a budget constraint K: they can select up to K sites.[1] The Researcher can choose any set S from the set of subsets of $\mathcal{P}$ that contain at most K elements: we write this set as Σ_K.

The Researcher must choose a proper subset $S \subset \mathcal{P}$ of sites, subject to an ℓ_0-norm, or *cardinality constraint* on S.[2] That is, we have that, for some $0 < K < ||P||_0$:

$$||S||_0 = \sum_{s=1}^{||P||_0} \mathbb{I}\{s_s \in S\} \leq K \tag{1}$$

[1] It is without loss of generality whether the Researcher's budget constraint is that they must choose exactly K sites, or up to K sites: all theory and algorithms apply to both cases.

[2] $||\cdot||_0$ is the ℓ_0 norm, so that $||S||_0 = \#\{s : s \in S\} = \sum_{s \in \mathcal{P}} \mathbb{I}\{s \in S\}$

Suppose that the Researcher is interested in either the Population Average Treatment Effect (PATE) or the Conditional Average Treatment Effect (CATE) over the full population[3]:

Definition 1 (Causal Estimands)

$$PATE = \mathbb{E}_{s\sim\mathcal{P},i\sim N_s}[Y_{is}(1) - Y_{is}(0)] \qquad CATE = \mathbb{E}_{s\sim\mathcal{P},i\sim N_s}[Y_{is}(1) - Y_{is}(0)|X = x]$$

Once the experiment is conducted, the Researcher will observe only the sample analogues to these quantities. But *which* sample analogue the Researcher will observe will depend on which sites are chosen for experimentation.

Definition 2 (Sample Quantities)

$$SATE = \mathbb{E}_{s\sim S,i\sim N_s}[Y_{is}(1) - Y_{is}(0)] \qquad SCATE = \mathbb{E}_{s\sim S,i\sim N_s}[Y_{is}(1) - Y_{is}(0)|X = x]$$

The Researcher's goal is to minimize the loss between the observed estimate of the PATE or CATE in the *selected* sample and the true population quantity. We want to select sites that yield estimates 'as close as possible' to the population causal quantities.

2.2 Causal Minimands

The Researcher wishes to minimize a *loss* criterion $\ell : \Sigma_K \to \mathbb{R}$. We can associate with each site selection a loss. N_s is the number of units in each site s.

The Mean Squared Error, which is the ℓ^2 distance between the sample and population quantities, for each estimand are as follows:

Definition 3 (Causal Estimands)

$$MSE_{PATE}(S) = \mathbb{E}_{s\sim S}\left[\Big(\mathbb{E}_{s\sim\mathcal{P},i\sim N_s}[Y_{is}(1) - Y_{is}(0)] - \mathbb{E}_{i\sim N_s}[Y_{is}(1) - Y_{is}(0)|s \in S]\Big)^2\right]$$

$$MSE_{CATE}(S) = \int_{\mathcal{X}} \mathbb{E}_{s\sim S}\left[\Big(\mathbb{E}_{s\sim\mathcal{P},i\sim N_s}[Y_{is}(1) - Y_{is}(0)|\boldsymbol{X}_i = \boldsymbol{x}] - \mathbb{E}_{i\sim N_s}[Y_{is}(1) - Y_{is}(0)|\boldsymbol{X}_i = \boldsymbol{x}, s \in S]\Big)^2\right] d\boldsymbol{x}$$

The second quantity is closely related to the Precision of Estimated Heterogeneous Effect of Hill (2011); Shalit et al. (2017).

The goal is to use a site selection procedure to select a set S of cardinality at most K that minimizes these losses. That is, our site selection problems can be written as follows:

Definition 4 (Site Selection Problems)

$$\min_{S:\,||S||_0 \leq K} MSE_{PATE}(S) \qquad \min_{S:\,||S||_0 \leq K} MSE_{CATE}(S)$$

[3]In what follows, I subscript expectations with what is being averaged over: $i \sim N_s$ denotes that we are averaging over units N_S; $s \sim \mathcal{P}$ denotes that we are averaging over sites in the population; $s \sim S$ that we are averaging over sites in the set S. This follows the convention in Roughgarden (2016)

These optimization problems are infeasible without further assumptions: we do not observe *either* potential outcome in sites we did not choose to experiment in. Further, randomization is minimax optimal for solving this problem *in the absence of any prior structural information* about the problem (Senn, 1994; Kallus, 2018).[4] This is related to the 'no free lunch' theorem of Wolpert and Macready (1997).

2.3 Using Covariates To Select Sites

We can, however, use covariates to construct feasible analogues of the above problems.

Suppose that Researcher observes site-level characteristics across multiple covariates for each potential site in the study. That is, we have a measurable space of covariates $\mathcal{X}$, which generates the covariates $\boldsymbol{X} = \{X_{js}\}_{j=1,s=1}^{J,P}$, where J is the dimension of covariates and P is the number of sites in the population $\mathcal{P}$. We also have a space of unobserved variables $\mathcal{U}$.

Under what conditions can the Researcher extrapolate about likely treatment effects, based on covariate information? We face the following restrictions.

First: covariates must explain some non-negligible amount of variation in treatment effects. That is, we require that covariates are *prognostic* (Hansen, 2008; Bicalho et al., 2022; Liao et al., 2024). Second, we require that covariates do not interact with unobserved confounders (Imai et al., 2008). Third, we require that potential outcomes do not vary in site specific ways: this is also called the contextual exclusion restriction (Egami and Hartman, 2023). We can formalize these requirements with the following assumptions:

Assumption 1 (Contextual exclusion restriction)

$$\forall i, s \neq s' : Y_i(Z = z, X = x, S = s) = Y_i(Z = z, X = x, S = s')$$

Assumption 2 (No unobserved interactions)
For unknown functions $\gamma : \mathcal{X} \to \mathbb{R}$, $\phi : \mathcal{U} \to \mathbb{R}$

$$\tau_{is} = \gamma(\boldsymbol{X}_i) + \phi(\boldsymbol{U}_i) + \epsilon_i$$

The functional form described above satisfies these criteria. In the ideal case, the variance explained by the collected covariates is large relative to the variance explained by uncollected covariates. We will now show how to construct feasible analogues to the site selection problems described above using covariate data.

3 Site Selection for the PATE

3.1 The Infeasible PATE Problem

Consider the problem of minimizing the expected MSE of the PATE:

Problem 1 (Researcher's Problem: PATE)

$$\min_{S} \quad \mathbb{E}_{s\sim S}\left[\left(\mathbb{E}_{s\sim\mathcal{P},i\sim N_s}[Y_{is}(1) - Y_{is}(0)] - \mathbb{E}_{i\sim N_s}[Y_{is}(1) - Y_{is}(0)|s \in S]\right)^2\right] \quad \text{s.t. } ||S||_0 \leq K$$

[4] A brief sketch of the argument of Theorem 1 in Kallus (2018) is as follows: suppose Nature adversarially chooses the distribution of site potential outcomes to maximize loss given some particular site selection. Then a completely random procedure is at least as good as any purposive selection procedure. See also Kasy (2016) and Eberhardt (2010).

We first show that, if we knew the true model of treatment effects, we would have a minimization problem for the PATE that is equivalent to choosing a set of sites S that minimizes a weighted difference of means between the covariates in the population and the covariates in the sample.

Proposition 1 (Infeasible site selection problem for the PATE)
Under Assumptions 1 and 2, the minimizer of:

$$\min_{S:||S||_0 \leq K} \quad \frac{1}{P}\sum_{s=1}^{P}\left(\mathbb{E}\Big[\gamma(\boldsymbol{X}^{\mathcal{P}})\Big] - \mathbb{E}\Big[\gamma(\boldsymbol{X}^{S})\Big]\right)^2$$

is the minimizer of Problem 1, conditional on unobserved covariates $\boldsymbol{U}$.

All proofs are contained in Appendix A.1.

The intuition is as follows. If we knew the function γ, we could write:

$$SATE = \mathbb{E}[\gamma(\boldsymbol{X}^{S})] \qquad PATE = \mathbb{E}[\gamma(\boldsymbol{X}^{\mathcal{P}})] \qquad PATE - SATE = \mathbb{E}[\gamma(\boldsymbol{X}^{\mathcal{P}})] - \mathbb{E}[\gamma(\boldsymbol{X}^{S})]$$

However, since we have only collected covariate data, we do not know the function γ. Thus, this minimization problem is *infeasible*: it depends on information we do not observe.

3.2 A Feasible PATE Problem

To make progress on this, consider plugging in an arbitrary function γ_0 in place of γ. By adding and subtracting terms, our expression for the MSE becomes:

$$\min_{S:||S||_0 \leq K} \quad \frac{1}{P}\sum_{s=1}^{P}\left\{\underbrace{\left(\mathbb{E}\Big[\gamma(\boldsymbol{X}^{\mathcal{P}})\Big] - \mathbb{E}\Big[\gamma_0(\boldsymbol{X}^{\mathcal{P}})\Big] + \mathbb{E}\Big[\gamma(\boldsymbol{X}^{S})\Big] - \mathbb{E}\Big[\gamma_0(\boldsymbol{X}^{S})\Big]\right)}_{\text{Bias due to } \gamma_0} + \underbrace{\left(\mathbb{E}\Big[\gamma_0(\boldsymbol{X}^{\mathcal{P}})\Big] - \mathbb{E}\Big[\gamma_0(\boldsymbol{X}^{S})\Big]\right)}_{\text{Shift in covariate distributions}}\right\}^2$$

This is analogous to a bias-variance decomposition. In practice, the first term is unobserved, since the oracle weights are unknown. This means that we cannot optimize over this quantity. The *feasible* optimization problem, given an arbitrary function γ_0 then becomes:

$$\min_{S:||S||_0 \leq K} \quad \frac{1}{P}\sum_{s=1}^{P}\left(\mathbb{E}[\gamma_0(\boldsymbol{X}^{\mathcal{P}})\Big] - \mathbb{E}\Big[\gamma_0(\boldsymbol{X}^{S})\Big]\right)^2 \tag{2}$$

3.3 Minimizing the Bayes Risk of a Site Selection for the PATE

How should we choose a function γ_0?

First note that we can interpret a particular choice of γ_0 as minimizing the Bayes Risk of the site selection procedure with respect to γ_0. That is, we *actually* solve the following Bayesian risk minimization problem:

$$\min_{S:||S||_0 \leq K} \int_{\gamma} MSE_{PATE}(S;\gamma,\boldsymbol{X})\ d\Lambda(\boldsymbol{\gamma})$$

By placing a prior Λ on $\boldsymbol{\gamma}$.

We consider the following choices of Λ. First, suppose we had data from a pilot study that allowed us to estimate the relationship $\mathbb{E}[Y(0)|X]$. Then, we could use as a plug-in estimate $\hat{\gamma}_0$. This is an empirical Bayes strategy (Efron, 2010). We study this case in 8.1.

Second, in the absence of any information about the outcome model, it is sensible to place a uniform, uninformative prior on the covariates $\boldsymbol{X}$. This can be achieved by replacing γ_0 with the identity function, which is equivalent to multiplying covariates by the weight vector $\mathbf{1}$. Plugging this in gives us the following feasible site selection problem for the PATE:

Problem 2 (Feasible Site Selection for the PATE)

$$\underset{S:||S||_0 \leq K}{min} \quad \frac{1}{P}\sum_{s=1}^{P}\Big(\mathbb{E}[\boldsymbol{X}^{\mathcal{P}}] - \mathbb{E}[\boldsymbol{X}^{S}]\Big)^2$$

This simply picks the subset of size K that minimizes the observed difference in means between the selected sites and the population.

The intuition here is that to choose sites that are best for the PATE, with no prior information about the moderating effects of covariates, our best bet is to find a set of sites that most closely matches the mean of the sites. This result is also discussed in Tipton and Mamakos (2023).

4 Site Selection for the CATE

We now consider the problem of selecting sites when heterogeneity is of interest. The argument in this section is somewhat involved, so I preview its main conclusions here.

First, I show that selecting sites for the CATE can be represented as a discrepancy minimization problem, in which we seek to chose a subset of sites S such that the distribution F^S is as similar as possible to the distribution of $F^{\mathcal{P}}$. Unlike in the previous section, we now seek to match all of the moments of F between sample an population. Heuristically, this is a different solution set: we now want a set that *fully represents* the population distribution, not just a set that represents the effect at the average. These two goals may be incompatible in practice: a set of sites that is representative of the full distribution may not be representative of the sites at the average, and vice versa.

I then show that we can solve this discrepancy minimization problem by minimizing the empirical Weighted Quantile Discrepancy (WQD) between the selected sites and the population (Fan et al., 2022). I show that minimizing the WQD^2 (the ℓ^2 WQD) is asymptotically equivalent to minimizing the Mean Squared Error of the CATE estimate as the quantile mesh over which the WQD is minimized becomes increasingly fine-grained.

Importantly, the only assumption we place on the population distribution is that it is P-Donsker, which is a very unrestrictive assumption in practice. This allows us to use empirical process theory to use a very general nonparametric strategy to minimize the discrepancy between the distribution in the population and the distribution of the selected subset (Kennedy, 2016; van der Vaart, 2000).

This motivates the objective functions which we shall use Best Subsets to solve in the next section.

4.1 A Feasible Site Selection Problem for the CATE

Consider the problem of minimizing the Mean Squared Error between treatment effects in the sample and in the population. Using our previous definition, we can write the Researcher's problem as follows:

Problem 3 (Researcher's Problem: CATE)

$$\underset{S:||S||_0 \leq K}{min} \quad \int_{\mathcal{X}} \mathbb{E}_{s\sim\mathcal{P}}\left[\Big(\mathbb{E}_{i\sim N_s}[Y_{is}(1) - Y_{is}(0)|\boldsymbol{X}_i = \boldsymbol{x}] - \mathbb{E}_{i\sim N_s}[Y_{is}(1) - Y_{is}(0)|\boldsymbol{X}_i = \boldsymbol{x}, s \in S]\Big)^2\right] d\boldsymbol{x}$$

Proposition 2 (Infeasible Site Selection for the CATE)
The infeasible minimizer of Problem 3 is the minimizer of:

$$\min_{S:||S||_0 \leq K} \quad \int_{\mathcal{X}} \left(\gamma(\boldsymbol{x}) \left[f^S(\boldsymbol{X} = \boldsymbol{x}) - f^{\mathcal{P}}(\boldsymbol{X} = \boldsymbol{x}) \right] \right)^2 d\boldsymbol{x}$$

We follow the same approach discussed in the previous section. The feasible minimization problem minimizes Bayes risk with respect to the degenerate prior $\gamma_0 = \mathbf{1}$ is then:

Problem 4 (Feasible Site Selection Problem for the CATE)

$$\min_{S:||S||_0 \leq K} \quad \int_{\mathcal{X}} [f^S(\boldsymbol{x}) - f^{\mathcal{P}}(\boldsymbol{x})]^2 \, d\boldsymbol{x}$$

This is a *discrepancy minimization problem*: here, we want to minimize the distance between two distributions, as opposed to two vectors of means (Banaszczyk, 1998; Bansal et al., 2017; Levy et al., 2017; Bansal et al., 2022; Chazelle, 2000).

4.2 Using the empirical Weighted Quantile Discrepancy to select sites

In this subsection, I prove that we can use the empirical weighted quantile discrepancy to solve the feasible site selection problem for the CATE. The idea is that, since empirical quantiles are good estimators of underlying quantiles for a wide set of distributions (Lemma 1), the resulting empirical estimate of the true approximation has error that disappears asymptotically as the number of quantiles increases (Lemma 2). This means that the minimizer of the empirical weighted quantile discrepancy is asymptotically equivalent to the minimizer of the MSE of the CATE Theorem 1.

First, define the Weighted Quantile Discrepancy (Fan et al., 2022). This is a nonparametric distance measure that depends on the sum of the ℓ^p-norm distance between the *quantiles* of the two distributions. Here, $\boldsymbol{\upsilon}$ is a set of user-specified weights[5], and Δ_m defines a subinterval $(m-1, m] \subset [0, 1]$ of length $1/M$:

Definition 5 (2-Weighted Quantile Discrepancy (WQD^2))

$$WQD^2[\mathcal{P}, S, \boldsymbol{\upsilon}] = \sum_{m=1}^{M} \upsilon_m ||Q_m^{\mathcal{P}}(x) - Q_m^S(x)||^2 \Delta_m$$

Where $\boldsymbol{\upsilon}$ is a weight vector.[6] Intuitively, this gives us a discretized, nonparametric approximation to the distance between two distributions. We can think of this as taking histograms of each distribution over equally-sized quantile meshes, and summing the differences in mass at each quantile (Scott, 1992).

Our next technical lemma states that we can think of the WQD^2 as a coarse approximation of the ℓ^2 risk described in Proposition 2.

As the resolution of the quantile mesh increases, the WQD^2 converges to the squared distance between distributions. Essentially, WQD^2 measures the distance between coarse 'histogram' approximations of the

[5] In Section 8.2, I show that suitable choices of $\boldsymbol{\upsilon}$ allows the researcher to assign greater weight to particular quantiles of the distribution, and under some assumptions, particular subgroups. This is useful if we care about the effects of a given policy in the lowest deciles of income, for instance.

[6] The idea is that user specified weights allow the researcher to estimate welfare effects: for instance, by choosing a rule like $\upsilon_t = 1$ if $t < \alpha$, the researcher can consider treatment effects for the bottom α^{th} quantiles of the distribution, assuming that covariates are appropriately coded, and some prior knowledge of the direction of the effects of moderators. In what follows, we take $\upsilon_t = \mathbf{1}_T$.

distributions – as those 'histograms' become increasingly accurate approximations of the true underlying distributions, so too the WQD^2 becomes a more accurate measure of the squared distance of distributions.

Lemma 1 (Quantile Approximation of the ℓ^2 Risk)

$$WQD^2[\mathcal{P}, S, \mathbf{1}] \to \int_{\mathcal{X}} [f^S(x) - f^{\mathcal{P}}(x)]^2 \, dx$$

.

The 2-Weighted Quantile Distance with uniform weights converges to the ℓ^2 risk as the quantile mesh length goes to zero.

This result tells us that the 2-Weighted Quantile Discrepancy is asymptotically equivalent to the ℓ-2 risk, as our quantile approximation of the CDF of $\boldsymbol{X}$ becomes increasingly fine-grained.

Now we want a guarantee on the behaviour of an empirical plug-in estimate of the WQD. This allows us to use the 2-WQD as the basis of our minimization procedure in the next section.

The empirical quantile function, $\hat{Q}_t(x) = \inf_x \{x : \hat{F}(x) \geq t\}$, is a natural plug-in estimator for the quantile function above. It is straightforward to show the convergence of empirical quantile functions to true quantile functions using arguments from empirical process theory, which is a powerful set of tools for demonstrating the convergence of (functionals of) empirical distribution functions (van der Vaart, 2000; Kennedy, 2016).

First, we require the following technical assumption:

Assumption 3
Suppose $F^{\mathcal{P}}$ belongs to a class $\mathcal{F}$ of measurable functions with finite bracketing integrals.[7] *Then, $F^{\mathcal{P}}$ is P-Donsker.*

This limits the set of possible distributions of the covariates to a large set of possible distributions. For instance, distributions belong to an exponential family, Lipschitz functions, and functions with finite Vapik-Chervonekis dimension are P-Donsker.

Lemma 2 (Weak Convergence of Empirical Weighted Quantile Discrepancy)
Let $\hat{Q}$ be the empirical quantile function, and consider the empirical Weighted Quantile Discrepancy $\widehat{WQD}^2 = \sum_{t=1}^{T} \upsilon_t ||\hat{Q}_t^{\mathcal{P}}(x) - \hat{Q}_t^S(x)||^2 \, \Delta_t$. Then, by 3, the empirical Weighted Quantile Discrepancy converges weakly to the Weighted Quantile Discrepancy.

This confirms the basic intuition that the empirical quantile function is a valid plug-in estimator for the true quantiles (and hence, for the Weighted Quantile Discrepancy).

We are now able to state our main result for this section, which is that the minimizer of the empirical Weighted Quantile Discrepancy with respect to a site selection of cardinality K weakly converges to the feasible minimizer of $MSE_{\text{CATE}}(S; X)$.

Theorem 1 (Minimizing the empirical WQD^2 minimizes MSE_{CATE})

$$\underset{S:||S||_0 \leq K}{arg\,min} \quad \lim_{\substack{||\Delta_m|| \to 0 \\ M \to \infty}} \widehat{WQD}^2[\mathcal{P}, S, \mathbf{1}] \;=\; \underset{S:||S||_0 \leq K}{arg\,min} \quad MSE_{CATE}(S; X)$$

Theorem 1 shows that choosing a subset that minimizes the empirical 2-Weighted Quantile Discrepancy is asymptotically equivalent to choosing a subset that minimizes the MSE of the CATE. Further, it does

[7] A function $f \in \mathcal{F}$ has a finite bracketing integral if...

so for a wide class of distribution functions: we require only that the distribution function is P-Donsker, which is an unrestrictive assumption.

This result motivates our use of the empirical 2-Weighted Quantile Discrepancy as a minimand in practice.

In Appendix A.4, I discuss the connection between the problem of minimizing the 2-Weighted Quantile Discrepancy and minimizing the Wasserstein Distance.

5 Using Best Subsets To Select Sites

The previous section considered *what* to minimize. In this section I describe *how* to minimize it.

I first briefly outline the Best Subset Selection Problem, its computational complexity, and recent work by Bertsimas et al. (2015b) that has made it possible to find solutions to the problem with provable optimality in a 'reasonable' computational time frame in the $n < p$ regime. I then show how to use Best Subsets to find solutions to the feasible minimization problems described in the previous section.

5.1 Best Subsets: A Brief History

Originally known as the variable selection problem, the classical Best Subset Selection Problem in statistics and machine learning is the problem of choosing a finite subset of predictors that best explain the outcome in a regression model (Hocking and Leslie, 1967; Beale et al., 1967; Breiman, 1995; Miller, 2002; James et al., 2021; Thompson, 2022). This is modelled as the optimization problem:

Problem 5 (Best Subset Selection Problem)

$$\min_{\beta} \quad \frac{1}{2}\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{J} X_{ij}\beta_j\right)^2 \qquad s.t. \quad ||\beta||_0 \leq K$$

Where J is the dimension of the covariates, K is an integer in $(0, \min\{n-1, J\})$. Best subsets has the attractive feature of imposing a K-*sparse* solution – that is, one in which exactly K coefficients are non-zero – on the regression problem.

Until recently, however, provably optimal solutions to this problem were generally computationally infeasible. This is because best subset selection is nonconvex and NP-hard (Natarajan, 1995), since there are generally $\binom{J}{K}$ models to search through. This grows combinatorically: a brute force approach to selecting 10 variables from 50 possible covariates would require evaluating empirical losses of $\approx 10^{10}$ models. The implementation `leaps`, based on a branch-and-bound technique, is infeasible for problems of size $p \geq 30$ (Furnival and Wilson, 1974). Greedy methods based on heuristics, such as stepwise search, are not guaranteed to find optimal solutions.

Best Subsets, however, is considered to have some attractive theoretical properties over LASSO and Ridge Regression, two close alternatives that represent convex relaxations of Problem 5: that is, where the the nonconvex cardinality constraint that β is K-sparse is replaced with a convex penality on the ℓ^1 or ℓ^2 norm of β. Zhang et al. (2014) shows that, under a restricted eigenvalue condition, Best Subsets has lower variance than LASSO; Bertsimas et al. (2015b) highlight the superior empirical performance of Best Subsets. Finally, optimality guarantees ensure consistency of the algorithm.

5.2 Best Subsets: Bertsimas et al.

Bertsimas et al. (2015b) use advances in Mixed Integer Optimization to develop an algorithm to implement best subsets with provably optimal solutions in the $N < P$ regime, and short runtimes. The empirical

performance of this method is studied by Hastie et al. (2020), who find that it generally outperforms the lasso in high signal-to-noise regimes. Two packages to implement best subsets using MIO solvers in R, `Best Subsets` and `rss` have been developed by Hastie et al. (2020) and Thompson (2022).

Let H_K be the *hard thresholding operator*, which sets all but the K largest elements of a vector to zero (Donoho and Johnstone, 1994). The hard-thresholding operator. For a vector $X = \{X_1, \ldots, X_n\}$ take the ordered vector $\{X_{(1)}, \ldots, X_{(n)}\}$. Then, for $K < n$, $H_K(X) = \{X_{(1)}, \ldots, X_{(K)}, 0, \ldots, 0\}$.

Then, the Algorithm of Bertsimas et al. (2015b) solves the following problem:

Problem 6

$$\begin{aligned} \min_{\beta} \quad & ||Y - X'\beta||^2 \\ s.t. \quad & ||\beta||_0 \leq K \\ & ||\beta||_\infty \leq M_U, ||\beta||_1 \leq M_\ell \\ & ||X'\beta||_\infty \leq M_U^{X\beta}, ||X'\beta||_1 \leq M_\ell^{X\beta} \end{aligned}$$

We note that the bounds $\{M_U, M_\ell, M_U^{X\beta}, M_\ell^{X\beta}\}$ do not affect the theoretical optimality of the procedure, since they can be estimated from data with provable optimality. It is also possible to write out the objective function without explicit reference to these bounds, though the above formulation is preferable for purposes of exposition.

Algorithm 1: Best Subsets: Bertsimas, King and Mazumder (2016)

function Best_Subsets (Y, X, K, λ, ϵ)
Input : Outcome Y, Matrix of predictors X, Subset size K, Learning rate λ, Tolerance ϵ
Output: Selected site indicators $\hat{S}$
Initialize β_1 such that $||\beta_1||_0 \leq K$ `# Choose initial vector of weights from warm-start.`
for $m \geq 1$ **do**
  Calculate $\mathcal{L}(\beta_m) = \sum_{i=1}^{n}(Y_i - X'\beta_m)^2$ `# Evaluate loss of weights.`
  Calculate $\beta_{m+1} \in H_k\left(\beta_m - \frac{1}{\lambda}\nabla\mathcal{L}(\beta_m)\right)$ `# Hard-threshold gradient update step.`
  if $\mathcal{L}(\beta_m) - \mathcal{L}(\beta_{m+1}) > \epsilon$ **then**
    Continue `# Continue until convergence threshold reached.`
  else
    $\hat{\mathrm{S}} \leftarrow [\mathbb{I}\{\beta_{m1} \neq 0\}, \ldots, \mathbb{I}\{\beta_{ms} \neq 0\}]$ `# When converged, store optimal support.`
  end
end
$\beta^* \leftarrow \underset{\beta:\,\mathrm{Support}(\beta)=\hat{S}}{arg\,min} ||Y - X'\beta||^2$ `# Solve original problem over selected support.`
Return β^*

Algorithm 1 is an example of *projected gradient descent* (Nesterov, 2004, 2012). It does gradient descent over coefficients, then *hard-thresholds* the losses: that is, it deletes all but the top-K coeffieicnts, or projects onto the set of K-sparse solutions.

Notice that the goal is to first learn the K-sparse support of β – that is, which covariates to include – then to solve the minimization problem, given the solution to which covariates to include. In the case of linear regression, the goal is to choose variables, then run a regression on the selected variables.

This algorithm converges to provably optimal solutions in the $n < p$ regime, and has guarantees on its suboptimality in the $p > n$ regime. Further detail on this method is contained in Section 9.5.

5.3 Using Best Subsets to Solve The Site Selection Problem

In order to use Best Subsets to solve the feasible site selection problems for the PATE and the CATE, we first need to write out an objective function that has the form of Problem 6.

Whereas Best Subsets is usually applied to a data structure that is $N \times J$, we here apply it to a data structure that is $J \times S$. That is, we treat sites as *predictors*, and statistics of the covariates as our outcome. The goal is to pick the K sites that best predict features of the covariate distribution.

In this section, we suppose that the observed set of sites is the true finite population of sites. We do this by picking sites that match statistics of the observed covariate data as closely as possible, for some set of weights with K-nonzero entries. The idea is to reweight sites so that they match statistics of the full observed population.

We use Algorithm 1 to solve a problem of the following form:

Problem 7 (Using Best Subsets To Select Sites)

$$\min_{\boldsymbol{w}} \sum_{j=1}^{J} \left(T(X_{js}) - \sum_{s=1}^{\mathcal{P}} X'_{js} w_s \right)^2 \quad s.t. \; ||\boldsymbol{w}||_0 \leq K$$

Where $T : \boldsymbol{X}^{J \times S} \to \mathbb{R}^J$ is a statistic that transforms the $J \times S$ matrix X into a real-valued vector of length J.

This allows us to select a K-sparse vector of weights that chooses a set of sites $\hat{S}$ that minimize the ℓ-2 distance between the selected subset and the statistic in the full sample. Recall that Algorithm 1 produces an active set $\hat{S}$, which tells us which covariate entries should be non-zero. Since we have 'rotated' our data structure, this set now tells us which *sites* should be given non-zero weight. In other words, a K-sparse solution to this site selection problem, it is therefore straightforward to extract the set of sites that 'best' predicted statistics of the covariates. These are the sites that solve the discrepancy minimization problems described above, up to a set of weights w.

Note that, in this setting, we are not actually interested in the vector of weights. We are instead just interested in which weights are non-zero: these tell us which sites to include.

5.4 Best Subsets for the PATE

To solve Problem 2, we want to minimize the distance between covariate vectors in the population and the selected sample at the average.

We implement the following estimation problem using Best Subsets:

$$\min_{w} \; \frac{1}{2} \sum_{j=1}^{J} (\bar{X}_j - X'_j \boldsymbol{w})^2 \qquad \text{s.t.} \quad ||\boldsymbol{w}||_0 \leq K \tag{3}$$

Note that we no longer explicitly constrain the cardinality of the set $\boldsymbol{S}$: instead, the weight vector w is forced to be K-sparse, which induces K-sparsity of the set $\boldsymbol{S}$. Second, the weights themselves are not meaningful in this problem: we are only interested in the variables that are in fact selected by the model.

While the above minimizes the distance at the mean, we can also implement the following minimization problem, using the median as a Huber-robust estimator of the mean (Huber, 1981):

$$\min_{w} \frac{1}{2}\sum_{j=1}^{J}(\text{median}(X_j) - X_j'\boldsymbol{w})^2 \qquad \text{s.t.} \quad ||\boldsymbol{w}||_0 \leq K \tag{4}$$

5.5 Best Subsets for the CATE

Minimizing the empirical WQD in practice entails minimizing the following objective function:

Problem 8 (Objective Function for the CATE)

$$\min_{w} \frac{1}{2}\sum_{t=1}^{T}\sum_{j=1}^{J}(\hat{Q}_t(X_j) - X_j'w)^2 \qquad s.t. \quad ||\boldsymbol{w}||_0 \leq K$$

This is recognizable as a K-quantiles problem (Hennig et al., 2019): the goal is to choose a set that minimizes the distance between each of the quantiles of the covariates in the selected subset and in the population.

A practical difficulty in implementing this is that we now have a *matrix* of outcomes $\hat{\boldsymbol{Q}}^{T\times J}$. This is the set of quantiles for each covariate over a quantile partition of size T. We can think of this as a J-dimensional histogram estimate of the distribution function of X. Our solution to this problem is to use Weighted Majority Voting, which we discuss in the next section.

5.6 Convergence of Algorithm 1

We want to show that Algorithm 1 does, in fact, yield the correct site selection.

The argument is a straightforward application of the theoretical optimality of Algorithm 1 to the site selection problems outlined in the previous sections. This is established in Theorem 3.1 of Bertsimas et al. (2015b). Since this result is not mine, I direct the reader to the proof in the original paper, and state the result below.

Theorem 2 (Bertsimas et al. (2015b), Theorem 3.1)
For an optimization problem:

$$\min_{\boldsymbol{w}} \ell(\boldsymbol{w}; Y) \quad s.t. \ ||\boldsymbol{w}||_0 \leq K$$

Where $\ell(\boldsymbol{w}) \geq 0$ is convex and has Lipschitz continuous gradient so that:

$$||\nabla\ell(\boldsymbol{w}; Y) - \nabla\ell(\tilde{\boldsymbol{w}}; Y)|| \leq \ L||\boldsymbol{w} - \tilde{\boldsymbol{w}}||$$

Defining a learning rate $\lambda > L$, and let $\boldsymbol{w}^*$ be a solution to the optimization problem above. Then, after B iterations, Algorithm 1 satisfies:

$$\min_{b\in B} ||\boldsymbol{w}^{(b+1)} - \boldsymbol{w}^{(b)}|| \leq \frac{2[\ell(\boldsymbol{w}^{(1)}) - \ell(\boldsymbol{w}^*)]}{M(\lambda - L)}$$

Where $\ell(\boldsymbol{w}^{(b)}) \searrow \ell(\boldsymbol{w}^*)$ as $b \to \infty$.

It is straightforward to apply this result to the objective functions described by Equation (4) and Problem 8 above.

Corollary 1
Take Y to be $T(X) = \bar{X}$, and $\hat{Q}_t(X)$, respectively. Then Algorithm 1 converges to the minimizers of Equation (4) and Problem 8 respectively.

6 Best Subsets Ensemble with Weighted Majority Voting

Our goal in this section is to develop an ensemble method with improved out-of-sample performance. Ensemble methods, in which the predictions from a base leaner are aggregated to form a final prediction, are a popular tool in machine learning and causal inference that are increasingly used in the social sciences (Zhou, 2012; Montgomery et al., 2012; Samii et al., 2016; Grimmer et al., 2017; Künzel et al., 2019; Grimmer et al., 2021; Athey et al., 2019).

6.1 Motivation: Improving Generalization

In the previous section, we assumed that we observed the full population of sites, and described a method to select sites that minimize an observed loss criterion over that population. In many possible applications, however, we observe only a subset of sites from a larger population. That is, we have the following nested structure of sets:

$$S \subset P^{\text{sub}} \subset \mathcal{P}$$

Where we observe a subpopulation P^{sub} from a larger unobserved population $\mathcal{P}$. This relates to the case where we observe a sample of sites from a given population, from which we must select a subset. In this case, we may be interested in selecting sites that are representative of the full, partially observed population, and not the observed subpopulation. This problem can be formalized as one of external validity, or *distribution shift*, where we wish to extrapolate from a distribution F^P to a distribution $F^{\tilde{P}}$

The classic best subsets problem was the focus of much early ensemble research, due to the computational intractability of finding exact solutions (Breiman, 1995, 1996). Here, we use ensemble methods as a form of implicit regularization: to prevent overfitting to observed data, and thereby to improve generalization performance. The approach is broadly inspired by the Weighted Majority Voting algorithm of Littlestone and Warmuth (1994).

We use Monte Carlo Cross-Validation to repeatedly fit Best Subsets models. The general idea is to generate many test-training sample splits, generate site selections for each training set, and evaluate them on our hold out testing set. We then take the observed empirical loss on the testing set, and use this to determine each submodel's contribution to the overall prediction of which sites to use.

The goal is to build robustness to endogenous distribution shifts observed in the data: instead of solving the site selection problem once, we induce distribution shifts by solving the site selection problem for subsets of the data, and then assessing the ability of that prediction to generalize to unseen data. At each iteration, we draw a new test-training split, in which we train the model on $\tau\%$ of the data and test it on $100 - \tau\%$. By default, we set $\tau = 90$.

6.2 Aggregation by Weighted Majority Voting

Formally, let $\hat{S}^{(b)}$ be the prediction of a Best Subsets procedure at an iteration $b \in B$, let $\beta^{(b)}$ be a weight vector. Let $\ell : \mathcal{X}, \theta \to \mathbb{R}$ be a loss function, $\hat{\ell}(X, \hat{S})$ be the observed empirical loss on our selected sites. Recall the hard-thresholding operator H_K as defined previously.

Our final site selection is a weighted ensemble of a site selections generated from individual iterations of Algorithm 1. That is, we have the aggregated site selection:

$$\hat{S} = H_K \left[\sum_{b \in B} \beta^{(b)} \hat{S}^{(b)} \right]$$

This is a Weighted Majority algorithm (Littlestone and Warmuth, 1994; Cesa-Bianchi and Lugosi, 2006).

Then 'voting' weights are defined as follows:

$$\beta_s^{(b)} = \begin{cases} e^{-\ell(X_{\text{test}}^{(b)}, \hat{S}^{(b)})} & \text{If site } s \text{ is selected at iteration } b \\ 0 & \text{If site } s \text{ is not selected at iteration } b \end{cases}$$

Which is a normalization of the empirical loss: larger losses mean that the contribution of that model to the eventual prediction is smaller. Weight is zero for any site that is not selected; and is larger when the empirical loss is smaller.

We use a loss function that balances: minimization of both the *overall discrepancy* between the test set and its selected subset; *and* on specific quantiles of the test set. This is intended to ensure good performance both overall and with respect to the specific quantile prediction task.

This motivates our loss function, for a given quantile t:

$$\hat{\ell}(X, \hat{S}) = \hat{W}_1[\hat{Q}_t(X^{\text{test}}), \hat{Q}_t(X^{\text{test}}(\hat{S}))]$$

Where $\hat{W}_1$ is the empirical Wasserstein distance.

6.3 Algorithms

We set $t = .5$ (i.e. the median) when selecting sites for the PATE, and iterate over a grid of quantiles when selecting sites for the CATE. We use the median in place of the mean, as it is a Huber-robust estimator of the mean (Huber, 1981). In this way, Algorithm 2 is a special case of Algorithm 3, evaluated at one quantile.

The CATE algorithm is a 'winner-of-winners' procedure: Each iteration generates a 'local' K-sparse solution, and then the performance of each solution is aggregated to yield a 'global' K-sparse solution.

Algorithm 2: Best Subsets Ensemble for the PATE

function BSE_PATE (X, $T(X) = \text{median}(X)$, K, B)

Input : Matrix of predictors X, Statistic $T(X)$, Subset size K, Iterations B

Output: Selected site indicators $\hat{S}$

Split X into B test-train folds.

for $b \in B$ **do**

$\hat{S}^{(b)} = \text{Best\_Subsets}(T(X^{(b)}_{\text{train}}), X^{(b)}_{\text{train}}, K)$ `# Apply Algorithm 1`

$\ell(X^{(b)}_{\text{test}}, \hat{S}^{(b)}) \leftarrow \hat{W}_1[\hat{Q}_{.5}(X^{\text{test}}), \hat{Q}_{.5}(X^{\text{test}}(\hat{S}))]$ `# Evaluate holdout loss`

$$\beta_s^{(b)} \leftarrow \begin{cases} e^{-\ell(X^{(b)}_{\text{test}}, \hat{S}^{(b)})} & s \in \hat{S}^{(b)} \\ 0 & o.w \end{cases}$$ `# Use loss to construct voting weight`

end

$\beta_s^{(b)} \leftarrow \dfrac{\beta_s^{(b)}}{(KB)^{-1} \sum_{s \in P} \sum_{b \in B} \beta_s^{(b)}}$ `# Normalize losses`

$\hat{\beta}^K \leftarrow H_K \left[\left\{ \sum_{b \in B} \beta_s^{(b)} \right\}_{s=1}^{S} \right]$ `# Aggregate votes, select` K `winners`

$\hat{S} \leftarrow \left(\mathbb{I}\{\hat{\beta}_s^K \neq 0\} \right)_{s=1}^{S}$ `# Select sites corresponding to` K `largest votes`

Return $\hat{S}$ `# Output site selection`

Algorithm 3: Best Subsets Ensemble for the CATE

function BSE_CATE $(X,\ T,\ K,\ B)$

Input : Matrix of site characteristics X, Vector of quantiles T, Subset size K, Iterations B

Output: Selected site indicators $\hat{S}$

Split X into B test-train folds.

for $b \in B$ **do**

  for $t \in T$ **do**

    $\hat{S}_t^{(b)} \leftarrow \text{Best\_Subsets}(X_{\text{train}}^{(b)},\ \hat{Q}_t(X),\ K)$

    $\ell(X_{\text{test}}^{(b)}, \hat{S}_t^{(b)}) \leftarrow \hat{W}_1[\hat{Q}_t(X^{\text{test}}), \hat{Q}_t(X^{\text{test}}(\hat{S}))]$

    $$\beta_{st}^{(b)} \leftarrow \begin{cases} e^{-\ell(X_{\text{test}}^{(b)}, \hat{S}^{(b)})} & s \in \hat{S}_t^{(b)} \\ 0 & o.w. \end{cases}$$

  end

  $$\beta_{st}^{(b)} \leftarrow \frac{\beta_{st}^{(b)}}{(KT)^{-1} \sum_{t \in T} \sum_{b \in B} \beta_{st}^{(b)}}$$

  $$\beta_s^{(b)} \leftarrow H_K \left[\left\{ \sum_{t \in T} \beta_{st}^{(b)} \right\}_{s=1}^{S} \right]$$

end

$$\beta_s^{(b)} \leftarrow \frac{\beta_s^{(b)}}{(KB)^{-1} \sum_{s \in P} \sum_{b \in B} \beta_s^{(b)}}$$

$$\hat{\beta}^K \leftarrow H_K \left[\left\{ \sum_{b \in B} \beta_s^{(b)} \right\}_{s=1}^{S} \right]$$

$$\hat{S} \leftarrow \left(\mathbb{I}\{\hat{\beta}_s^K \neq 0\} \right)_{s=1}^{S}$$

Return $\hat{S}$

6.4 Properties of Algorithms 2 and 3

Algorithms 2 and 3 are closely related to the 'Super Learner' approach of van der Laan et al. (2007); Grimmer et al. (2017). Cross-validation benefits from finite sample and asymptotic performance guarantees described in Dudoit and van der Laan (2005). The asymptotic consistency of weighted majority voting is studied in Berend and Kontorovich (2014); Cesa-Bianchi and Lugosi (2006). I study the empirical performance of Algorithms 2 and 3 in the next section.

7 Empirical Application: Auerbach and Thachil, 2018

The goal is to assess both in-sample performance and out-of-sample performance. That is, how well the selected sites describe the observed covariate data, under the assumption of no unobserved moderators; and how well the sites describe unobserved sites. Out-of-sample performance gives us some sense of how well the observed sites generalize to unseen examples.

Throughout, we compare the performance of our model selection tool to a the Synthetic Purposive Sampling (SPS) method proposed by Egami and Lee (2024). I also calculate an infeasible oracle estimate, based on unobserved, synthetic treatment effects, generated under a linear model.

7.1 Settlements: Auerbach and Thachil, 2018

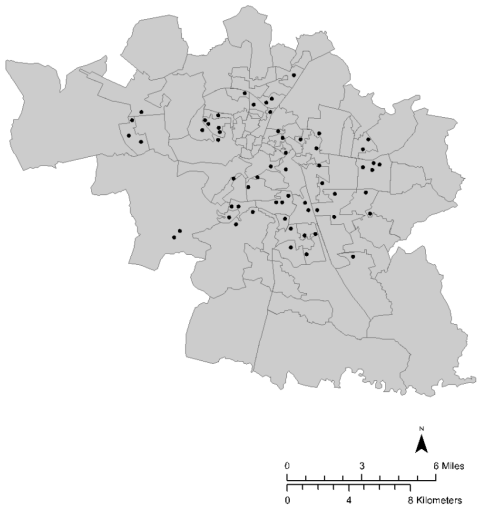

Figure 1: Bhopal

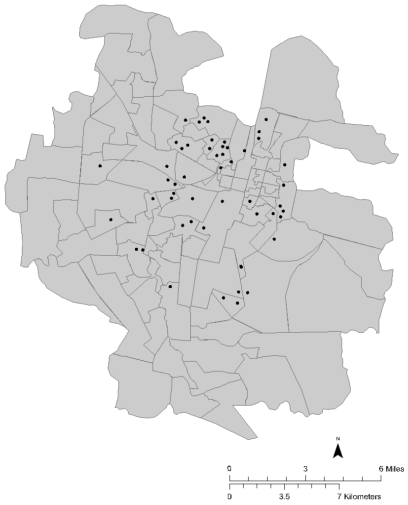

Figure 2: Jaipur

I design a naturalistic simulation study, based on Auerbach and Thachil (2018), who conducted a conjoint experiment across a sample of 110 informal urban settlements in Bhopal and Jaipur, India. Their goal was to study resident preferences for political brokers, whose role is to create linkages between political parties and local voters. A main finding from their paper is that residents were around 13.3% more likely to prefer a broker with a bachelor's degree, which Auerbach and Thachil interpret as a measure of a broker's capability to extract resources from the governing party.

This setting is a good test of our method for several reasons. First, the authors measured a variety of settlement-level covariates, some of which are predictive of the outcome. This allows us to generate synthetic treatment effects as a function of observed covariates to use as the 'ground truth' in our simulation. Second, because the authors conducted a study in 110 settlements, we have a large and well-defined study population. We can then partition this population into an observed subpopulation, and an unobserved larger population, to assess how well the sites selected in the subpopulation represent sites the unobserved larger population. This creates a naturalistic case study of distribution shift: covariate values will necessarily shift across selections into site. It is plausible that we observe both sampling uncertainty and distributional uncertainty across both contexts (Rothenhäusler and Bühlmann, 2023).

7.2 Simulation approach

The simulation is structured as follows:

1. Denote the set of all sites as our population, $\mathcal{P}$.
2. Generate ITEs using individual-level covariate data, and treat these ITEs as the ground truth. This gives us both the PATE and the CATE.

   For $i \in I$:

   3. Randomly sample a **subpopulation** of sites $\mathcal{P}^{(i)} \subset \mathcal{P}$. This is taken to be the population of interest, for which the analyst observes aggregated site-level covariate data.
   4. Use a *site selection method* to select a subset of K sites from the subpopulation $\mathcal{P}^{(i)}$
   5. a) **CATE loss**: Record *the empirical 1-Wasserstein distance* between:
      - *In-sample loss*: The (unobserved) distribution of ITEs in the subpopulation and the distribution of ITEs in the selected sample.
      - *Out-of-sample loss* The (unobserved) distribution of ITEs in the population and the distribution of ITEs in the selected sample.

      b) **PATE loss**: Record the empirical 1-Wasserstein distance between:
      - *In-sample loss*: The (unobserved) distribution of SATEs in the subpopulation and the distribution of SATEs in the selected sample.
      - *Out-of-sample loss* The (unobserved) distribution of SATEs in the population and the distribution of SATEs in the selected sample.

      c) **Oracle loss**: By brute force search, find the site selection that minimizes each of the above losses with respect to *the unobserved distribution of synthetic treatment effects*. This procedure is infeasible in general because treatment effects are not observed, units outside the population are not observed, *and* brute force search is computationally infeasible for larger sample sizes.
6. Aggregate losses across all iterations.

7.3 Results

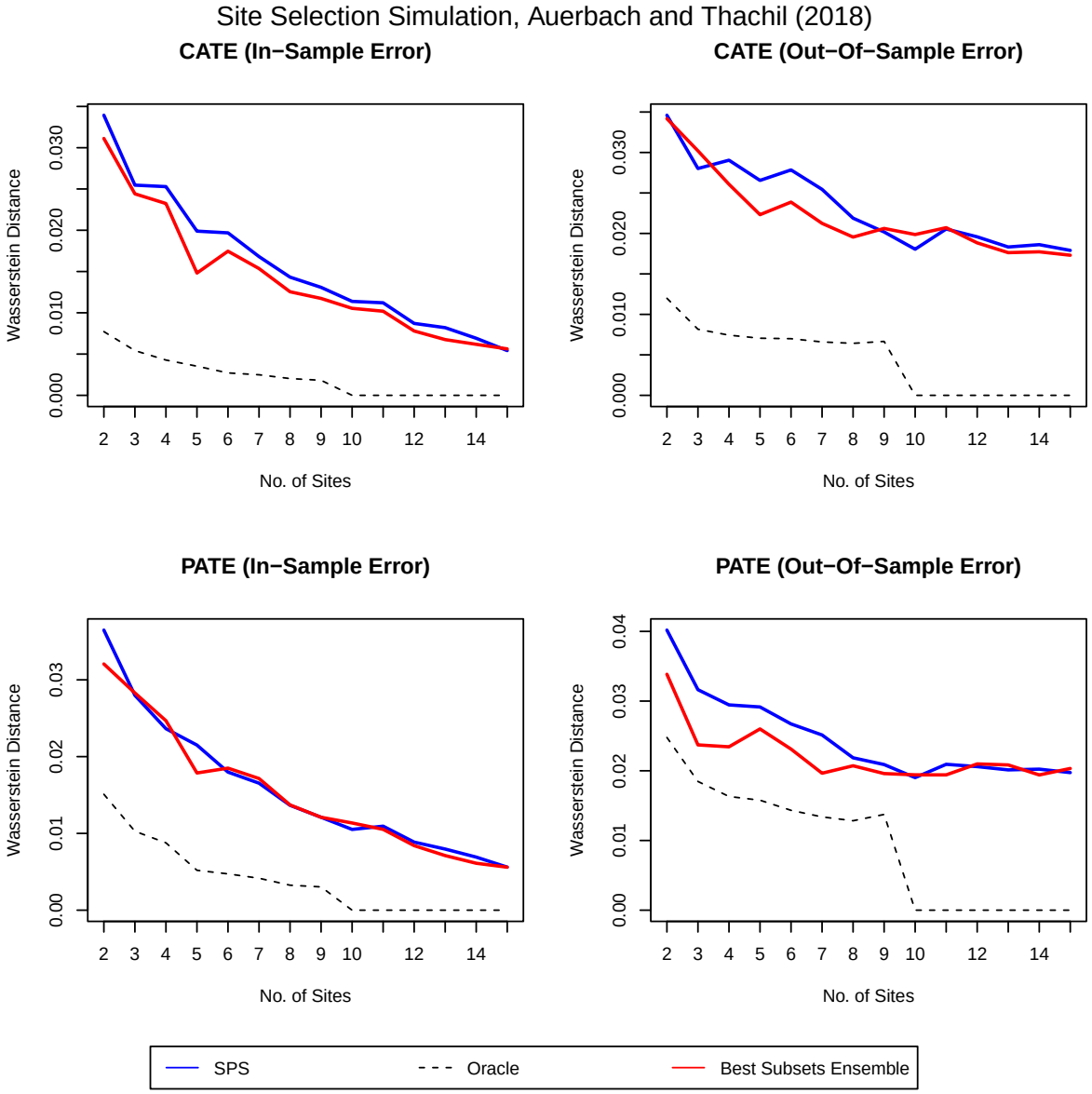

Table 1: Simulation Results, Auerbach and Thachil (2018)

	SPS	Best Subsets Ensemble
PATE (In-sample)	0.01576215	**0.01508495**
PATE (Out-of-Sample)	0.02469087	**0.02217600**
CATE (In-sample)	0.01573571	**0.01396278**
CATE (Out-of-Sample)	0.02332683	**0.02228081**

The proposed method has better average performance when estimating the PATE and the CATE; and

in both in-sample and out-of-sample contexts, compared to Synthetic Purposive Sampling.

In Appendix A.2, we assess the performance of the method using synthetic treatment effect data in Hill (2011) and Louizos et al. (2017). My method also outperforms SPS in average case performance in these simulation studies.

8 Extensions: Incorporating Prior Information

8.1 Prior Information About Pretreatment Outcomes

So far, we have assumed that Researchers do not observe outcome data. In the absence of any information about the model of treatment effects, we have minimized unweighted covariates.

We now consider the case where the Researcher observes pretreatment outcomes for a sample of units. That is, we suppose they observe a sample $\{Y_i(0), X_i\}_{i=1}^N$ drawn at random from the population. We assume that the Contextual Exclusion Restriction holds also for this sample, so that this sample contains no site-specific information.

From this sample, we can form a provisional estimate of the weights on X in the true outcome model:

$$\hat{\gamma} = \underset{\gamma}{arg\,min} \sum_{i=1}^{N} \left(Y_i(0) - X_i'\gamma\right)^2$$

We can then use these weights $\hat{\gamma}$ as a plug-in estimate of the weights in the *infeasible* site selection problems above.

This gives us two new optimization problems:

Problem 9 (Site Selection for the PATE with Pilot Data)

$$\underset{S\,:\,||S||_0 \leq K}{min} \mathbb{E}\left[\hat{\boldsymbol{\gamma}}(\boldsymbol{X}^S - \boldsymbol{X}^{\mathcal{P}})\right]$$

Problem 10 (Site Selection for the CATE with Pilot Data)

$$\underset{S\,:\,||S||_0 \leq K}{min} \int \left[f^S(\hat{\boldsymbol{\gamma}}'\boldsymbol{x}) - f^{\mathcal{P}}(\hat{\boldsymbol{\gamma}}'\boldsymbol{x})\right]^2 d\boldsymbol{x}$$

And their corresponding objective functions:

$$\underset{\boldsymbol{w}}{min} \sum_{j=1}^{J} \left(\hat{\gamma}_j X_j - (\hat{\gamma}_j X_j)'\boldsymbol{w}\right)^2 \quad \text{s.t. } ||\boldsymbol{w}||_0 \leq K$$

$$\underset{\boldsymbol{w}}{min} \sum_{t=1}^{T} \sum_{j=1}^{J} \left(\hat{Q}_t(\hat{\gamma}_j X_j) - (\hat{\gamma}_j X_j)'\boldsymbol{w}\right)^2 \quad \text{s.t. } ||\boldsymbol{w}||_0 \leq K$$

8.2 Prior Information About Welfare

We now consider the case where a Researcher is interested in selecting sites that are optimal for a *welfare-weighted* treatment effect. In particular, we consider the problem of using user-specified weights to diferentially weight quantiles of treatment effect moderators.

The user-weighted Quantile Discrepancy, for user weights $\boldsymbol{v} = \{v_{tj}\}_{(t=1)}^{(T)}$, is:

$$WQD^2(\mathcal{P}, S, \boldsymbol{\upsilon}) = \sum_{t=1}^{T}\sum_{j=1}^{J} \upsilon_t ||\hat{Q}_t(X_k) - X_j'\boldsymbol{w}||^2 \quad \text{s.t.} \quad ||\boldsymbol{w}||_0 \leq K$$

Plugging this into the objective function for the CATE, above, gives us the user-weighted minimization problem:

$$\min_{\boldsymbol{w}} \frac{1}{2}\sum_{t=1}^{T}\sum_{j=1}^{J} \upsilon_t \left(\hat{Q}_t(X_k) - X_j'\boldsymbol{w}\right)^2 \quad \text{s.t.} \quad ||\boldsymbol{w}||_0 \leq K$$

This is equivalent to solving the following minimization problem:

Problem 11 (Site Selection for the CATE with Welfare Weights)

$$\min_{S\,:\,||S||_0 \leq K} \int \left[\boldsymbol{\upsilon} \circ f^S(\boldsymbol{x}) - \boldsymbol{\upsilon} \circ f^{\mathcal{P}}(\boldsymbol{x})\right]^2 d\boldsymbol{x}$$

Where $\upsilon \circ f$ represents a transformation of the distribution of f.

8.2.1 Example: α-Lexicographic Welfare Weighting

Consider the following example. Let:

$$\upsilon_t = \begin{cases} 1 & \text{if } t < \alpha \\ 0 & \text{o.w.} \end{cases}$$

Then $\boldsymbol{\upsilon} \circ f$ is the *truncation* of f at the α^{th} quantile, and solving Problem 11 selects sites that minimize the discrepancy between the distributions at the lowest $\alpha\%$ of covariate values marginally.

Suppose individual pretreatment outcomes demonstrate increasing differences with respect to covariates (Milgrom and Shannon, 1994), so that:

Assumption 4 (Increasing pretreatment differences)

$$\boldsymbol{x}_i' \succeq \boldsymbol{x}_i \implies Y_i(0|\boldsymbol{x}') \geq Y_i(0|\boldsymbol{x})$$

This is a reasonable assumption in many policy evaluation contexts: for instance, we may think that pretreatment health status or education attainment increase monotonically in household socioeconomic characteristics.

Then, under increasing pretreatment differences, the α-truncated distribution represents the lowest α-quantiles of welfare in the population. Problem 11 then selects sites based on the units with lowest pre-treatment welfare.

Truncation of f represents weighting based on an α-lexicographic welfare criterion (Dickerson et al., 2014): that is, we find sites that minimize the site selection risk at the lowest α-quantiles of the distribution.

8.3 Prior Information About Variances: Shrinkage and Neyman-type Allocation

So far we have assumed that we observe only a matrix of covariate mean values in the population. We now consider the case where we also have some information about site-specific covariate variances. This could be the case where we have census information about a set of unit-level covariates, such that we believe this information to both prognostic with respect to treatment effects, and to vary within sites.

Neyman Allocation of groups is the procedure by which selection probabilities are proportional to the variance of the characteristic in the group (Neyman, 1934; Zhao, 2023). The idea is that we want more

weight assigned to groups with larger within-variances: they are more likely to include relevant population values, as well as comprising greater shares of the population as a whole.

We propose a Neyman-type procedure to incorporate information about variances into the above site selection problems.

Suppose we observe a matrix of population variances, and site sample sizes $\{n_s, \hat{\sigma}_{js}\}_{(j=1,s=1)}^{(J,\mathcal{P})}$. This information is new: we observed only site-level covariate means, previously, and standardized covariates across sites.

Define:

$$\hat{\zeta}_{js} = \frac{n_s \hat{\sigma}_{js}^2}{\sum_{s=1}^{\mathcal{P}} n_s \hat{\sigma}_{js}^2}$$

This "shrinkage factor" normalizes the variance of covariate j in site s relative to the variance of covariate j in all other sites s. For instance, if the variance in site s is much larger than the variance of all of other sites, $\hat{\zeta}_{js} \to 1$, while $\hat{\zeta}_{j\_s} \to 0$ for all other sites $\_s$.

This allows us to weight our objective function by the relative precision of estimates in site s. The idea is that sites with larger variance or larger sample sizes should be given more weight in our objective function: they are more likely to be internally diverse, which is useful for both PATE and CATE estimation.

We can rewrite our objective functions as follows:

$$\min_{\boldsymbol{w}} \quad \sum_{j=1}^{J} \left(T(\hat{\zeta}_{js} X_{js}) - \sum_{s=1}^{\mathcal{P}} (\hat{\zeta}_{js} X_{js})' w_s \right)^2 \quad \text{s.t. } ||\boldsymbol{w}||_0 \leq K \tag{5}$$

This is a *shrinkage estimator*, because it shrinks the weights on sites with smaller variances towards zero (Rosenman and Miratrix, 2022). As Tipton and Mamakos (2023) point out, in order to minimize variance with respect to the CATE, we want to both minimize the discrepancy of the two distributions, and select sites that are as heterogeneous as possible.

9 Discussion

9.1 Other Applications

While the above method is designed for the case of selecting a fixed number of experimental *sites* from a well-defined population, with studies in political science in mind, a number of other causal inference problems can be recast as discrepancy minimization problems where we want to select a subset that solves an optimization problem. We note the possibility of applying the above methods to three additional cases:

- **Individual-level data.** The most straightforward setting is one in which we wish to experiment on a subset of units. This setting is studied in Addanki et al. (2022). Here, we wish to minimize error with respect to included units i rather than sites s.
- **Treatment assignment** The problem of choosing two treatment assignment vectors that balance covariate distributions across two samples is a discrepancy minimization problem (Hainmueller, 2012; Ben-Michael et al., 2021).
- **Cluster Randomization** We can also use the shrinkage estimator proposed in the previous section to assign treatment incorporating both information about covariates and knowledge of cluster-specific variances.

9.2 When Should You Use This Method?

9.2.1 When Covariates Are Informative

If covariates are uninformative about potential outcomes, there is little point in using an optimization procedure to pick sites. Further, if relevant moderators of treatment effects are excluded

This is not a new point, as was stated by Bowley in 1926, in the earliest work on purposive sampling. Bowley wrote that the question of practical importance was "how far the precision of the measurements [here, of τ] is increased by correlation [of X and τ]".

9.2.2 When K Is Much Smaller Than P

The performance gap between randomization and optimization has been studied by (Bertsimas et al., 2015a; Kallus, 2018).

Randomization has worse performance when we must select a small number of units from a large potential population. When both S and K are large, there may be relatively little precision to be gained from purposive selection rather than random sampling, a point made by Neyman (1934). Random sampling guarantees representativeness asymptotically, so as K approaches S, and as S approaches ∞, the performance gap between random sampling and purposive site selection decreases.

9.2.3 When Between-Site Variance Is Large

When sites have high between-site variation and low within-site variation, randomization can perform this well. Here the 'risk' of a bad sample is larger, because the randomly selected sites can be further from the median; or closer together.

This is analogous to the argument is made by Pashley and Miratrix (2022) in the case of blocking: blocking improves inference, relative to complete randomization, when sites are heterogeneous, but have relatively little within-site variation.

9.3 Selective Inference

One concern raised at the use optimization procedures in experimental design is that the Researcher risks 'selecting on the dependent variable' (Geddes, 1990). Interest in selective inference has grown significantly over the last decade (Kuchibhotla et al., 2022), especially in the context of machine learning tools that reuse data to compare multiple models (Fithian et al., 2017).

The nature of the optimization objective matters, however. To allay selective inference concerns, what is important is that treatment assignment is independent of site selection. Because the experimental design proceeds in two stages: first select sites, *then* randomize units within site, the consistency of treatment effect estimation is independent of the site selection procedure. It is possible to have a purposive selection model *and* valid inference for treatment effect estimates. To put it another way, an optimization-based procedure can only make it more likely to discover a true treatment effect, given that there is in fact a treatment effect. It cannot make it more likely to discover a true treatment effect given that there is no treatment effect.

9.4 The Price of Robustness: Representativeness Versus Generalizability

Interestingly, site selection objectives for the CATE and for the PATE are distinct. For the PATE, we want to match sample means as closely as possible, while for the CATE, we want to match the whole distribution as closely as possible. Since these are different objectives, a solution that is optimal for one

is not necessarily optimal for the other; this is known as *the price of robustness* in Operations Research (Bertsimas and Sim, 2004).

This result has a clear substantive interpretation for the applied researcher: for the PATE, we want to find *bellweathers* – sites that are as representative as possible of the population at the average.

Note that this conclusion differs from conventional wisdom around study selection, in which it is suggested to diversify a selected sample as much as possible. Diversification may be useful if one is interested in the external validity or robustness of an estimate, or in learning about the CATE. But if the goal is to assess the PATE as accurately as possible, diversification may be worse from a Mean Squared Error perspective than selected sites closest to the average. This is an example of what is known in operations research as *the price of robustness* (Bertsimas and Sim, 2004). In the social sciences, this is sometimes colloquially referred to as 'the trade off between internal and external validity'.

9.5 Conclusions

I propose a set of novel methods for selecting experimental sites based on observable covariate data based around using Best Subsets. I apply these methods to selecting sites optimal for the PATE, and for the CATE. I develop an ensemble method intended to improve generalization performance when out-of-sample loss minimization is of interest. I show how to incorporate three kinds of prior information into the site selection problem of interest: pre-treatment outcomes, welfare weights, and variances.

I show that this method has better average case performance than existing alternative methods. I note that it can be applied to a range of other problems where the analyst seeks to solve a discrepancy minimization problem subject to a cardinality constraint, and describe the situations in which our method should be expected to perform better than randomization. Randomization is preferable when the number of selected sites is large, so that asymptotic approximations are likely to be good approximations. Purposive selection is preferable when we have prognostic covariates, can conduct experiments in only a small number of sites, and when sites have significant between-variance. Since this well describes common political science contexts, there is ample scope for adoption of this and related methods in cases where we can 'do better' than randomization.

References

Abadie, A., Diamond, A., and Hainmueller, J. (2015). Comparative politics and the synthetic control method. *American Journal of Political Science*, 59(2):495–510.

Addanki, R., Arbour, D., Mai, T., Musco, C., and Rao, A. (2022). Sample constrained treatment effect estimation.

Alberto Abadie, A. D. and Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505.

Arbour, D., Dimmery, D., and Rao, A. (2021). Efficient balanced treatment assignments for experimentation. In Banerjee, A. and Fukumizu, K., editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3070–3078. PMLR.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2):1148 – 1178.

Auerbach, A. and Thachil, T. (2018). How clients select brokers: Competition and choice in india's slums. *American Political Science Review*, 112(4):775–791.

Banaszczyk, W. (1998). Balancing vectors and gaussian measures of n-dimensional convex bodies. *Random Structures & Algorithms*, 12(4):351–360.

Bansal, N., Dadush, D., Garg, S., and Lovett, S. (2017). The gram-schmidt walk: A cure for the banaszczyk blues.

Bansal, N., Laddha, A., and Vempala, S. S. (2022). A unified approach to discrepancy minimization.

Beale, E. M. L., Kendall, M. G., and Mann, D. W. (1967). The discarding of variables in multivariate analysis. *Biometrika*, 54(3/4):357–366.

Ben-Michael, E., Feller, A., Hirshberg, D. A., and Zubizarreta, J. R. (2021). The balancing act in causal inference.

Berend, D. and Kontorovich, A. (2014). Consistency of weighted majority votes.

Bertsimas, D., Johnson, M., and Kallus, N. (2015a). The power of optimization over randomization in designing experiments involving small samples. *Operations Research*, 63(4):868–876.

Bertsimas, D., King, A., and Mazumder, R. (2015b). Best subset selection via a modern optimization lens.

Bertsimas, D. and Sim, M. (2004). The price of robustness. *Operations Research*, 52(1):35–53.

Bicalho, C., Bouyamourn, A., and Dunning, T. (2022). Conditional balance tests: Increasing sensitivity and specificity with prognostic covariates.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, 37(4):373–384.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):123–140.

Brooks-Gunn, J., ruey Liaw, F., and Klebanov, P. K. (1992). Effects of early intervention on cognitive function of low birth weight preterm infants. *The Journal of Pediatrics*, 120(3):350–359.

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, Learning, and Games*. Cambridge University Press.

Chazelle, B. (2000). *The Discrepancy Method: Randomness and Complexity*. Randomness and Complexity. Cambridge University Press.

Deaton, A. and Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science Medicine*, 210:2–21. Randomized Controlled Trials and Evidence-based Policy: A Multidisciplinary Dialogue.

Dickerson, J. P., Procaccia, A. D., and Sandholm, T. (2014). Price of fairness in kidney exchange. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems*, AAMAS '14, page 1013–1020, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81:425–455.

Dudoit, S. and van der Laan, M. J. (2005). Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Statistical Methodology*, 2(2):131–154.

Dunning, T., Grossman, G., Humphreys, M., Hyde, S. D., McIntosh, C., Nellis, G., Adida, C. L., Arias, E., Bicalho, C., Boas, T. C., Buntaine, M. T., Chauchard, S., Chowdhury, A., Gottlieb, J., Hidalgo, F. D., Holmlund, M., Jablonski, R., Kramon, E., Larreguy, H., Lierl, M., Marshall, J., McClendon, G., Melo, M. A., Nielson, D. L., Pickering, P. M., Platas, M. R., Querubín, P., Raffler, P., and Sircar, N. (2019). Voter information campaigns and political accountability: Cumulative findings from a preregistered meta-analysis of coordinated trials. *Science Advances*, 5(7):eaaw2612.

Eberhardt, F. (2010). Causal discovery as a game. In Guyon, I., Janzing, D., and Schölkopf, B., editors, *Proceedings of Workshop on Causality: Objectives and Assessment at NIPS 2008*, volume 6 of *Proceedings of Machine Learning Research*, pages 87–96, Whistler, Canada. PMLR.

Efron, B. (2010). *Large-Scale Hypothesis Testing*, page 15–29. Institute of Mathematical Statistics Monographs. Cambridge University Press.

Egami, N. and Hartman, E. (2023). Elements of external validity: Framework, design, and analysis. *American Political Science Review*, 117(3):1070–1088.

Egami, N. and Lee, D. D. I. (2024). Designing multi-site studies for external validity: Site selection via synthetic purposive sampling.

Fan, Z., Xu, Q., Jiang, C., and Ding, S. X. (2022). Weighted quantile discrepancy-based deep domain adaptation network for intelligent fault diagnosis. *Knowledge-Based Systems*, 240:108149.

Findley, M. G., Jensen, N. M., Malesky, E. J., and Pepinsky, T. B. (2016). Can results-free review reduce publication bias? the results and implications of a pilot study. *Comparative Political Studies*, 49(13):1667–1703.

Fithian, W., Sun, D., and Taylor, J. (2017). Optimal inference after model selection.

Furnival, G. M. and Wilson, R. W. (1974). Regressions by leaps and bounds. *Technometrics*, 16(4):499–511.

Gechter, M., Hirano, K., Lee, J., Mahmud, M., Mondal, O., Morduch, J., Ravindran, S., and Shonchoy, A. S. (2024). Selecting experimental sites for external validity.

Geddes, B. (1990). How the cases you choose affect the answers you get: Selection bias in comparative politics. *Political Analysis*, 2:131–150.

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

Grimmer, J., Roberts, M. E., and Stewart, B. M. (2021). Machine learning for social science: An agnostic approach. *Annual Review of Political Science*, 24(Volume 24, 2021):395–419.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46.

Hansen, B. B. (2008). The prognostic analogue of the propensity score. *Biometrika*, 95(2):481–488.

Hartman, E. (2021). *Generalizing Experimental Results*, page 385–410. Cambridge University Press.

Hastie, T., Tibshirani, R., and Tibshirani, R. (2020). Best Subset, Forward Stepwise or Lasso? Analysis and Recommendations Based on Extensive Comparisons. *Statistical Science*, 35(4):579 – 592.

Hennig, C., Viroli, C., and Anderlucci, L. (2019). Quantile-based clustering. *Electronic Journal of Statistics*, 13(2):4849 – 4883.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240.

Hocking, R. R. and Leslie, R. N. (1967). Selection of the best subset in regression analysis. *Technometrics*, 9(4):531–540.

Huber, P. J. (1981). *Robust Statistics*. Wiley.

Imai, K., King, G., and Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171(2):481–502.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2021). *An Introduction to Statistical Learning: with Applications in R*. Springer US.

Kallus, N. (2018). Optimal a priori balance in the design of controlled experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):85–112.

Kasy, M. (2016). Why experimenters might not always want to randomize, and what they could do instead. *Political Analysis*, 24(3):324–338.

Kennedy, E. H. (2016). Semiparametric theory and empirical processes in causal inference.

Kuchibhotla, A. K., Kolassa, J. E., and Kuffner, T. A. (2022). Post-selection inference. *Annual Review of Statistics and Its Application*, 9(1):505–527.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165.

Levy, A., Ramadas, H., and Rothvoss, T. (2017). Deterministic discrepancy minimization via the multiplicative weight update method.

Li, X. and Ding, P. (2016). General forms of finite population central limit theorems with applications to causal inference.

Liao, L. D., Zhu, Y., Ngo, A. L., Chehab, R. F., and Pimentel, S. D. (2024). Prioritizing variables for observational study design using the joint variable importance plot. *The American Statistician*, 78(3):318–326.

Littlestone, N. and Warmuth, M. (1994). The weighted majority algorithm. *Information and Computation*, 108(2):212–261.

Louizos, C., Shalit, U., Mooij, J., Sontag, D., Zemel, R., and Welling, M. (2017). Causal effect inference with deep latent-variable models. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6449–6459, Red Hook, NY, USA. Curran Associates Inc.

Milgrom, P. and Shannon, C. (1994). Monotone comparative statics. *Econometrica*, 62(1):157–180.

Miller, A. (2002). *Subset Selection in Regression*. Chapman and Hall/CRC.

Montgomery, J. M., Hollenbach, F. M., and Ward, M. D. (2012). Improving predictions using ensemble bayesian model averaging. *Political Analysis*, 20(3):271–291.

Murray, M. K. and Rice, J. W. (1993). *Differential geometry and statistics*. Chapman & Hall/CRC Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, Philadelphia, PA.

Natarajan, B. K. (1995). Sparse approximate solutions to linear systems. *SIAM J. Comput.*, 24:227–234.

Nesterov, Y. (2004). *Introductory Lectures on Convex Optimization*. Springer US.

Nesterov, Y. (2012). Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161.

Neyman, J. (1934). On the Two Different Aspects of the Representative Method : The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society*, 97(4):558–606.

Olea, J. L. M., Prallon, B., Qiu, C., Stoye, J., and Sun, Y. (2024). Externally valid selection of experimental sites via the k-median problem.

Pashley, N. E. and Miratrix, L. W. (2022). Block what you can, except when you shouldn't. *Journal of Educational and Behavioral Statistics*, 47(1):69–100.

Rosenman, E. T. R. and Miratrix, L. (2022). Designing experiments toward shrinkage estimation.

Rothenhäusler, D. and Bühlmann, P. (2023). Distributionally robust and generalizable inference.

Roughgarden, T. (2016). *Twenty Lectures on Algorithmic Game Theory*. Cambridge University Press, USA, 1st edition.

Samii, C., Paler, L., and Daly, S. Z. (2016). Retrospective causal inference with machine learning ensembles: An application to anti-recidivism policies in colombia.

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley.

Senn, S. (1994). Fisher's game with the devil. *Statistics in Medicine*, 13(3):217–230.

Shadish, W., Cook, T., and Campbell, D. (2002). *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Number v. 1 in Experimental and Quasi-experimental Designs for Generalized Causal Inference. Houghton Mifflin.

Shalit, U., Johansson, F. D., and Sontag, D. (2017). Estimating individual treatment effect: generalization bounds and algorithms.

Slough, T. and Tyson, S. A. (2023). External validity and meta-analysis. *American Journal of Political Science*, 67(2):440–455.

Sun, L., Ben-Michael, E., and Feller, A. (2023). Using multiple outcomes to improve the synthetic control method.

Thompson, R. (2022). Robust subset selection. *Computational Statistics amp; Data Analysis*, 169:107415.

Tipton, E. (2013). Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *Journal of Educational and Behavioral Statistics*, 38(3):239–266.

Tipton, E. and Mamakos, M. (2023). Designing randomized experiments to predict unit-specific treatment effects.

van der Laan, M. J. and Petersen, M. L. (2007). Causal effect models for realistic individualized treatment and intention to treat rules. *The International Journal of Biostatistics*, 3(1).

van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1).

van der Laan, M. J. and Rose, S. (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer New York.

van der Vaart, A. (2000). *Asymptotic Statistics*. Asymptotic Statistics. Cambridge University Press.

Villani, C. (2003). *Topics in Optimal Transportation*. American Mathematical Society.

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

Wolpert, D. and Macready, W. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82.

Xu, Y. (2017). Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis*, 25(1):57–76.

Zhang, Y., Wainwright, M. J., and Jordan, M. I. (2014). Lower bounds on the performance of polynomial-time algorithms for sparse linear regression.

Zhao, J. (2023). Adaptive neyman allocation.

Zhou, Z.-H. (2012). *Ensemble Methods: Foundations and Algorithms*. Chapman & Hall/CRC, 1st edition.

Appendix

A.1: Proofs of Results

Proof of Proposition 1

Proof.

$$
\begin{aligned}
&\min_{S:||S||_0 \leq K} \quad \mathbb{E}_{s\sim\mathcal{P}}\left[\Big(\mathbb{E}_{i\sim\mathcal{P}}[Y_{is}(1)-Y_{is}(0)] - \mathbb{E}_{i\sim\mathcal{P}}[Y_{is}(1)-Y_{is}(0)|s\in S]\Big)^2\right] \\
=&\min_{S:||S||_0 \leq K} \quad \mathbb{E}_{s\sim\mathcal{P}}\left[\Big(\mathbb{E}_{i\sim\mathcal{P}}[\tau_{is}] - \mathbb{E}_{i\sim\mathcal{P}}[\tau_{is}|s\in S]\Big)^2\right] \\
=&\min_{S:||S||_0 \leq K} \quad \frac{1}{P}\sum_{s=1}^{P}\Big(\mathbb{E}_{i\sim\mathcal{P}}[\tau_{is}] - \mathbb{E}_{i\sim\mathcal{P}}[\tau_{is}|s\in S]\Big)^2 \\
=&\min_{S:||S||_0 \leq K} \quad \frac{1}{P}\sum_{s=1}^{P}\Big(\mathbb{E}_{\mathcal{X}}\Big[\mathbb{E}_{i\sim\mathcal{P}}[\tau_{is}|\boldsymbol{X}=\boldsymbol{x}]\Big|\,\boldsymbol{X}_i=\boldsymbol{x}_i\Big] - \mathbb{E}_{\mathcal{X}}\Big[\mathbb{E}_{i\sim\mathcal{P}}[\tau_{is}|\boldsymbol{X}=\boldsymbol{x}, s\in S]\Big|\,\boldsymbol{X}_i=\boldsymbol{x}_i\Big]\Big)^2 \\
=&\min_{S:||S||_0 \leq K} \quad \frac{1}{P}\sum_{s=1}^{P}\left(\int\left[\int \tau_{is} f(\tau_{is}|\boldsymbol{X}=\boldsymbol{x})\,di\right] f^P(\boldsymbol{X}=\boldsymbol{x})\,d\boldsymbol{x} - \int\left[\int \tau_i f(\tau_i|\boldsymbol{X}=\boldsymbol{x})\,di\right] f^S(\boldsymbol{X}=\boldsymbol{x})\;d\boldsymbol{x}\right)^2 \\
=&\min_{S:||S||_0 \leq K} \quad \frac{1}{P}\sum_{s=1}^{P}\left(\int \mathbb{E}[\tau_{is}|\boldsymbol{X}=\boldsymbol{x}] f^P(\boldsymbol{X}=\boldsymbol{x})\,d\boldsymbol{x} - \int \mathbb{E}[\tau_{is}|\boldsymbol{X}=\boldsymbol{x}] f^S(\boldsymbol{X}=\boldsymbol{x})\;d\boldsymbol{x}\right)^2 \\
=&\min_{S:||S||_0 \leq K} \quad \frac{1}{P}\sum_{s=1}^{P}\left(\int \gamma(\boldsymbol{x}) f^P(\boldsymbol{X}=\boldsymbol{x})\,d\boldsymbol{x} - \int \gamma(\boldsymbol{x}) f^S(\boldsymbol{X}=\boldsymbol{x})\;d\boldsymbol{x}\right)^2 \\
=&\min_{S:||S||_0 \leq K} \quad \frac{1}{P}\sum_{s=1}^{P}\left(\mathbb{E}[\gamma(\boldsymbol{X}^{\mathcal{P}})\Big] - \mathbb{E}\Big[\gamma(\boldsymbol{X}^{S})\Big]\right)^2
\end{aligned}
$$

□

Proof of Proposition 2

Proof.

$$
\begin{aligned}
&\min_{S:||S||_0 \leq K} \quad \mathrm{MSE}_{\mathrm{CATE}}(S;X) \\
=&\min_{S:||S||_0 \leq K} \quad \int_X \Big(\mathbb{E}[\tau|X=x, s\in S] - \mathbb{E}[\tau|X=x, s\in\mathcal{P}]\Big)^2 dx \\
=&\min_{S:||S||_0 \leq K} \quad \int_X \Big(\tau^S(x) - \tau^{\mathcal{P}}(x)\Big)^2 dx \\
=&\min_{S:||S||_0 \leq K} \quad \int_X \Big(\tau(x) f^S(x) - \tau(x) f^{\mathcal{P}}(x)\Big)^2 dx \\
=&\min_{S:||S||_0 \leq K} \quad \int_X \Big(\tau(x)[f^S(x) - f^{\mathcal{P}}(x)]\Big)^2 dx
\end{aligned}
$$

□

Proof of Lemma 1

Proof. Consider the $WQD^2[\mathcal{P}, S]$ with uniform weights. We write the 'coarse' quantiles associated with a mesh $\mathcal{M}$ as $\tilde{Q}_m$. Then, for any quantile mesh with $||\Delta_m|| = 1/M$, we have the following:

$$\begin{aligned}\int_X [f^{\mathcal{P}}(x) - f^S(x)]^2\, dx &= \int_0^1 ||Q_t^P(x) - Q_t^S(x)||^2\, dt \\ &= \sum_{m=0}^{M} ||\tilde{Q}_m^{\mathcal{P}}(x) - \tilde{Q}_m^S(x)||^2\, \Delta_m + \left(\sum_{m=1}^{M} \int_{m-1}^{m} ||Q_t^{\mathcal{P}}(x) - Q_t^S(x)|| - ||\tilde{Q}_m^{\mathcal{P}}(x) - \tilde{Q}_m^S(x)||^2\, \Delta_m\, dt \right) \\ &= WQD^2[\mathcal{P}, S, \mathbf{1}] + \underbrace{\left(\sum_{m=1}^{M} \int_{m-1}^{m} ||Q_t^{\mathcal{P}}(x) - Q_t^S(x)||^2 - ||\tilde{Q}_m^{\mathcal{P}}(x) - \tilde{Q}_m^S(x)||^2\, \Delta_m\, dt \right)}_{\text{Quantization error / Approximation error}}\end{aligned}$$

Where:

$$\begin{aligned}\lim_{\substack{||\Delta_m|| \to 0 \\ M \to \infty}} &\left(\sum_{m=1}^{M} \int_{m-1}^{m} ||Q_t^{\mathcal{P}}(x) - Q_t^S(x)||^2 - ||\tilde{Q}_m^{\mathcal{P}}(x) - \tilde{Q}_m^S(x)||^2 \Delta_m\, dt \right) \\ &= \int_0^1 ||Q_t^{\mathcal{P}}(x) - Q_t^S(x)||^2 - ||Q_t^{\mathcal{P}}(x) - Q_t^S(x)||^2\, dt = 0\end{aligned}$$

So that:

$$\lim_{\substack{||\Delta_m|| \to 0 \\ M \to \infty}} WQD^2[\mathcal{P}, S, \mathbf{1}] = \int_X [f^S(x) - f^{\mathcal{P}}(x)]^2\, dx$$

. □

Proof of Lemma 2

Proof. Since $F^{\mathcal{P}}$ is P-Donsker, the sequence of empirical distribution function $\hat{F}_n^{\mathcal{P}}$ converges almost surely to the distribution function $F^{\mathcal{P}}$ by Donsker's Theorem. Lemma 21.2 in van der Vaart (2000) proves that that $\hat{F}_n^{\mathcal{P}}$ is weakly consistent for $F^{\mathcal{P}}$. $\hat{F}_n^{\mathcal{P}}$ is weakly consistent for $F^{\mathcal{P}}$ if and only if $(\hat{F}_n^{\mathcal{P}})^{-1}$ is weakly consistent for $(F^{\mathcal{P}})^{-1}$, which implies that $(\hat{F}_n^{\mathcal{P}})^{-1}$ is weakly consistent for $(F^{\mathcal{P}})^{-1}$, and hence that $\hat{Q}_t^{\mathcal{P}}(x)$ is weakly consistent for $Q_t^{\mathcal{P}}(x)$. □

Proof of Theorem 1

Proof. This collects the results of this section. We have:

$$\begin{aligned}\underset{S:||S||_0 \leq K}{arg\,min} \lim_{\substack{||\Delta_m|| \to 0 \\ M \to \infty}} \widehat{WQD}[\mathcal{P}, S, \mathbf{1}] &\xrightarrow{d} \underset{S:||S||_0 \leq K}{arg\,min} \lim_{\substack{||\Delta_m|| \to 0 \\ M \to \infty}} WQD^2[\mathcal{P}, S, \mathbf{1}] \\ &= \underset{S:||S||_0 \leq K}{arg\,min} \int \left[f^S(x) - f^{\mathcal{P}}(x) \right] dx \\ &= \underset{S:||S||_0 \leq K}{arg\,min}\ MSE_{\text{CATE}}(S; X)\end{aligned}$$

□

A.2: Additional Simulation Results

IHDP: Hill, 2011

Our second study uses data from the Infant Health and Development Program (IHDP), which has become a benchmark dataset for assessing the performance of models that estimate heterogeneous treatment effects (Brooks-Gunn et al., 1992; Hill, 2011).

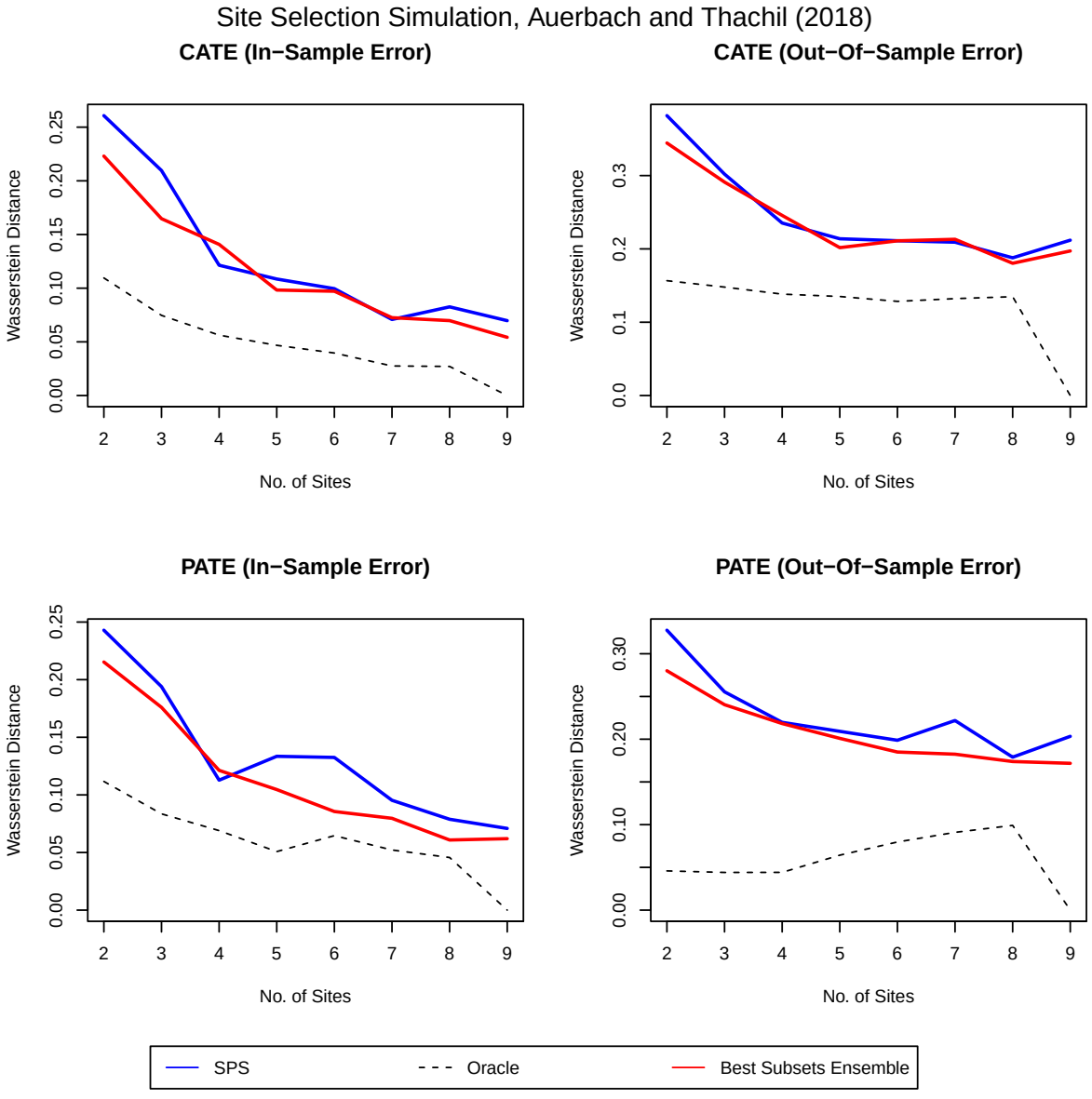

Table 2: Error (Wasserstein Distance), Hill (2011)

	SPS	Best Subsets Ensemble
PATE (In-sample)	0.1325840	**0.1131592**
PATE (Out-of-Sample)	0.2268514	**0.2065952**
CATE (In-sample)	0.1278788	**0.1150756**
CATE (Out-of-Sample)	0.2441207	**0.2355789**

Twins: Louizos, 2017

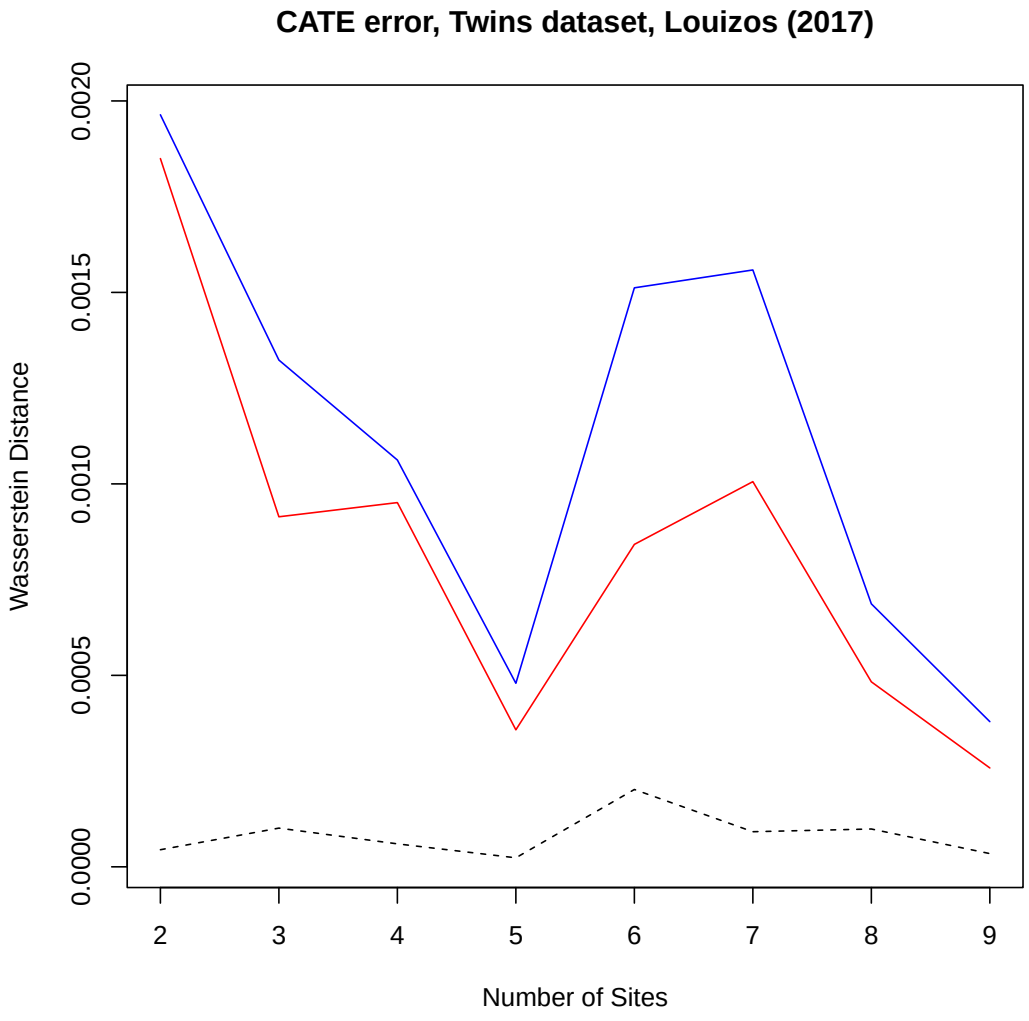

Figure 3: Twins simulation. Blue: SPS, Red: Best Subsets, Black: Oracle

Table 3: Test error (Wasserstein distance) on *Twins* dataset, Louizos (2017)

	SPS	*BestSubsetsEnsemble*
$K = 2$	0.0016831510	**0.0001232097**
$K = 3$	0.0014473130	**0.0006458922**
$K = 4$	0.0014986730	**0.0002082345**
$K = 5$	0.0002831783	**0.0001110586**
$K = 6$	0.0013974690	**0.0003896494**
$K = 7$	0.0023271380	**0.0012880440**
$K = 8$	0.0005380903	**0.0004933354**
$K = 9$	0.0003607147	**0.0001697653**
Mean	0.0011001370	**0.0004286486**

A.3: Comparison with Egami and Lee

My method differs from that in Egami and Lee (2024) in several ways.

First, I study the CATE, in addition to the PATE, which is a distinct problem and a main contribution of my paper.

Second, for the PATE, the goal of the method is slightly different. They consider the problem of choosing sites with improved external validity, which in practice means minimizing the probability that the PATE is not within the convex hull of observed site ATEs.

I study the problem of directly minimizing the MSE subject to a degenerate prior on the contribution of covariates to treatment effects. This is the optimal approach when covariates are fully informative. I am somewhat more agnostic about the challenge of external validity, and aim to explore in subsequent work how to incorporate tools from the distributional robustness literature into the site selection problem. My approach to improving out-of-sample performance is the ensemble method, which takes a 'winner-of-winners' approach to selecting sites, with the eventual site selection being one that is selected across multiple subsamples of covariates.

Second, the optimization problems differ: Egami and Lee (2024) propose a convex objective function with three components: i) minimizing ℓ^2 distance between unselected and the weighted combination of unselected sites ii) a penalty minimizing the distance between selected and unselected sites amd iii) a penalty term on the weights themselves.

I instead solve the k-means problem, by minimizing the distance between the weighted selection of sites and the mean vector of the covariates. The goal is to choose sites that well predict the mean of the covariates, rather than well-predicting the values of not included sites.

To illustrate the difference, imagine that sites vary across one dimension, and that we are interested in the PATE. The set of sites that are MSE-optimal are simply those closest to the PATE. But the set of sites that maximize the convex hull of the sites are the extremes, since the convex hull is the full line:

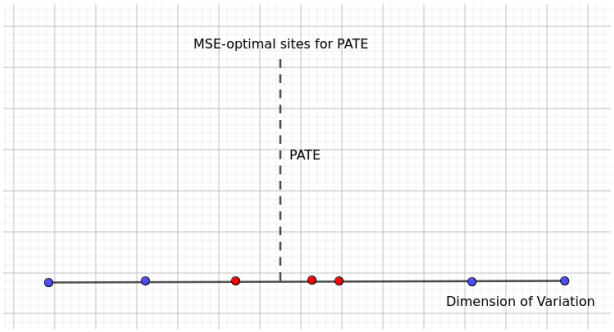

Figure 4: MSE-optimal

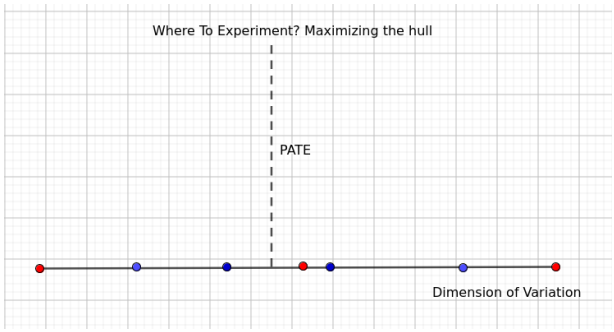

Figure 5: Hull-maximization

The point of the penalty term in Egami and Lee (2024) is to 'shrink' sites closer to the mean. Their objective function can therefore be understood as a convex combination of the two extremes presented in the two images, depending on choice of parameters.

My approach to improving external validity is to opt for the first method, and to induce distribution shift through resampling covariates, and combining the resulting models.

Thirdly, the implementation of my method benefits from the optimality guarantees, and fast runtimes of the method of Bertsimas et al. (2015b). In practice, Algorithm 1 has significantly faster runtimes than SPS.

A.4: Weighted Quantile Discrepancy and Optimal Transport

I briefly note a connection between minimizing the WQD and minimizing the Wasserstein Distance.

The Wasserstein Distance is frequently used in the optimal transport literature as a measure of the distance between two distributions: it has an intuitive interpretation as the amount of 'probability mass' that must be shifted from one distribution to another to make them equal (Villani, 2003).

It is defined as follows:

Definition 6 (p-Wasserstein Distance)

$$WD^p(\mathcal{P}, S) = \Big(\int_0^1 |F^{\mathcal{P}}(x) - F^S(x)|^p \, dx \Big)^{\frac{1}{p}}$$

We note that the 2-WQD is asymptotically equal to the 2-Wasserstein Distance.

Remark 1

The 2-Wasserstein Distance is defined $WD^2(\mathcal{P}, S) = \left(\int_0^1 |F^{\mathcal{P}}(x) - F^{\mathcal{P}}(x)|^2 \, dx \right)^{1/2}$.

So we also have that:

$$\underset{S:||S||_0 \leq K}{arg\,min} \lim_{\substack{||\Delta_m|| \to 0 \\ M \to \infty}} WQD^2[\mathcal{P}, S, \mathbf{1}] = \underset{S:||S||_0 \leq K}{arg\,min} \, WD^2(\mathcal{P}, S)$$

Which entails that, asymptotically, a minimizer of the empirical 2-WQD is a minimizer of the 2-Wasserstein Distance. In other words, if find a subset that minimizes the 2-WQD, we find a subset that solves the optimal transport problem of finding the subset S that minimizes the transport cost from F^S to $F^{\mathcal{P}}$.