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

Adam Zaid Austin Bouyamourn

UC Berkeley

Berkeley

Motivation: Where To Experiment?

Consider a Researcher who faces the following decision problem:

- Select some set of sites S from a universe $\mathcal{P}$, where $S \subset \mathcal{P}$.
- Run an experiment in each site to learn the PATE or CATE.

'**Double randomization**' may be undesirable in practice, as it controls bias and variance of population estimation quantity *in expectation*, but may have poor finite sample performance.

We may be able to do better by directly optimizing sites to minimize *discrepancy* between selected and unselected sites by optimization procedures (Kasy, 2016; Bansal et al., 2022; Harshaw et al., 2023; Egami and Lee, 2024).

What to Minimize?

Suppose a Researcher has a budget constraint: they can run experiments in exactly K sites.

For the PATE, we want to solve the following problem:

$$\begin{aligned}
&\min_{S:||S||_0 \leq K} \left(\mathbb{E}_{i\sim\mathcal{P}}[Y_i(1) - Y_i(0)] - \mathbb{E}_{i\sim\boldsymbol{S}}[Y_i(1) - Y_i(0)]\right)^2 \\
=&\min_{S:||S||_0 \leq K} \left(\mathbb{E}[\mathbb{E}_{i\sim\mathcal{P}}[Y_i(1) - Y_i(0)]|X = x] - \mathbb{E}[\mathbb{E}_{i\sim\boldsymbol{S}}[Y_i(1) - Y_i(0)]|X = x]\right)^2 \\
=&\min_{S:||S||_0 \leq K} \left(\int_X \int_i Y_i(1) - Y_i(0)|X = x \; d\left[f^{\mathcal{P}}(X = x) - f^{S}(X = x)\right]\right)^2
\end{aligned}$$

Since we have not yet run the experiment, the first term can be ignored, and we instead wish to minimize:

$$\min_{S} \int_X d[f^{\mathcal{P}}(X = x) - f^{S}(X = x)] \quad \text{s.t. } ||S||_0 \leq K$$

This is a *discrepancy minimization problem*, with an ℓ_0-norm constraint on S.

Weighted Quantile Discrepancy Minimization

We have the following approximation to this discrepancy minimization problem:

$$\begin{aligned}
\int f^{P}(X) - f^{\boldsymbol{S}}(X)\, dx &= \lim_{||\Delta_{X_t}|| \to 0} \sum_{t=1}^{T} [F^{P}(X) - F^{\boldsymbol{S}}(X)]\Delta_{X_t} \\
&= \sum_{t=1}^{T} [Q_t(X) - Q_t(X(\boldsymbol{S}))] \\
&\quad + \sum_{t=1}^{T} \int_{t-1}^{t} [F^{P}(X) - F^{\boldsymbol{S}}(X)] - [Q_t(X) - Q_t(X(\boldsymbol{S}))]\, dx \\
&\approx \sum_{t=1}^{T} [Q_t(X) - Q_t(X(\boldsymbol{S}))]
\end{aligned}$$

We can use the empirical analogues $\hat{Q}_t(X)$ above, and our problem then becomes to minimize the observed distance between the two empirical CDFs. This motivates our loss function, the Weighted Quantile Discrepancy (Fan et al., 2022):

$$WQD^2(X, \boldsymbol{X}(S), v) \equiv \sum_{t=1}^{T} v_t \left[Q_t(X) - Q_t(X(S))\right]^2$$

Best Subsets: A Modern Approach

Recall the classic Best Subsets problem:

$$\min_{S} ||Y - X'\beta||^2 \quad \text{s.t. } ||\beta||_0 \leq K$$

Best Subsets generates K-sparse solutions, but is nonconvex and NP-hard (Natarajan, 1995). Implementations of Best Subsets have either worked only for small sample sizes (`leaps`), or lacked optimality guarantees (stepwise selection).

Using Best Subsets To Select Sites

Bertsimas et al. (2015) use recent progress in Mixed Integer Optimization to develop a Best Subsets algorithm that generates provably optimal solutions for practical size problem instances with short computation time.

Whereas Best Subsets is typically used on data structures that are of dimension $\{I, I \times J\}$, we first generate a statistic of our observed covariates to use as our outcome vector, and 'rotate' our data, so that we consider a data matrix that is $\{J, J \times S\}$. That is, we use site information to predict functions of covariates.

Consider the following problem:

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

We iterate over a grid of quantiles, and choose a set of sites that minimizes the discrepancy over all quantiles.

An Ensemble: Subsampling and Weighted Majority Voting

Ensemble methods can be used to improve the generalization performance of a model (Zhou, 2012), by reducing overfitting. The goal is to simulate endogenous distribution shift, by splitting our observed data into smaller subsamples, which will have subsample moments contained within a small neighborhood of overall moments. In our setup, they can also be used to reduce the dependence of the model on specific covariates.

We use subsampling, an approach studied by Wager and Athey (2018). Define:

The *Hard-thresholding operator* $H_K(\cdot)$ (Donoho and Johnstone, 1994)

The *1-Wasserstein Distance*: $W_1(P, S) = \int_0^1 |F_{\mathcal{P}}^{-1}(x) - F_S^{-1}(x)|\, dx$

Algorithm: Ensemble Best Subsets for Site Selection

Input: Covariates X, Vector of quantiles T, Subset size K, Iterations B

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

For $b \in B$:

1. Split X into $X^{(b)}_{\text{test}}$, $X^{(b)}_{\text{train}}$

   For $t \in T$:

   2. $\hat{S}_t \leftarrow$ Best_Subsets($X^{(b)}_{\text{train}}$, $\hat{Q}_t(X^{(b)}_{\text{train}})$ K)
   3. $\ell(X^{(b)}_{\text{test}}, \hat{S}^{(b)}) \leftarrow W_1[X^{(b)}_{\text{test}}, X^{(b)}_{\text{test}}(\hat{\boldsymbol{S}})] + W_1[\hat{Q}_t(X^{(b)}_{\text{test}}), \hat{Q}_t(X^{(b)}_{\text{test}}(\hat{\boldsymbol{S}}))]$
   4. $\beta^{(b)}_{st} \leftarrow \begin{cases} 1 - \text{Softmax}_t\left(\sqrt{\dfrac{e^{\ell(X^{(b)}_{\text{test}}, \hat{S}^{(b)})}}{1 + e^{\ell(X^{(b)}_{\text{test}}, \hat{S}^{(b)})}}}\right) & s \in \hat{S}^{(b)}_t \\ 0 & o.w \end{cases}$

   end

end

5. $\hat{\beta}^K \leftarrow H_K\left[\left\{\sum_{b\in B} \beta^{(b)}_s\right\}_{s=1}^{S}\right]$
6. Return $\hat{S} \leftarrow \left(\mathbb{I}\{\hat{\beta}^K_s \neq 0\}\right)^S$

References

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

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

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

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.

Harshaw, C., Sävje, F., Spielman, D., and Zhang, P. (2023). Balancing covariates in randomized experiments with the gram-schmidt walk design.

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

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

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.

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

Application: Auerbach and Thachil, 2018

(a) Bhopal (b) Jaipur

Figure 1. 110 settlements studied in Auerbach and Thachil, 2018. We use these sites as a finite population to study the performance of our method. We partition the sites into observed and unobserved, run our method on the observed sites, and estimate in-sample and out-of-sample error, based on the Wasserstein distance between the (unobserved, synthetic) treatment effects in the selected sites and the treatment effects in the relevant comparison population.

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 $b \in B$:

   3. Randomly sample a **subpopulation** of sites $\mathcal{P}^{(b)} \subset \mathcal{P}$. 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 $\mathcal{P}^{(b)}$
   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 replications.

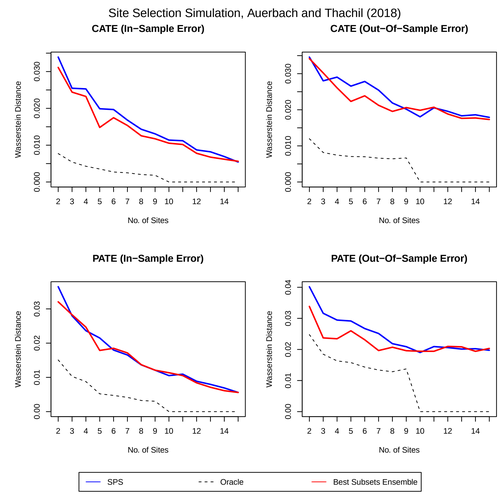

Table 1. Simulation Results, Auerbach and Thachil (2018)

	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**