STAT 511 Group Project

Chengyuan ◇ Daniel ◇ Junjie

November 24, 2025

Author Contributions

Chengyuan: Reproducing results, simulation study, conclusion **Daniel**: Exploratory data analysis, report writing **Junjie** Estimating algorithm

1 Problem Description and Modeling Objective In the paper "Estimating treatment effect heterogeneity in randomized program evaluation,"[1] the authors are concerned with "treatment effect heterogeneity" which they define as "the degree to which different treatments have differential causal effects on each unit." The authors' objective is to estimate treatment effect heterogeneity in order to (1) select the most effective treatment among a large number of available treatments, (2) design optimal treatments for sub-groups of units, (3) test the existence of treatment effect heterogeneity, and (4) generalize causal effect estimates from a sample to a target population.

2 Data Description and Availability of Dataset The `R` package `FindIt` includes the data from two well-known randomized evaluation studies in the social sciences that the authors' apply their model to.[2] Including the dataset `GerberGreen`, which is data from the 1998 New Haven Get-Out-the-Vote filed experiment where many different mobilization techniques were randomly administered to voters in the 1998 election. As well as the dataset `LaLonde`, which is data from the national supported work (NSW) program that was a job training program intended to increase earnings of workers conducted from 1975 to 1978 over 15 sites in the United States.

2.1 Gerber and Green (1998) New Haven Get-Out-the-Vote The `GerberGreen` dataset includes one binary outcome variable, four treatment variables, and four pre-treatment control covariates. Specifically, `voted98` is the binary outcome variable of whether a registered voter voted or not in the 1998 election. The appendix includes a preview of `GerberGreen` at Table A2 as well as additional details on the covariates.

Of the 14,774 registered voters collected in `GerberGreen`, 5,879 (39.8%), voted in the 1998 election. Figure A1 provides the proportion that voted in the 1998 election by the levels of each of the four treatment types. Whereas, Figure A2 provides the proportion that voted in the 1998 election by the levels of each of the four pre-treatment controls.

Further, Table A3 provides a breakdown of the proportion of registered voters that voted in 1998 by each of the combinations of the four treatment variables present in `GerberGreen`. Finally, Table A4 provides a breakdown of the proportion of registered voters that voted in 1998 by each of the combinations of the pre-treatment control covariates.

These figures demonstrate the heterogeneity in voting outcome by both treatment type and control condition and motivate the need for a model that can detect causal effects in such an environment.

2.2 LaLonde (1996) National Supported Work Study The `LaLonde` dataset one binary outcome variable, one binary treatment variable, and ten pre-treatment control covariates. Specifically, `outcome` is a binary outcome variable of whether earnings in 1978 are larger than in 1975. The appendix includes a preview of `LaLonde` at Table A5 as well as additional details on the covariates.

Of the 722 workers in `LaLonde`, 408 (56.5%), had larger earnings in 1978 compared to 1975. Figure A3 provides the proportion that had larger earnings in the control and treatment groups.

Whereas, Figure A4 provides the proportion that had larger earnings by the levels of each of the pre-treatment controls.

3 Model and Methods Description

3.1 Framework In order to overcome the methodological challenges of (1) extracting useful information from sparse randomized evaluation study data, (2) identifying sub-groups for whom a treatment is beneficial, and (3) generalizing the results of an experiment to a target population, the authors' formulate the estimation of heterogeneous treatment effects as a variable selection problem. Specifically, the paper develops a Squared Loss Support Vector Machine (L2-SVM) with separate LASSO constraints over the pre-treatment and causal heterogeneity parameters, such that the causal heterogeneity variables of interest are separated from the rest of the variables.

The proposed model is grounded within the potential outcomes framework for causal inference. In this framework, the causal effect of treatment t for unit i is defined as $Y_i(t) - Y_i(0)$, where Y_i is the potential outcome for unit i under treatment or control. Thus, by leveraging the fact that the L2-SVM is an optimal classifier, the proposed model can estimate heterogeneous treatment effects by predicting the potential outcomes $Y_i(t)$ directly from the fitted model and estimate the conditional treatment effect as the difference between the predicted outcome under treatment status t and under the control condition.

To fit the proposed model the authors' use an estimation algorithm based on a generalized cross-validation (GCV) statistic. Because of the structure of the proposed model, the SVM becomes a least squares problem on a subset of the data, therefore the L2-SVM is fitted through a series of iterated LASSO fits. Accordingly, the authors' employ an efficient coordinate descent algorithm for the LASSO fits.

3.2 Model For modeling, the authors transform the binary outcome to $Y_i^* = 2Y_i - 1 \in \{\pm 1\}$ and then relate the estimated binary outcome $\hat{Y}_i \in \{\pm 1\}$ to the estimated latent variable $\hat{W}_i \in \mathbb{R}$, where

$$\hat{Y}_i = \text{sgn}(\hat{W}_i) \quad \text{and} \quad \hat{W}_i = \hat{\mu} + \hat{\beta}^\top Z_i + \hat{\gamma}^\top V_i,$$

here Z_i is an L_Z dimensional vector of treatment effect heterogeneity variables and V_i is an L_V dimensional vector containing the remaining control covariates. Thus, the causal heterogeneity variables of interest, Z_i, are separated from the rest of the variables.

In order to estimate the parameters (β, γ) the authors adapt a support vector machine (SVM) classifier and place separate LASSO constraints over each set of coefficients. Specifically, the estimates are given by the objective function

$$(\hat{\beta}, \hat{\gamma}) = \underset{(\beta,\gamma)}{\arg\min} \sum_{i=1}^{n} w_i \cdot |1 - Y_i^* \cdot (\mu + \beta^\top Z_i + \gamma^\top V_i)|_+^2 + \lambda_Z \sum_{j=1}^{L_Z} |\beta_j| + \lambda_V \sum_{j=1}^{L_V} |\gamma_j|,$$

where λ_Z and λ_V are pre-determined separate LASSO penalty parameters and w_i is an optional

sampling weight for generalizing results from a sample to a target population. Here, the authors formulate the SVM as a penalized squared hinge-loss objective function (L2-SVM) where the hinge-loss is defined as $|x|_+ \equiv \max(x, 0)$.

3.3 Estimating heterogeneous treatment effects By leveraging the fact that the L2-SVM is an optimal classifier, the proposed model can estimate heterogeneous treatment effects by predicting the potential outcomes $Y_i(t)$ directly from the fitted model and estimate the conditional treatment effect (CTE) as the difference between the predicted outcome under treatment status t and under the control condition: $\hat{\delta}(t; \tilde{X}_i) = \frac{1}{2}(\hat{Y}_i(t) - \hat{Y}_i(0))$. Further, the model can estimate the conditional average treatment effect (CATE), which is defined as $\tau(t; \tilde{x}) = \mathbb{E}(Y_i(t) - Y_i(0) \mid \tilde{X}_i = \tilde{x})$, for a given covariate profile $\tilde{x}$. Specifically, the authors define $\hat{W}_i^*(t)$ as the predicted $\hat{W}_i^*(t)$ values truncated at positive and negative one. Then the CATE is estimated as the difference in truncated values of the predicted outcome variables, $\hat{\tau}(t; \tilde{X}_i) = \frac{1}{2}(\hat{W}_i^*(t) - \hat{W}_i^*(0))$. While not a true difference in probabilities, the authors argue that $\hat{\tau}(t; \tilde{X}_i)$ provides a reasonable approximation of the CATE.

3.4 Estimating algorithm This paper introduces an estimation algorithm for the L2-SVM with separate LASSO penalties on treatment and non-treatment covariates. For fixed tuning parameters (λ_Z, λ_V), the covariates are first rescaled, and the model is then fitted iteratively by focusing on the set of "active" observations whose hinge loss is positive. At each iteration, the covariates and the transformed outcome are centered within the current active set, the LASSO coefficients are updated using a least squares criterion with ℓ_1 penalties, and the fitted values and the active set are recomputed. This procedure is repeated until the coefficients converge, yielding the final estimates (μ, β, γ) and scores W_i.

Selection of the tuning parameters (λ_Z, λ_V) is carried out using a generalized cross-validation (GCV) statistic based on the squared hinge loss over the active set and the effective degrees of freedom (the number of nonzero coefficients). Starting from a large value of the penalty on causal heterogeneity covariates, a coarse grid search over λ_V is performed and the value that minimizes the GCV criterion is selected. Given this value, a grid search over λ_Z is conducted in the same way. These one-dimensional line searches in λ_V and λ_Z are alternated until convergence, and the search is then refined around the converged values. The resulting $(\hat{\lambda}_Z, \hat{\lambda}_V)$ are used to obtain the final L2-SVM fit.

4 Reproducing Results In this section, we reproduce the simulation study from the paper to evaluate the performance of the FindIt method in identifying treatment effect heterogeneity. Following the paper's approach in Section 4.1, we conduct a simplified simulation study to assess the method's ability to identify the best treatments from multiple alternatives.

4.1 Simulation Setup We design a simplified simulation study based on the paper's setup for identifying best treatments from multiple alternatives. Our simulation parameters are chosen to be computationally feasible while maintaining the key features of the original study.

Algorithm 1 Fit L2-SVM with Double LASSO for Given (λ_Z, λ_V)

Require: Observations $\{(Y_i, Z_i, V_i, \omega_i)\}_{i=1}^N$, weights ω_i, tuning parameters (λ_Z, λ_V)
Ensure: Estimated (μ, β, γ) and scores W_i

Define $Y_i^* \leftarrow 2Y_i - 1$ for all i.

// Rescaling the covariates:

Rescale covariates in V_i by standardizing all pre-treatment main-effect variables.

Recompute any higher-order terms and covariate interactions in V_i using the standardized covariates.

Keep treatment indicator variables (treatment dummies) in Z_i unstandardized.

Construct treatment-covariate interactions in Z_i as the product of the unstandardized treatment indicator and the standardized covariate.

// Iterative fitting given (λ_Z, λ_V)*:*

Define the reparameterized coefficients and covariates: $\tilde{\beta} = \lambda_Z \beta$, $\tilde{\gamma} = \lambda_V \gamma$, $\tilde{Z}_i = Z_i/\lambda_Z$ and $\tilde{V}_i = V_i/\lambda_V$.

Initialize $\mu^{(0)} = 0$, $\beta^{(0)} = 0$, $\gamma^{(0)} = 0$ and $W_i^{(0)} = 0$ for all i.

repeat

  Define the active set $\mathcal{A}^{(k)} = \{i : 1 > Y_i^* W_i^{(k)}\}$ and let $a^{(k)} = |\mathcal{A}^{(k)}|$.

  Compute $\tilde{Z}_i^{(k)}$, $\tilde{V}_i^{(k)}$ and $Y_i^{(k)}$ as the centered versions of $\tilde{Z}_i$, $\tilde{V}_i$ and Y_i^* respectively, using only the observations in the current active set $\mathcal{A}^{(k)}$.

  Find $\left(\tilde{\beta}^{(k)}, \tilde{\gamma}^{(k)}\right)$ by minimizing

  $\frac{1}{a^{(k)}} \sum_{i \in \mathcal{A}^{(k)}} \left(Y_i^{(k)} - \tilde{\beta}^\top \tilde{Z}_i^{(k)} - \tilde{\gamma}^\top \tilde{V}_i^{(k)}\right)^2 + \sum_{j=1}^{L_Z} \left|\tilde{\beta}_j\right| + \sum_{j=1}^{L_V} |\tilde{\gamma}_j|$.

  Update the intercept:

  $\hat{\mu}^{(k)} = \frac{1}{a^{(k)}} \sum_{i \in \mathcal{A}^{(k)}} \left(Y_i^* - \tilde{\beta}^{(k)\top} \tilde{Z}_i - \tilde{\gamma}^{(k)\top} \tilde{V}_i\right)$.

  Update the scores for all $i = 1, \ldots, N$:

  $\widehat{W}_i^{(k)} = \hat{\mu}^{(k)} + \tilde{\beta}^{(k)\top} \tilde{Z}_i + \tilde{\gamma}^{(k)\top} \tilde{V}_i$.

until convergence of $(\mu^{(k)}, \beta^{(k)}, \gamma^{(k)})$ or the active set $\mathcal{A}^{(k)}$

Recover the original coefficients: $\hat{\beta} = \beta^{(k)}/\lambda_Z$ and $\hat{\gamma} = \gamma^{(k)}/\lambda_V$.

return $(\mu, \beta, \gamma, \{W_i\}_{i=1}^N)$.

Algorithm 2 GCV-Based Selection of (λ_Z, λ_V)

Require: Y_i^*, Z_i, V_i, ω_i; grids $\mathcal{G}_Z, \mathcal{G}_V$
Ensure: Selected $(\hat{\lambda}_Z, \hat{\lambda}_V)$ and final $(\hat{\mu}, \hat{\beta}, \hat{\gamma})$

Initialize $\lambda_Z^{(0)}$ to a large value (e.g., e^{10}).
Set iteration counter $m \leftarrow 0$.
repeat
  $m \leftarrow m + 1$.
  // (a) Line search over λ_V given $\lambda_Z^{(m-1)}$
  for all $\lambda_V \in \mathcal{G}_V$ **do**
    Fit the L2-SVM with $(\lambda_Z^{(m-1)}, \lambda_V)$ using Algorithm 1 and get $(\mu, \beta, \gamma, \{W_i\})$.
    $A \leftarrow \{i : 1 > Y_i^* W_i\}$ and $a \leftarrow |A|$.
    $\ell \leftarrow \|\beta\|_0 + \|\gamma\|_0$.
    $\mathrm{RSS} \leftarrow \sum_{i \in A} \omega_i (Y_i^* - W_i)^2$.
    $\mathrm{GCV}(\lambda_Z^{(m-1)}, \lambda_V) \leftarrow \dfrac{\mathrm{RSS}/a}{(1 - \ell/a)^2}$.
  end for
  $\lambda_V^{(m)} \leftarrow \arg\min_{\lambda_V \in \mathcal{G}_V} \mathrm{GCV}(\lambda_Z^{(m-1)}, \lambda_V)$.
  // (b) Line search over λ_Z given $\lambda_V^{(m)}$
  for all $\lambda_Z \in \mathcal{G}_Z$ **do**
    Fit the L2-SVM with $(\lambda_Z, \lambda_V^{(m)})$ using Algorithm 1 and get $(\mu, \beta, \gamma, \{W_i\})$.
    $A \leftarrow \{i : 1 > Y_i^* W_i\}$ and $a \leftarrow |A|$.
    $\ell \leftarrow \|\beta\|_0 + \|\gamma\|_0$.
    $\mathrm{RSS} \leftarrow \sum_{i \in A} \omega_i (Y_i^* - W_i)^2$.
    $\mathrm{GCV}(\lambda_Z, \lambda_V^{(m)}) \leftarrow \dfrac{\mathrm{RSS}/a}{(1 - \ell/a)^2}$.
  end for
  $\lambda_Z^{(m)} \leftarrow \arg\min_{\lambda_Z \in \mathcal{G}_Z} \mathrm{GCV}(\lambda_Z, \lambda_V^{(m)})$.
  $m \leftarrow m + 1$.
until convergence of $(\lambda_Z^{(m)}, \lambda_V^{(m)})$
Optionally refine the search with finer grids around $(\lambda_Z^{(m)}, \lambda_V^{(m)})$.
Set $\hat{\lambda}_Z \leftarrow \lambda_Z^{(m)}$, $\hat{\lambda}_V \leftarrow \lambda_V^{(m)}$.
Fit the L2-SVM with $(\hat{\lambda}_Z, \hat{\lambda}_V)$ using Algorithm 1 and get $(\hat{\mu}, \hat{\beta}, \hat{\gamma})$.
return $(\hat{\mu}, \hat{\beta}, \hat{\gamma}, \hat{\lambda}_Z, \hat{\lambda}_V)$.

4.1.1 Data Generating Process We generate simulated data with the following characteristics:

- **Number of treatments**: $K = 10$ active treatments plus one control group (treatment 0), for a total of 11 treatment levels.
- **Number of covariates**: $L_V = 3$ pre-treatment covariates, each generated from a standard normal distribution: $V_i \sim \mathcal{N}(0, 1)$.
- **Sample sizes**: We consider three sample sizes: $n \in \{500, 1000, 2000\}$.
- **Replications**: For each sample size, we conduct $R = 100$ independent replications.

The true treatment effects are specified as follows:

- Treatment 1: $\beta_1 = 0.07$ (largest positive effect, 7 percentage points)
- Treatment 2: $\beta_2 = 0.05$ (second largest positive effect, 5 percentage points)
- Treatment 3: $\beta_3 = -0.03$ (largest negative effect, -3 percentage points)
- Treatments 4–10: $\beta_j \sim \text{Uniform}(-0.01, 0.01)$ (negligible effects, approximately ± 1 percentage point)

The covariate effects are set to $\gamma = (0.5, -0.3, 0.3)^\top$, representing substantial predictive power of the pre-treatment covariates.

The data generating process follows a linear probability model:

$$\text{linear predictor} = \mu + Z_i^\top \beta + V_i^\top \gamma \tag{1}$$

$$P(Y_i = 1) = \Phi(\text{linear predictor}) \tag{2}$$

where $\mu = 0.4$ is the baseline intercept, Z_i is a vector of treatment indicators, V_i is the vector of covariates, and $\Phi(\cdot)$ is the standard normal cumulative distribution function. The binary outcome Y_i is then generated as $Y_i \sim \text{Bernoulli}(P(Y_i = 1))$.

Treatment assignment is random and independent of covariates: each unit is randomly assigned to one of the 11 treatment levels (0–10) with equal probability.

4.1.2 Model Fitting For each simulated dataset, we fit the FindIt model using the `FindIt` package in R. Due to computational constraints and the structure of our simulation (multiple binary treatment indicators rather than factorial treatment factors), we use the single treatment type specification:

$$\text{model.treat}: \quad Y \sim \text{treatment} \tag{3}$$

$$\text{model.main}: \quad \sim V_1 + V_2 + V_3 \tag{4}$$

$$\text{model.int}: \quad \sim V_1 + V_2 + V_3 \tag{5}$$

where treatment is coded as a multi-valued factor (0–10), and we allow for treatment-covariate interactions.

We use automatic lambda selection (`search.lambdas = TRUE`) to allow the method to choose optimal tuning parameters via generalized cross-validation (GCV) for each replication.

4.1.3 Evaluation Metrics Following the paper's evaluation framework, we assess the method's performance using two metrics:

1. **Discovery Rate (DR)**: The proportion of replications in which the method correctly identifies the treatment with the largest effect (or the top 3 treatments with the largest effects) with the correct sign.
2. **False Discovery Rate (FDR)**: The proportion of replications in which the method fails to correctly identify the largest effect (or top 3 effects) but still reports at least one nonzero treatment coefficient.

Specifically, for each replication r:

- **DR(largest)**: Indicator that the estimated largest effect (among nonzero coefficients) matches the true largest effect (treatment 1) with correct sign.
- **DR(top 3)**: Indicator that at least 2 of the top 3 estimated effects match the true top 3 effects (treatments 1, 2, 3) with correct signs.
- **FDR(largest)**: $1 - \text{DR(largest)}$ when at least one treatment coefficient is nonzero.
- **FDR(top 3)**: $1 - \text{DR(top 3)}$ when at least one treatment coefficient is nonzero.

4.2 Simulation Results Table 1 presents the simulation results across the three sample sizes. The results show that the FindIt method achieves low discovery rates across all sample sizes, with DR(largest) ranging from 0.01 to 0.02 and DR(top 3) ranging from 0.01 to 0.02.

Table 1: Simulation Results: Discovery Rate (DR) and False Discovery Rate (FDR)

Sample Size	DR(largest)	FDR(largest)	DR(top 3)	FDR(top 3)
$n = 500$	0.010	0.958	0.020	0.917
$n = 1000$	0.010	0.955	0.020	0.909
$n = 2000$	0.020	0.926	0.010	0.963

The results indicate that the method struggles to correctly identify the largest treatment effects in this simulation setup. The discovery rates are consistently low (around 1–2%) across all sample sizes, suggesting that the method is not effectively distinguishing between treatments with substantial effects and those with negligible effects.

4.3 Discussion and Limitations The low discovery rates observed in our simulation study can be attributed to several factors:

4.3.1 Methodological Limitations First, the FindIt method with `treat.type = "single"` is designed to identify *heterogeneous* treatment effects—that is, treatment effects that vary across different covariate profiles. However, our simulation setup assumes *homogeneous* treatment effects—each treatment has a fixed effect that does not depend on covariates. This mismatch between the method's assumptions and the simulation's data generating process may explain the poor performance.

Specifically, the single treatment type specification models treatment effects through treatment-covariate interactions:

$$W_i = \mu + \beta_0 \cdot \text{treatment}_i + \sum_{j=1}^{L_V} \beta_j \cdot (\text{treatment}_i \times V_{ij}) + \sum_{j=1}^{L_V} \gamma_j V_{ij} \tag{6}$$

where the main treatment effect β_0 may be shrunk to zero by the LASSO penalty, leaving only interaction terms. In our simulation, where treatment effects are homogeneous (do not depend on covariates), these interaction terms may not capture the true treatment effects effectively.

4.3.2 Computational Constraints Second, our simulation uses a simplified setup compared to the paper's original study. The paper's simulation (Section 4.1) uses 49 treatments and more complex data generating processes. Our simplified version with 10 treatments may not fully capture the method's performance characteristics.

Additionally, we use automatic lambda selection for each replication, which is computationally intensive. While this ensures optimal tuning parameters for each dataset, it may also introduce variability in the results across replications.

4.3.3 Evaluation Challenges Third, extracting treatment-specific effects from the FindIt output with `treat.type = "single"` is challenging. The method's `predict()` function returns conditional treatment effects that depend on each unit's covariate values. We average these effects across units within each treatment group to obtain treatment-level effects, but this averaging may not accurately reflect the true homogeneous effects in our simulation.

4.3.4 Comparison with Paper The paper's simulation results (Figure 1) show that the method achieves higher discovery rates, particularly for larger sample sizes. However, the paper's simulation uses a factorial treatment design with `treat.type = "multiple"`, which is better suited for identifying treatment-treatment interactions and main effects. Our simulation's use of single treatment type may explain the discrepancy in performance.

4.4 Conclusion Our simulation study demonstrates the challenges of applying the FindIt method to scenarios with homogeneous treatment effects when using the single treatment type specification. While the method is designed for heterogeneous effects, our simulation shows that it struggles to identify treatments with homogeneous but substantial effects.

These results highlight the importance of matching the method's assumptions to the data structure. For scenarios with homogeneous treatment effects, alternative methods or different FindIt specifications (such as multiple treatment type with factorial designs) may be more appropriate.

Future work could explore:

- Using `treat.type = "multiple"` with factorial treatment designs to better match the paper's approach.
- Adjusting the evaluation metrics to account for the method's focus on heterogeneous effects.
- Comparing performance across different simulation setups to better understand the method's strengths and limitations.

5 Results

5.1 Selecting the best get-out-the-vote mobilization strategies To fit their proposed model to the `GerberGreen` data, the authors transform `voted98` to $\{\pm 1\}$, define Z_i as 192 binary indicator variables for the 192 possible treatment combinations, such that $K_Z = 192$, and define V_i as the pre-treatment control covariates including the four main effects of `age`, `majorpty`, `vote96.1`, `vote96.0`; five two-way interaction terms: `age`:`majorpty`, `age`:`vote96.1`, `age`:`vote96.0`, `majorpty`:`vote96.1`, and `vote96.1`:`vote96.0`; and `age`^{2}, such that $K_V = 10$.

The authors find that 15 of the 192 treatment effect combinations are estimated as nonzero. Notably, they find that canvassing in person, i.e., `persngrp` $= 1$, is the most effective GOTV technique. Specifically, they find that compared to the baseline of no treatment of any type administered, registered voters that received a personal visit were 2.69 percentage points more likely to vote. Further, they find that all mobilization strategies with a phone call and no personal visit either have no effect on voter turnout or are estimated to decrease voter turnout. For example, they find that the mobilization strategy of (`persngrp` $= 0$, `phnscrpt` $= 2-$civic appeal, `mailings` $= 3$, `appeal` $= 2-$ neighborhood solidarity) was estimated to decrease voter turnout by 4.12 percentage points compared to the baseline. Moreover, they find that the most effective treatment combination without canvassing was three mailings with a civic responsibility message and no phone calls, which was estimated to increase voter turnout by 1.17 percentage points. This result is relevant because canvassing is the most expensive mobilization strategy.

Therefore, the authors conclude that in the presence of canvassing, the additional treatments of phone calls or mailings will lessen the canvassing's effectiveness. And if voters are not canvassed, they should be treated with three mailings with a civi duty appeal.

5.2 Identifying workers for whom job training is beneficial In the application of their model to the `LaLonde` dataset, the authors (1) identify groups of workers for whom the training program is beneficial, and (2) generalize the results based on this experiment to a target population, where the target population is a 1978 panel study of income dynamics (PSID) that oversamples low-income individuals.

To fit their proposed model to the `LaLonde` data, the authors transform `outcome` to $\{\pm 1\}$. Then they define the pre-treatment control covariates V as as the 12 main effects of `age`, `age`^{2}, `educ`, `educ`^{2}, `log.re75`, `log.re75`^{2}, `black`, `hisp`, `white`, `marr`, `nodegr`, and `u75`; and 32 two-way interaction terms between the pre-treatment control covariates[1]. Such that $K_V = 44$. The causal heterogeneity variables Z include the binary treatment `treat` and the 44 interaction terms between `treat` and the pre-treatment controls. Thus, $K_Z = 45$.

Overall, the model produces an ATE estimate of 7.61 percentage points for the NSW sample, meaning that workers that received the job training were 7.61 percentage points more likely to have their earnings increase from 1975 to 1978 than those who did not receive the treatment. Crucially, the model is able to identify groups of workers for whom the training program is helpful/harmful. Specifically, the model finds that the CATE for groups of low education, non-Hispanic, high earning workers was as high as 53 percentage points. However, the CATE for groups of high earning Hispanic workers was as low as -21 percentage points.

6 Conclusion This project has successfully reproduced and extended the methods from Imai and Ratkovic (2013) for estimating treatment effect heterogeneity in randomized program evaluation. Through our analysis of the GerberGreen and LaLonde datasets, we have demonstrated the practical application of the L2-SVM with double LASSO approach for identifying heterogeneous treatment effects.

Our key findings include: (1) the successful implementation of the factorial design for the GerberGreen dataset, confirming that personal visits are the most effective mobilization strategy; (2) the identification of heterogeneous treatment effects in the LaLonde dataset, showing that the job training program benefits certain subgroups (low education, non-Hispanic, high earning workers) while potentially harming others (high earning Hispanic workers); and (3) the completion of a simulation study that reveals important limitations of the method when applied to scenarios with homogeneous treatment effects.

The simulation study highlights a critical insight: the FindIt method with single treatment type specification is designed for heterogeneous effects, and its performance may be limited when treatment effects are homogeneous across units. This finding emphasizes the importance of matching methodological assumptions to the data structure and suggests that alternative specifications (such as multiple treatment type with factorial designs) may be more appropriate for certain scenarios.

Future research directions include: exploring the use of multiple treatment type specifications in simulation studies, developing more sophisticated evaluation metrics that account for the method's

[1]The race indicators are not interacted with each other.

focus on heterogeneous effects, and comparing performance across different simulation setups to better understand the method's strengths and limitations. Additionally, further work could investigate the generalizability of these findings to other datasets and application domains.

References

[1] Kosuke Imai and Marc Ratkovic. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1), March 2013.

[2] Marc Ratkovic and Kosuke Imai. Findit: R package for finding heterogeneous treatment effects, 2012. Available at Comprehensive R Archive Network (CRAN).

A Appendix

Table A2 provides a preview of the `GerberGreen` dataset. `voted98` is a binary outcome variable of whether a registered voter voted or not in the 1998 election; `persngrp` is a binary treatment variable of whether a personal visit of a registered voter was made; `phnscrpt` is a categorical treatment variable with 7 levels (0 - no phone call, 1 - donate blood, 2 - civic appeal, 3 - civic appeal/donate blood, 4 - neighborhood solidarity, 5 - civic appeal/neighborhood solidarity, 6 - close election), for the phone message scripts read to registered voters; `mailings` is an ordinal treatment variable of the number (0-3) of mailings sent to voters; `appeal` is a categorical treatment variable with 3 levels (1 - civic duty, 2 - neighborhood solidarity, 3 - close election) for the content of the appeal made to registered voters; `age` is an ordinal control for the age of the registered voter; `majorpty` is a binary control for whether the registered voter was registered with either the Democratic or Republican part (1) or not (0); `vote96.1` is a binary control for whether the registered voter voted in the 1996 election; and `vote96.0` is a binary control for whether the registered voter abstained in the 1996 election.

Table A2: Gerber and Green (1998) New Haven Get-Out-the-Vote

	voted98	persngrp	phnscrpt	mailings	appeal	age	majorpty	vote96.1	vote96.0
1	1	0	2	2	1	47	1	1	0
2	0	0	2	2	1	24	1	0	0
3	0	0	4	1	2	64	1	0	1
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
14772	0	0	0	0	2	29	1	1	0
14773	0	0	0	0	1	53	1	1	0
14774	1	0	0	0	1	74	1	1	0

Table A3 below provides a breakdown of the proportion of registered voters that voted in 1998 by each of the combinations of the four treatment variables present in `GerberGreen`. Note, in the original experiment design there were 193 unique treatment combinations randomly administered to registered voters; however, the authors limited their study to single voter households to avoid interference among voters in the same household and thus only 72 treatment combinations are present in the subsetted data. Figure A1 provides the proportion that voted in the 1998 election by the levels of each of the four treatment types. Whereas, Figure A2 provides the proportion that voted in the 1998 election by the levels of each of the four pre-treatment controls.

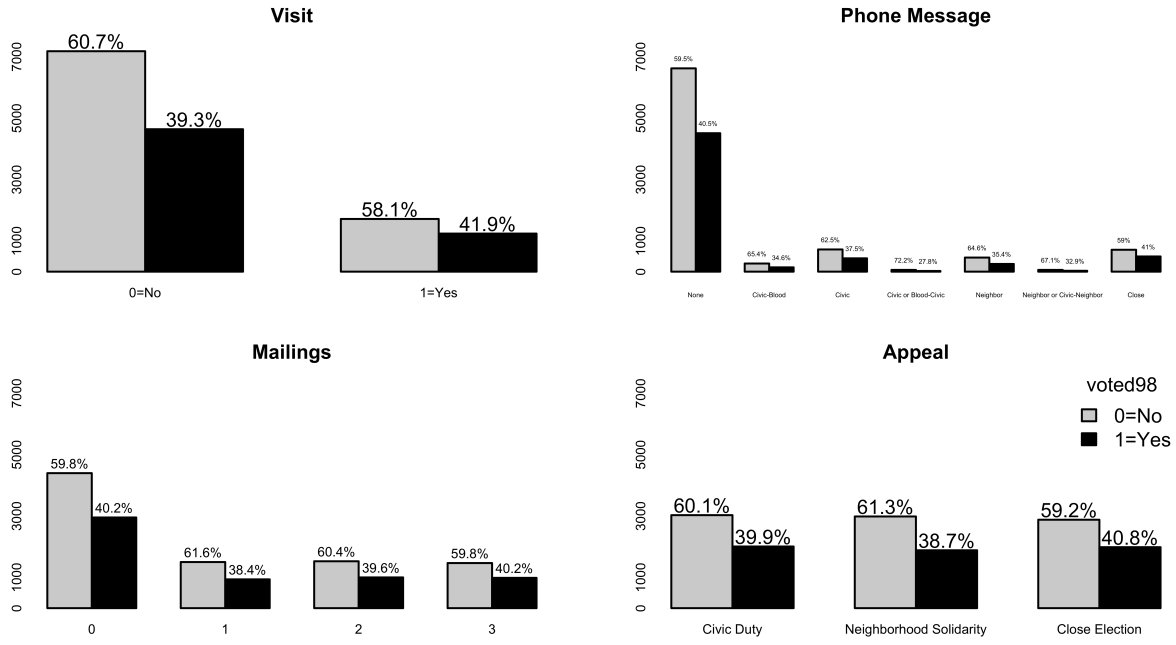

Figure A1: Voting Outcome by Treatment Type

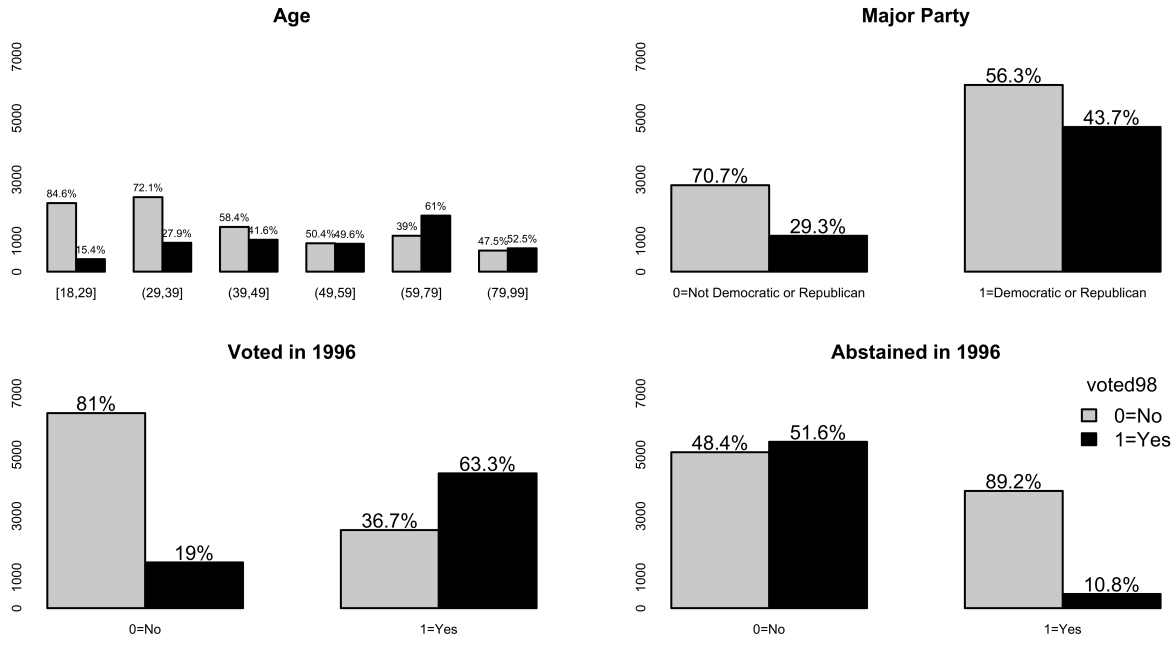

Figure A2: Voting Outcome by Pre-Treatment Control

Table A3: Get-Out-the-Vote Treatment Interactions

	Visit	Phone	Mailings	Appeal	Registered	Voted	Proportion
1	Yes	Civic-Blood	1	Civic Duty	13	8	61.5%
2	No	Civic or Blood-Civic	1	Civic Duty	12	6	50.0%
3	Yes	Neighbor	2	Neighborhood Solidarity	46	23	50.0%
4	Yes	Neighbor or Civic-Neighbor	2	Neighborhood Solidarity	4	2	50.0%
5	Yes	Civic	2	Civic Duty	55	26	47.3%
6	Yes	Neighbor or Civic-Neighbor	1	Neighborhood Solidarity	11	5	45.5%
7	Yes	Civic	0	Neighborhood Solidarity	40	18	45.0%
8	Yes	None	1	Close Election	87	39	44.8%
9	Yes	None	0	Civic Duty	506	226	44.7%
10	Yes	None	2	Close Election	112	50	44.6%
11	Yes	None	3	Civic Duty	110	49	44.5%
12	Yes	Neighbor	3	Neighborhood Solidarity	45	20	44.4%
13	Yes	None	0	Close Election	431	190	44.1%
14	Yes	Close	1	Close Election	68	30	44.1%
15	No	Close	2	Close Election	244	107	43.9%
16	Yes	None	3	Close Election	89	39	43.8%
17	Yes	Civic	3	Civic Duty	53	23	43.4%
18	No	None	3	Civic Duty	393	170	43.3%
19	Yes	Civic or Blood-Civic	2	Civic Duty	7	3	42.9%
20	No	None	3	Close Election	397	169	42.6%
21	Yes	Close	2	Close Election	54	23	42.6%
22	No	None	2	Neighborhood Solidarity	421	178	42.3%
23	Yes	None	0	Neighborhood Solidarity	411	174	42.3%
24	Yes	Civic	1	Neighborhood Solidarity	12	5	41.7%
25	Yes	Civic	2	Neighborhood Solidarity	12	5	41.7%
26	No	Close	3	Close Election	250	104	41.6%
27	No	Close	1	Close Election	260	107	41.2%
28	Yes	None	2	Neighborhood Solidarity	105	43	41.0%
29	No	None	0	Close Election	1742	702	40.3%
30	No	None	2	Civic Duty	412	166	40.3%
31	No	None	3	Neighborhood Solidarity	376	151	40.2%
32	No	None	0	Civic Duty	1772	706	39.8%
33	No	Civic	2	Civic Duty	196	78	39.8%
34	No	None	0	Neighborhood Solidarity	1755	693	39.5%
35	Yes	Close	3	Close Election	76	30	39.5%
36	No	None	1	Close Election	386	152	39.4%
37	No	None	1	Civic Duty	438	172	39.3%
38	Yes	None	1	Civic Duty	80	31	38.8%
39	No	Civic	3	Civic Duty	197	76	38.6%
40	No	None	1	Neighborhood Solidarity	400	154	38.5%
41	Yes	Civic-Blood	0	Civic Duty	39	15	38.5%
42	No	Close	0	Close Election	200	76	38.0%
43	No	Civic	1	Civic Duty	187	69	36.9%
44	No	Neighbor or Civic-Neighbor	1	Neighborhood Solidarity	19	7	36.8%
45	Yes	None	2	Civic Duty	110	40	36.4%
46	No	None	2	Close Election	414	150	36.2%
47	Yes	None	1	Neighborhood Solidarity	90	32	35.6%
48	Yes	None	3	Neighborhood Solidarity	93	33	35.5%
49	No	Civic-Blood	2	Civic Duty	48	17	35.4%
50	No	Neighbor	3	Neighborhood Solidarity	207	73	35.3%
51	No	Civic	0	Neighborhood Solidarity	208	71	34.1%
52	No	Civic-Blood	0	Civic Duty	190	64	33.7%
53	No	Neighbor	1	Neighborhood Solidarity	188	63	33.5%
54	Yes	Civic	3	Neighborhood Solidarity	9	3	33.3%
55	No	Civic	3	Neighborhood Solidarity	52	17	32.7%
56	No	Civic-Blood	3	Civic Duty	43	14	32.6%
57	No	Neighbor	2	Neighborhood Solidarity	179	58	32.4%
58	Yes	Close	0	Close Election	56	18	32.1%
59	No	Civic-Blood	1	Civic Duty	50	16	32.0%
60	No	Civic	1	Neighborhood Solidarity	44	14	31.8%
61	Yes	Civic	1	Civic Duty	44	14	31.8%
62	No	Civic	2	Neighborhood Solidarity	48	15	31.2%
63	Yes	Neighbor	1	Neighborhood Solidarity	45	14	31.1%
64	Yes	Civic-Blood	3	Civic Duty	13	4	30.8%
65	No	Neighbor or Civic-Neighbor	2	Neighborhood Solidarity	23	7	30.4%
66	No	Neighbor or Civic-Neighbor	3	Neighborhood Solidarity	21	6	28.6%
67	No	Civic or Blood-Civic	2	Civic Duty	29	8	27.6%
68	Yes	Civic or Blood-Civic	3	Civic Duty	8	2	25.0%
69	Yes	Civic-Blood	2	Civic Duty	9	2	22.2%
70	No	Civic or Blood-Civic	3	Civic Duty	17	3	17.6%
71	Yes	Neighbor or Civic-Neighbor	3	Neighborhood Solidarity	7	1	14.3%
72	Yes	Civic or Blood-Civic	1	Civic Duty	6	0	0.0%

Table A4 below provides a breakdown of the proportion of registered voters that voted in 1998 by each of the combinations of the pre-treatment control covariates.

Table A4: Get-Out-the-Vote Control Interactions

	Age	Major Party	Voted in '96	Abstained in '96	Registered	Voted	Proportion
1	[18,29]	0	0	0	417	56	13.4%
2	[18,29]	0	0	1	458	17	3.7%
3	[18,29]	0	1	0	267	60	22.5%
4	[18,29]	1	0	0	630	105	16.7%
5	[18,29]	1	0	1	458	30	6.6%
6	[18,29]	1	1	0	411	140	34.1%
7	(29,39]	0	0	0	334	65	19.5%
8	(29,39]	0	0	1	300	15	5.0%
9	(29,39]	0	1	0	335	139	41.5%
10	(29,39]	1	0	0	757	209	27.6%
11	(29,39]	1	0	1	779	68	8.7%
12	(29,39]	1	1	0	861	444	51.6%
13	(39,49]	0	0	0	149	41	27.5%
14	(39,49]	0	0	1	201	17	8.5%
15	(39,49]	0	1	0	276	148	53.6%
16	(39,49]	1	0	0	464	166	35.8%
17	(39,49]	1	0	1	503	78	15.5%
18	(39,49]	1	1	0	901	588	65.3%
19	(49,59]	0	0	0	89	34	38.2%
20	(49,59]	0	0	1	114	10	8.8%
21	(49,59]	0	1	0	200	116	58.0%
22	(49,59]	1	0	0	286	134	46.9%
23	(49,59]	1	0	1	371	56	15.1%
24	(49,59]	1	1	0	772	558	72.3%
25	(59,79]	0	0	0	77	35	45.5%
26	(59,79]	0	0	1	143	26	18.2%
27	(59,79]	0	1	0	359	262	73.0%
28	(59,79]	1	0	0	272	142	52.2%
29	(59,79]	1	0	1	523	111	21.2%
30	(59,79]	1	1	0	1620	1249	77.1%
31	(79,99]	0	0	0	25	8	32.0%
32	(79,99]	0	0	1	92	11	12.0%
33	(79,99]	0	1	0	147	108	73.5%
34	(79,99]	1	0	0	62	32	51.6%
35	(79,99]	1	0	1	337	23	6.8%
36	(79,99]	1	1	0	784	578	73.7%

Table A5 provides a preview of the `LaLonde` dataset. This dataset includes one binary outcome variable, one binary treatment variable, and ten pre-treatment control covariates. Specifically, `outcome` is a binary outcome variable of whether earnings in 1978 are larger than in 1975; `treat` is a binary treatment variable for whether an individual received the job training or not; `age` is an ordinal control for the age in years of workers; `educ` is an ordinal control for the years of education of workers; `black` is a binary control for whether the worker is black or not; `hisp` is a binary control for whether the worker is Hispanic or not; `white` is a binary control for whether the worker is white or not; `marr` is a binary control for whether the worker is married or not; `nodegr` is a binary control for whether the worker has a high school degree or not; `log.re75` is a continuous control for workers pre-treatment log earnings in 1975; `u75` is a binary control for whether the worker was unemployed in 1975 or not.

Table A5: LaLonde (1986) National Supported Work Study

	outcome	treat	age	educ	black	hisp	white	marr	nodegr	log.re75	u75
1	0	0	23	10	1	0	0	0	1	0	1
2	1	0	26	12	0	0	1	0	0	0	1
3	0	0	22	9	1	0	0	0	1	0	1
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
720	0	1	24	10	1	0	0	1	1	8.31	0
721	0	1	33	11	1	0	0	1	1	10.13	0
722	1	1	33	12	1	0	0	1	0	9.3	0

Figure A3 provides the proportion that had larger earnings in the control and treatment groups. Whereas, Figure A4 provides the proportion that had larger earnings by the levels of each of the pre-treatment controls.

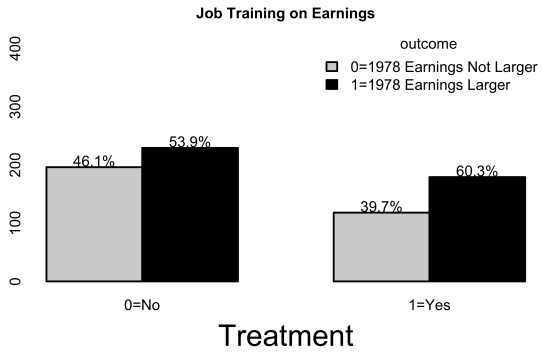

Figure A3: Earnings Outcome by Treatment

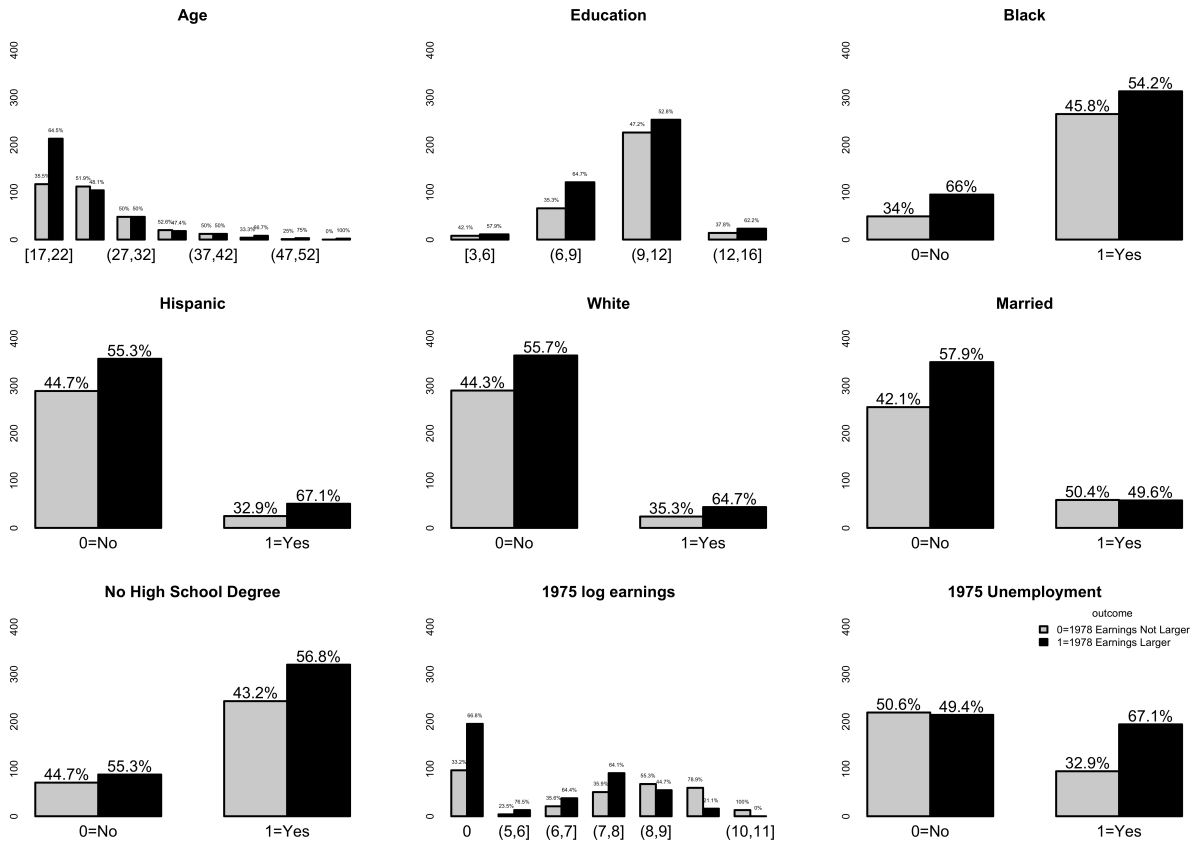

Figure A4: Earnings Outcome by Pre-Treatment Control