Journal of Machine Learning Research 27 (2026) 1-56 Submitted 4/23; Revised 8/25; Published 2/26

A causal fused lasso for interpretable heterogeneous treatment effects estimation

Oscar Hernan Madrid Padilla OSCAR.MADRID@STAT.UCLA.EDU
Department of Statistics and Data Sciences
Univeristy of California, Los Angeles
Los Angeles, CA 90095.

Yanzhen Chen[*] IMYANZHEN@UST.HK
Department of ISOM
Hong Kong University of Science and Technology
Clear Water Bay, Hong Kong.

Carlos Misael Madrid Padilla[†] CARLOSMISAEL@WUSTL.EDU
Department of Statistics and Data Science
Washington University in St Louis
St Louis, MO 63130.

Gabriel Ruiz RUIZG@UCLA.EDU
Department of Statistics
Univeristy of California, Los Angeles
Los Angeles, CA 90095.

Editor: Eric Laber

Abstract

We propose a novel method for estimating heterogeneous treatment effects based on the fused lasso. By first ordering samples based on the propensity or prognostic score, we match units from the treatment and control groups. We then run the fused lasso to obtain piecewise constant treatment effects with respect to the ordering defined by the score. Similar to the existing methods based on discretizing the score, our methods yield interpretable subgroup effects. However, existing methods fixed the subgroup a priori, but our causal fused lasso forms data-adaptive subgroups. We show that the estimator consistently estimates the treatment effects conditional on the score under very general conditions on the covariates and treatment. We demonstrate the performance of our procedure using extensive experiments that show that it can be interpretable and competitive with state-of-the-art methods.

Keywords: Causality, propensity score, prognostic score, total variation

1. Introduction

Causal inference focuses on the causal relationships between covariates and their outcomes, yet is deeply rooted in and advances the way we understand the world. Applications of

[*] These authors contributed equally

[†] These authors contributed equally

causal inference include medical tests and personalized medicine (Zhao et al., 2017; Imbens and Rubin, 2015), and economic and public policy evaluations (Angrist and Pischke, 2008; Ding et al., 2016; Shalit et al., 2017). Our paper proposes a powerful tool for estimating heterogeneous treatment effects under general assumptions.

We adopt the potential outcomes framework (Neyman, 1923; Rubin, 1974), where each subject has an observed outcome variable caused by a treatment indicator and a set of other predictors. The main challenge in estimating heterogeneous treatment effects stems from the fact that for any given subject we only observe the outcome under treatment or control but not both. This is also summarized as a "missing data" problem (Ding et al., 2018). To address this problem, some pioneering works relied on matching via pre-specified groups (Assmann et al., 2000; Pocock et al., 2002; Cook et al., 2004). However, these approaches are sensitive to subject grouping which are largely selected using domain expertise. Our approach, in contrast, does not rely on pre-specified subgroups. Rather, it simultaneously identifies the subgroups and their associated treatment effects. While our approach is similar in spirit to Abadie et al. (2018), it is more flexible allowing for discontinuous treatment effect functions.

Our estimator integrates the merits of similarity scores with the fused lasso method using a simple two-step approach. First, we construct a statistic for each unit and sort observations according to it. The intuition of this step is to summarize the similarities among units using the statistics constructed. In the second step, we perform matching of units of the treatment and control groups based on the statistics generated in the first step. The differences in observed outcomes between the matched pairs guide the fused lasso method, which is a one-dimensional nonparametric regression method as introduced in Mammen et al. (1997) and Tibshirani et al. (2005), to estimate the treatment effects for different units. A key difference between our causal fused lasso approach and the usual fused lasso is that in the latter, there is a given input signal y and an ordering associated to it. In contrast, in causal inference there are no measurements available associated with the individual effects, which is the reason behind our two-step approach.

To be more specific, for the first step of our proposed method, we capture similarities among units using widely adopted statistics such as the propensity score (Rosenbaum and Rubin (1983, 1984)), and the prognostic score (see e.g Hansen (2008); Abadie et al. (2018)). For the former, we fit a parametric model such as logistic regression. For the prognostic score method, we regress the outcome on the covariates using the control group data only.

Despite being a simple, our method enjoys the following properties:

1. From a theoretical perspective, we establish that our method consistently estimates treatment effects under minimal assumptions. These assumptions include a general random design for covariates and bounded variation of the conditional mean of the outcome, conditional on both the subgroup and treatment assignment.

2. Our estimator is computationally efficient, with overall complexity on the order of $O(nd + n^2)$, where n is the number of units and d is the number of covariates. The nd term corresponds to fitting a linear regression model and can be considered linear in n when d is small. The n^2 term, which may be dominant, arises from a matching step that requires computing pairwise scores between all units. This is comparable

to the complexity of a K-nearest neighbors algorithm. However, this computational cost can be significantly reduced through parallel computing.

3. Unlike many nonparametric methods, our estimator offers interpretability. Moreover, experimental results demonstrate that it either outperforms or matches state-of-the-art approaches in estimating heterogeneous treatment effects, as measured by mean squared error for estimating the conditional treatment effects.

1.1 Previous work

A substantial body of statistical research has focused on estimating heterogeneous treatment effects, with many studies building on the seminal Bayesian additive regression tree (BART) framework introduced by Chipman et al. (2010b). The core idea behind BART-based methods is to impose the prior commonly used in BART on both the regression function of the control group and that of the treatment group, as exemplified in Hill (2011); Green and Kern (2012); Hill and Su (2013). More recently, Hahn et al. (2020) introduced a BART-based approach designed to handle small effect sizes and confounding by observables. Beyond BART, other Bayesian methods include the linear model prior proposed by Heckman et al. (2014) and the Bayesian nonparametric framework developed in Taddy et al. (2016).

In a separate line of research, tree-based regression methods have been developed to estimate heterogeneous treatment effects, with regression trees emerging as a key approach. This direction was initiated by Su et al. (2009), who proposed an estimator based on the widely used CART method Breiman et al. (1984). More recently, Athey and Imbens (2016) and Wager and Athey (2018) extended this work by employing the random forest framework from Breiman (2001) to construct estimators. A significant contribution of Wager and Athey (2018) was the introduction of an inferential framework for treatment effect estimates using the infinitesimal jackknife. Building on this foundation, Athey et al. (2019) developed a generalized random forest approach, improving robustness in practical applications compared to the estimator from Wager and Athey (2018).

Beyond regression tree-based methods, various machine learning approaches have been explored for estimating heterogeneous treatment effects. Crump et al. (2008) introduced nonparametric tests to detect treatment effect heterogeneity. Imai et al. (2013) proposed a method that integrates hinge loss (Wahba, 2002) with lasso regularization (Tibshirani et al., 2005) to enhance estimation accuracy. Tian et al. (2014) developed an approach capable of handling a high-dimensional feature space while capturing interactions between treatment and covariates.

Other contributions include Weisberg and Pontes (2015), who designed a method based on variable selection, and Taddy et al. (2016), who developed Bayesian nonparametric techniques applicable to both linear regression and tree-based models. Syrgkanis et al. (2019) introduced a flexible framework that can incorporate any machine learning method and, under valid instrumental variables, account for unobserved confounders. From a theoretical perspective, Gao and Han (2020) analyzed the fundamental limits of estimating heterogeneous treatment effects under Hölder smoothness conditions.

A different line of work focuses on meta-learning strategies in causal inference. Künzel et al. (2019) proposed a meta-learner framework that can leverage various estimators and is particularly effective when treatment group sizes are highly imbalanced. We note that the term meta-learning is sometimes used differently in the broader machine learning literature. For example, it can refer to transfer learning settings where metadata across tasks is used to inform model training Vanschoren (2019). Our use of meta-learning follows the convention in causal inference, where meta-algorithms aggregate supervised learners to estimate treatment effects.

For comprehensive reviews of testing procedures and estimation methods for individual treatment effects, see Willke et al. (2012) and Caron et al. (2022).

In this paper, we propose a two-step method for estimating heterogeneous treatment effects. First, we construct a one-dimensional score that serves as the basis for ordering observations. We then apply the one-dimensional fused lasso, following Tibshirani et al. (2005), to identify subgroups. By design, our approach directly estimates subgroups of individuals with the same treatment effect without requiring heuristic arguments or pre-specifying subgroups, as in Abadie et al. (2018). Moreover, our method is highly interpretable—arguably even simpler than CART—since it relies solely on a single learned covariate.

Finally, we highlight related work regarding the fused lasso, the nonparametric tool that we use in this paper. Also known as total variation denoising, the fused lasso first appeared in the machine learning literature (Rudin et al., 1992), and then in the statistical literature (Mammen et al., 1997). A discretized version of total variation regularization was introduced by Tibshirani et al. (2005). Since then, multiple authors have used the fused lasso for nonparametric regression in different frameworks. Tibshirani et al. (2014) proved that the fused lasso can attain minimax rates for estimation of a one-dimensional function that has bounded variation. Guntuboyina et al. (2020) provided minimax results for the fused lasso when estimating piecewise constant functions. Hütter and Rigollet (2016); Chatterjee and Goswami (2019) studied the convergence rates of the fused lasso for denoising of grid graphs. Wang et al. (2016); Padilla et al. (2018) considered extensions of the fused lasso to general graphs structures. Padilla et al. (2020) proposed the fused lasso for multivariate nonparametric regression and showed adaptivity results for different levels of the regression function. Ortelli and van de Geer (2019) studied further connections between the lasso and fused lasso.

1.2 Notation

For two random variables X and Y, we use the notation $X \perp\!\!\!\perp Y$ to indicate that they are independent. We write $a_n = O(b_n)$ if there exist constants N and C such that $n \geq N$ implies that $a_n \leq C b_n$, for sequences $\{a_n\}, \{b_n\} \subset \mathbb{R}$. In addition, when $a_n = O(b_n)$ and $b_n = O(a_n)$ we use the notation $a_n \asymp b_n$ and sometimes the notation $a_n = \Theta(b_n)$. For a sequence of random variables $\{X_n\}$, we denote $X_n = O_{\mathbb{P}}(a_n)$ if for every $\epsilon > 0$ there exist positive N and C such that $\mathbb{P}(|X_n| > C a_n) < \epsilon$ for all $n \geq N$.

Finally, for a random vector $X \in \mathbb{R}^d$ we say that X is sub-Gaussian(C) for $C > 0$ if

$$\|X\|_{\psi_2} := \sup_{v \in \mathbb{R}^d \,:\, \|v\|=1} \|v^\top X\|_{\psi_2} < C,$$

where for a random variable u we have

$$\|u\|_{\psi_2} := \sup_{k \geq 1} k^{-1/2} \left\{ E\left(|u|^k\right)^{1/k} \right\}.$$

1.3 Outline

The rest of the paper proceeds as follows. Section 2 describes the mathematical setup of the paper and presents the proposed class of estimators. Section 3 then develops theory for the corresponding estimators based on propensity and prognostic scores. Section 4 provides extensive comparisons with state-of-the-art methods in the literature on heterogeneous treatment effects estimation. All the proofs of the theoretical results can be found in the Supplementary Material.

2. Methodology

In this section, we introduce the main methods of the paper. We begin in Section 2.1 with a predecessor estimator that serves as both a foundation and motivation for our approach. In Section 2.2, we develop a prognostic score-based estimator suited for completely randomized experiments, followed by a propensity score-based method in Section 2.3, designed for observational studies.

2.1 A predecessor estimator

Consider independent draws $\{Z_i, X_i, Y_i(1), Y_i(0)\}_{i=1}^n$, where $Z_i \in \{0, 1\}$ is a binary treatment indicator, $X_i \in \{1, \ldots, K\}$ is a discrete and ordinal covariate, and $Y_i \in \mathbb{R}$ is an outcome. Under the Stable Unit Treatment Value Assumption (SUTVA) (see, e.g., Imbens and Rubin (2015)), the observed outcome can be expressed as:

$$Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0). \tag{1}$$

Under the unconfoundedness assumption $Z_i \perp\!\!\!\perp \{Y_i(1), Y_i(0)\} \mid X_i$ and overlap (see Imbens (2004)), we can write the subgroup causal effect as

$$\begin{aligned} \tau_{[k]} &= E\{Y_i(1) - Y_i(0) \mid X_i = k\} \\ &= E\{Y_i(1) \mid Z_i = 1, X_i = k\} - E\{Y_i(0) \mid Z_i = 0, X_i = k\} \\ &= E(Y_i \mid Z_i = 1, X_i = k) - E(Y_i \mid Z_i = 0, X_i = k), \end{aligned}$$

which can be identified by the joint distribution of the observed data $\{(Z_i, X_i, Y_i)\}_{i=1}^n$.

We can estimate $\tau_{[k]}$ by the sample moments:

$$\hat{\tau}_{[k]} = \frac{1}{n_{[k]1}} \sum_{Z_i=1, X_i=k} Y_i - \frac{1}{n_{[k]0}} \sum_{Z_i=0, X_i=k} Y_i = \bar{Y}_{[k]1} - \bar{Y}_{[k]0}, \tag{2}$$

where $n_{[k]z}$ is the sample size of units with covariate value k under the treatment arm z.

However, the estimates $\hat{\tau}_{[k]}$ are well-behaved only when the sample sizes $n_{[k]z}$ are sufficiently large. When sample sizes are small, particularly as K increases, many of the estimates $\hat{\tau}_{[k]}$ can become highly variable, resulting in noisy approximations of the true parameters.

When we expect that many subgroup causal effects are small or even zero, it is reasonable to shrink some $\hat{\tau}_{[k]}$'s to zero, similar to the idea of the lasso estimator (Tibshirani, 1996). Moreover, we may also expect that some subgroup causal effects are close or even identical, then it is reasonable to shrink some $\hat{\tau}_{[k]}$'s to the same value, similar to the idea of the fused lasso (Tibshirani et al., 2005). Motivated by these considerations, define

$$\tilde{\tau} \,=\, \underset{b \in \mathbb{R}^K}{\arg\min} \,\left\{ \frac{1}{2} \sum_{k=1}^{K} (b_k - \hat{\tau}_{[k]})^2 + \lambda \sum_{k=1}^{K-1} |b_k - b_{k+1}| \right\}, \tag{3}$$

for some tuning parameter $\lambda > 0$. As in Tibshirani et al. (2005) and Tibshirani et al. (2014), the second term in (3) is the fused lasso penalty to enforce a piecewise constant structure of the estimates. We use ℓ_1 regularization instead of $\sum_{k=1}^{K-1} |b_k - b_{k+1}|^2$ because the latter would result in linear estimator that is not locally adaptive, see Donoho and Johnstone (1998).

Notice that the term $\sum_{k=1}^{K} |b_k - b_{k+1}|$ applies the same penalty to each difference $b_k - b_{k+1}$. At first glance, this might suggest that the categories of X must be equally spaced. Specifically, if categories 1 and 2 are closer than categories 2 and 3 in a given application, one might wonder whether $|b_1 - b_2|$ should be penalized more than $|b_2 - b_3|$. The answer is no—the fused lasso penalty is adaptive in the sense that it can adjust to the true signal's structure, even when the locations of the jumps are unknown or unevenly spaced. For further details, see Tibshirani et al. (2014) and Guntuboyina et al. (2020).

While $\tilde{\tau}$ seems appealing, it requires that the covariates X_i are univariate and categorical. Also, it requires that there is an order of the covariates under which treatment effects are piecewise constant. Both of these assumptions are usually not met in practice, as typically X_i is a vector that can have continuous random variables. The next section proposes a general class of estimators to handle such situations.

2.2 Prognostic-based estimator

In this subsection, we focus on completely randomized experiments where $Z_i \perp\!\!\!\perp \{Y_i(1), Y_i(0), X_i\}$, with $X_i \in \mathbb{R}^d$. representing the covariate vector. We define $E\{Y(0)|X\}$ as the prognostic score (Hansen, 2008). Furthermore, let $g(X) = X^\top \theta^*$ be the best linear approximation of the prognostic score, where

$$\theta^* \,=\, \underset{\theta \in \mathbb{R}^d}{\arg\min}\, E\left\{ \left(Y - X^\top \theta \right)^2 |Z = 0 \right\}. \tag{4}$$

Abadie et al. (2018) assumed a piecewise structure for treatment effects when seen as a function of the approximate prognostic score. However, their approach relied on a fixed number (typically five) of subgroups, determined by discretizing an estimator of the approximate prognostic score. Inspired by Abadie et al. (2018), we assume that the treatment effects, as a function of the prognostic score, are piecewise constant. However, rather than pre-specifying the number of subgroups, we assume that $\tau(s) = E\{Y(1) - Y(0) \mid g(X) = s\}$ is either piecewise constant with an unknown number of segments or has bounded variation. See Section 3 for the precise definition. Let $\sigma : \{1, \ldots, n\} \to \{1, \ldots, n\}$ be the permutation such that

$$g\{X_{\sigma(1)}\} \leq \cdots \leq \cdots \leq g\{X_{\sigma(n)}\}.$$

If $g(s)$ is known and the individual causal effects $Y_i(1) - Y_i(0)$ are known, a natural estimator for $\tau_i^* = \tau\{g(X_i)\}$, for $i = 1, \ldots, n$, would be the solution to

$$\underset{b \in \mathbb{R}^n}{\text{minimize}} \left[\frac{1}{2} \sum_{i=1}^{n} \{Y_i(1) - Y_i(0) - b_i\}^2 + \lambda \sum_{i=1}^{n-1} \left| b_{\sigma(i)} - b_{\sigma(i+1)} \right| \right], \tag{5}$$

for a tuning parameter $\lambda > 0$. This is similar to the standard fused lasso for one-dimensional nonparametric regression (Tibshirani et al., 2005). The first term in (5) provides a measure of fit to the data, and the second term penalizes the total variation to enforce a piecewise constant structure of the estimated treatment effects. The goal is to adaptively estimate subgroups where the treatment effect, conditional on the prognostic score, remains constant. This approach is conceptually similar to Morucci et al. (2023), where the authors constructed a neighborhood for each unit based on a score derived from covariates and then estimated both $E\{Y(1) \mid g(X)\}$ and $E\{Y(0) \mid g(X)\}$ using a nearest-neighbor-type estimator. However, a key distinction is that the fused lasso Tibshirani et al. (2014) is well known for its ability to adapt to discontinuities in the regression function, whereas the method in Morucci et al. (2023) relies on a Lipschitz continuity condition.

In practice, neither $g(s)$ or the individual treatment effects $Y_i(1) - Y_i(0)$ are known. We first estimate the prognostic scores as $\hat{g}(X_i) = X_i^\top \hat{\theta}$ where

$$\hat{\theta} \;=\; \underset{\theta \in \mathbb{R}^d}{\arg\min} \; \left\{ \frac{1}{m} \sum_{i=1}^{m} \left(Y_i' - X_i'^\top \theta \right)^2 \right\}, \tag{6}$$

where $\{(X_i', Y_i')\}_{i=1}^m$ are independent copies of (X, Y) conditional on $Z = 0$ Additionally, to establish our theoretical results, we require that $\{(X_i', Y_i')\}_{i=1}^m$ be independent of $\{(Z_i, X_i, Y_i)\}_{i=1}^n$. This condition can be satisfied through sample splitting, where the data is evenly divided into two subsets of equal size. Hence, $\hat{\theta}$ is the estimated vector of coefficients when regressing the outcome variable on the covariates conditioning on the treatment assignment being the control group. Based on the estimated prognostic score, we find the permutation $\hat{\sigma}$ satisfying

$$\hat{g}\{X_{\hat{\sigma}(1)}\} \leq \cdots \leq \hat{g}\{X_{\hat{\sigma}(n)}\}. \tag{7}$$

We then match the units to impute the missing potential outcomes. Define

$$\widetilde{Y}_i \;=\; Y_{N(i)}, \quad \text{with} \;\; N(i) \;=\; \underset{j \,:\, Z_j \neq Z_i}{\arg\min} \, |\, \hat{g}(X_i) - \hat{g}(X_j)|. \tag{8}$$

So if $Z_i = 1$, then $\widetilde{Y}_i$ is the imputed $Y_i(0)$ and the imputed individual effect is $Y_i - \widetilde{Y}_i$; if $Z_i = 0$, then $\widetilde{Y}_i$ is the imputed $Y_i(1)$ and the imputed individual effect is $\widetilde{Y}_i - Y_i$. With these ingredients, we define the estimator

$$\hat{\tau} = \underset{b \in \mathbb{R}^n}{\arg\min} \left[\frac{1}{2} \sum_{i=1}^{n} \left\{ Y_i - \widetilde{Y}_i + (-1)^{Z_i} b_i \right\}^2 + \lambda \sum_{i=1}^{n-1} \left| b_{\hat{\sigma}(i)} - b_{\hat{\sigma}(i+1)} \right| \right]. \tag{9}$$

The optimization problem in (9) can be solved in $O(n)$ operations by employing the algorithm from Johnson (2013) or that of Barbero and Sra (2014). Therefore, $\hat{\tau}$ is our

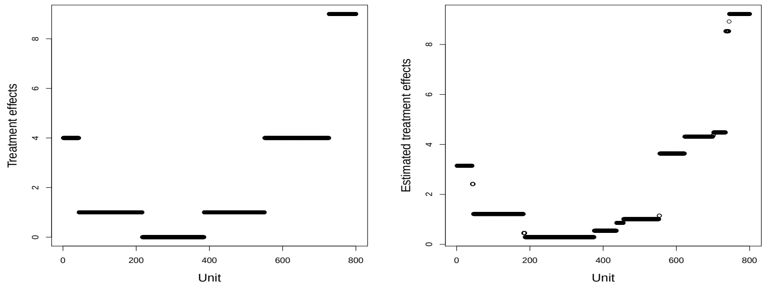

Figure 1: The left panel shows a plot of $\tau^*_\sigma = (\tau^*_{\sigma(1)}, \ldots, \tau^*_{\sigma(n)})^\top$, where σ is a permutation satisfying $X^\top_{\sigma(1)}\boldsymbol{e}_1 < X^\top_{\sigma(2)}\boldsymbol{e}_1 < \cdots < X^\top_{\sigma(n)}\boldsymbol{e}_1$. The right panel then shows a plot of $(\hat{\tau}_{\hat{\sigma}(1)}, \ldots, \hat{\tau}_{\hat{\sigma}(n)})^\top$, where $\hat{\sigma}$ is the order based on the estimated prognostic score as defined in (7).

final estimator of the vector of subgroup treatment effects τ^*, where $\tau^*_i = \tau\{g(X_i)\}$, for $i = 1, \ldots, n$, with $g(x) = x^\top\theta^*$. Similarly, for $x \notin \{X_1, \ldots, X_n\}$ we can estimate $\tau\{g(x)\}$ with $\hat{\tau}_i$ where X_i is the closest to x among $X_1, \ldots, X_n$. A related prediction rule was used in a different context by Padilla et al. (2020).

Because $\hat{\tau}$ is piecewise constant, we can think of its different pieces as data-driven subgroups of units. These so-called subgroups are estimated adaptively and do not need to be prespecified. See Figure 1 for a visual example of $\hat{\tau}$.

Notice that we have used a different data set for estimating the prognostic score $\hat{\theta}$ than the one for which we estimate the treatment effects. This can be achieved in practice by sample splitting. The reason why we proceed in this way is to prevent $\hat{\sigma}$ from being correlated to $\{(Z_i, X_i, Y_i)\}_{i=1}^n$.

Regarding the choice of λ, we proceed as in Tibshirani et al. (2012). Thus, for each value of λ from a list of choices, we compute the estimator in (9) and its corresponding degrees of freedom as in Tibshirani et al. (2012). Then we select the value of λ with the smaller Bayesian information criterion (BIC).

Example 1 *To illustrate the behavior of $\hat{\tau}$ defined in (9) we consider a simple example. We generate $\{(Z_i, X_i, Y_i)\}_{i=1}^n$ with $n = 800$, and $d = 10$, from the model*

$$\begin{aligned} Y_i(0) &= f_0(X_i) + \epsilon_i, \\ Y_i(1) &= f_0(X_i) + \tau(X_i) + \epsilon_i, \\ \mathbb{P}(Z_i = 1|X_i) &= 0.5, \\ f_0(x) &= \sin\{2(4\pi x^\top\boldsymbol{e}_1 - 2)\} + 2.5(4\pi x^\top\boldsymbol{e}_1 - 2) + 1, \quad \forall x \in [0,1]^d, \end{aligned} \tag{10}$$

$$\begin{aligned} \tau(x) &= \left\lfloor \frac{10}{1+\exp\left(\frac{f_0(x)}{15}-\frac{1}{30}\right)} - 5 \right\rceil^2, \quad \forall x \in [0,1]^d, \\ X_i &\overset{\text{ind}}{\sim} U[0,1]^d, \\ \epsilon_i &\overset{\text{ind}}{\sim} \mathcal{N}(0,1), \end{aligned} \tag{11}$$

where $\boldsymbol{e}_1 = (1,0,\ldots,0)^\top \in \mathbb{R}^d$*.*

Notice that, by construction, τ *can be interpreted as a piecewise constant function of* $x^\top \boldsymbol{e}_1$*. Figure 1 illustrates a plot of the vector of treatment effects* $\tau^* = (\tau(X_1),\ldots,\tau(X_n))^\top$*. In addition, Figure 1 shows that our estimator* $\hat{\tau}$ *can reasonably estimate* τ^* *when choosing the tuning parameter via BIC. This is surprising to an extent since the ordering that makes* τ^* *piecewise constant is unknown, as both the propensity score and the function* f_0 *are unknown.*

2.3 Propensity score based estimator

While the prognostic score-based approach can, in principle, be applied to observational studies, it is primarily suited for randomized experiments. As stated in Theorem 2, the estimate $\hat{\tau}$ obtained from Equation (9) targets the quantity

$$\rho_i^* = E\{Y(1)|g(X) = g(X_i), Z=1\} - E\{Y(0)|g(X) = g(X_i), Z=0\},$$

for $i = 1,\ldots,n$. However, to have ρ_i^* to equals the conditional treatment effect $E\{Y(1) - Y(0)|g(X) = g(X_i)\}$, we must assume $Y(0), Y(1) \perp\!\!\!\perp Z \,|\, g(X)$. This ignorability condition may not hold in observational settings. Therefore, the prognostic score-based method described in Section 2.2 is intended for use in randomized experiments, consistent with the estimator proposed in Abadie et al. (2018).

To adapt our approach for observational studies, we follow Rosenbaum and Rubin (1983) and shift focus from the prognostic score to the propensity score. Accordingly, we define the subgroup treatment effect as:

$$\tau(s) = E\left\{Y(1) - Y(0) \,|\, e(X) = s\right\}, \tag{12}$$

where $e(X) = \mathbb{P}(Z=1 \mid X)$ is the propensity score (Rosenbaum and Rubin, 1983). We can define an analogous estimator for $\tau = (\tau\{e(X_i)\})_{i=1}^n$ based on the estimated propensity scores. However, our estimator based on the propensity score is specifically designed for observational studies, as the true propensity score is constant in completely randomized experiments. In such settings, conditioning on the propensity score offers no meaningful variation, and thus it does not make sense to estimate heterogeneous treatment effects as a function of the propensity score.

To be specific, we first estimate the propensity score based on logistic regression to obtain $\hat{e}(X_i) = F(X_i^\top \hat{\theta})$ where $F(x) = \exp(x)/\{1+\exp(x)\}$ and

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^d}{\arg\max} \left(\sum_{i=1}^m \left[Z_i' \log F(X_i'^\top \theta) + (1 - Z_i') \log\{1 - F(X_i'^\top \theta)\} \right] \right). \tag{13}$$

Again the sequence $\{(X_i', Z_i')\}_{i=1}^m$ consists of independent copies of (X, Z) independent of $\{(X_i, Z_i, Y_i)\}_{i=1}^n$. Obtain the permutation $\hat{\sigma}$ that satisfies

$$\hat{e}\{X_{\hat{\sigma}(1)}\} \leq \cdots \leq \hat{e}\{X_{\hat{\sigma}(n)}\}, \tag{14}$$

with $\hat{P}$ being the associated permutation matrix. Use matching to impute the missing potential outcomes based on

$$\widetilde{Y}_i \,=\, Y_{N(i)}, \quad \text{with} \;\; N(i) \,=\, \underset{j\,:\,Z_j \neq Z_i}{\arg\min}\, |\,\hat{e}(X_i) - \hat{e}(X_j)|. \tag{15}$$

The final estimator for τ becomes

$$\hat{\tau} = \hat{P}^T \left(\underset{b \in \mathbb{R}^n}{\arg\min} \left[\frac{1}{2} \sum_{i=1}^{n} \left\{ (-1)^{Z_{\hat{\sigma}(i)}+1} \left(Y_{\hat{\sigma}(i)} - \widetilde{Y}_{\hat{\sigma}(i)} \right) - b_i \right\}^2 + \lambda \sum_{i=1}^{n-1} |b_i - b_{i+1}| \right] \right), \tag{16}$$

for a tuning parameter $\lambda > 0$.

One should be cautious in interpreting the propensity score based estimator defined in (16). Specifically, (16) estimates $\rho\{e(X_i)\}$ for $i = 1, \ldots, n$ where

$$\rho(s) \,:=\, f_1(s) \,-\, f_0(s),$$

with

$$f_0(s) := E\{Y | Z = 0,\, e(X) = s\}, \quad \text{and} \quad f_1(s) := E\{Y | Z = 1,\, e(X) = s\}.$$

However, in general $\rho(s) \neq \tau(s)$ with τ as in (12). Although $\rho(s) = \tau(s)$ for all s provided that $Y(0), Y(1) \perp\!\!\!\perp Z \,|\, e(X)$. See Section 3.2 for a more detailed discussion.

3. Theory

3.1 Main result for prognostic score based estimator

We start by studying the statistical behavior of the prognostic score based estimator defined in (9). Towards that end, we first introduce some assumptions used in our proofs to arrive at our first result.

Assumption 1 (Overlap) *The propensity score $e(X)$ has support $[e_{\min}, e_{\max}] \subset (0, 1)$.*

Assumption 2 (Surrogate prognostic score) *We define*

$$\theta^* \,=\, \underset{\theta \in \mathbb{R}^d}{\arg\min}\; L(\theta),$$

with $L(\theta) = E\left\{\|Y - X^\top \theta\|^2 | Z = 0\right\}$, and suppose that L has a unique minimizer $\theta^* \neq 0$. In addition, we assume that the probability density function of $\tilde{g}(X) := X^\top \theta^*$ is bounded from below and above. The support of $\tilde{g}(X)$ is denoted as $[g_{\min}, g_{\max}]$.

We note that Assumption 2 states that the linear surrogate population prognostic score is well behaved and uniquely defined. This allows us to understand the statistical properties of $\hat{\theta}$ defined in (6), which is potentially a misspecified maximum likelihood estimator (White, 1982).

Assumption 3 (Sub-Gaussian errors) *Define $V(z,x) = E\{Y|Z = z, g(X) = g(x)\}$ and $\epsilon_i = Y_i - V(Z_i, X_i)$ for $i = 1, \ldots, n$, with $g(x) = E\{Y(0) \,|\, X = x\}$. Then the vector $(\epsilon_1, \ldots, \epsilon_n)^\top$ has independent coordinates that are mean zero sub-Gaussian(v) for some constant $v > 0$.*

The previous assumption requires that the errors are independent, and mean zero sub-Gaussian. This condition is standard in the analysis of total variation denoising; see for instance Padilla et al. (2018). The resulting condition allows for general models such as normal, bounded distributions, etc. In addition, Assumption 3 allows for the possibility of heteroscedastic errors.

Our next assumption has to do with behavior of the mean functions of the outcome variable, when conditioning on treatment assignment and prognostic score. We start by recalling the definition of bounded variation. For a function $f : [l,u] \to \mathbb{R}$, we define its total variation as

$$\mathrm{TV}(f) = \sup_{r \geq 1} \mathrm{TV}(f,r), \tag{17}$$

where

$$\mathrm{TV}(f,r) = \sup_{l \leq a_1 \leq \ldots \leq a_r \leq u,} \sum_{l=1}^{r-1} |f(a_l) - f(a_{l+1})|.$$

We say that f has bounded variation if $\mathrm{TV}(f) < \infty$. For a fixed C, the collection

$$\mathcal{F}_C = \{f : [0,1] \to \mathbb{R} : \mathrm{TV}(f) \leq C\},$$

is a rich class of functions that contains, among others, Lipschitz continuous functions, piecewise constant and piecewise Lipschitz functions. We refer the reader to Mammen et al. (1997); Tibshirani et al. (2014) for comprehensive studies of nonparametric regression on the class $\mathcal{F}_C$.

Assumption 4 *The functions $f_1(s) = E\{Y \,|\, Z = 1, g(X) = s\}$, and $f_0(s) = E\{Y \,|\, Z = 0, g(X) = s\}$ for $s \in [g_{\min}, g_{\max}]$ are bounded and have bounded variation. The latter means that $t_l = \mathrm{TV}(f_l, n)$, for $l \in \{0,1\}$, satisfy $\max\{t_0, t_1\} = O(1)$.*

Importantly, Assumption 4 allows for the possibility that functions f_0 and f_1 can have discontinuities. Our next condition imposes a relationship between the prognostic score g defined in Assumption 3 and its surrogate $\tilde{g}$ defined in Assumption 2.

Assumption 5 *Let $\tilde{\sigma}$ and σ be random permutations such that*

$$\tilde{g}\{X_{\tilde{\sigma}(1)}\} \leq \cdots \leq \tilde{g}\{X_{\tilde{\sigma}(n)}\},$$

and

$$g\{X_{\sigma(1)}\} \leq \cdots \leq g\{X_{\sigma(n)}\},$$

respectively, with $\tilde{g}$ as defined in Assumption 2. Then we write

$$\mathcal{K}_j = \left\{ i : \left(\sigma^{-1}(i) < \sigma^{-1}(j) \text{ and } \tilde{\sigma}^{-1}(j) < \tilde{\sigma}^{-1}(i)\right) \text{ or } \left(\sigma^{-1}(j) < \sigma^{-1}(i) \text{ and } \tilde{\sigma}^{-1}(i) < \tilde{\sigma}^{-1}(j)\right) \right\},$$

set $\kappa_j := |\mathcal{K}_j|$ for $j = \{1, \dots, n\}$, and

$$\kappa_{\max} := \max_{j=1,\dots,n} \kappa_j,$$

and require that $\kappa_{\max} = O_{\mathbb{P}}(\overline{\kappa}_n)$, where $\overline{\kappa}_n$ is a deterministic sequence.

Assumption 5 allows to quantify the discrepancy between the order statistics of the prognostic score at the samples and the corresponding order statistics based on the surrogate prognostic score. The following remark further clarifies this.

Remark 1 *Notice that κ_j can be thought as the number of units that have different relative orderings in the rankings induced by σ and $\tilde{\sigma}$. In addition, notice that the parameter $\overline{\kappa}_n$ gives an upper bound on the entries of the vector $(\kappa_1, \dots, \kappa_n)$. In fact, $\overline{\kappa}_n$ can be thought as an ℓ_∞ version of the Kendall-Tau distance between the permutations σ^{-1} and $\tilde{\sigma}^{-1}$. Such Kendall-Tau distance is given as $\sum_{j=1}^n \kappa_j$ (see Kumar and Vassilvitskii (2010) for an overview). Readers can also consider cases in which $\tilde{g}$ and g induce the same ordering. In such cases, $\overline{\kappa}_n$ can be taken as zero.*

Our next assumption is a condition on the covariates.

Assumption 6 *The random vectors $X_1, \dots, X_n$ are independent copies of X which has support $[a, b] \subset \mathbb{R}^d$, for some fixed points $a, b \in \mathbb{R}^d$. In addition, the following holds:*

- *The probability density function of X, p_X, is bounded. This amounts to*

  $$p_{\min} \leq \inf_{x \in [a,b]} p_X(x) \;\leq\; \sup_{x \in [a,b]} p_X(x) \leq \; p_{max},$$

  for some positive constants $p_{\min}$ and $p_{\max}$.

- *There exist $D_{\max}, C_{\min} > 0$ such that*

  $$C_{\min} < \Lambda_{\min}\{E(XX^\top)\} \leq \Lambda_{\max}\{E(XX^\top)\} < D_{\max}, \tag{18}$$

  where $\Lambda_{\min}(\cdot)$ and $\Lambda_{\max}(\cdot)$ are the minimum and maximum eigenvalue functions.

We emphasize that the first condition in Assumption 6 is standard in nonparametric regression, see Padilla et al. (2020). It is slightly more general than assuming that the covariates are uniformly drawn in $[0, 1]^d$ as in the nonparametric regression models in Györfi et al. (2006), and the heterogenous treatment effect setting from Wager and Athey (2018).

With these assumptions, we are now ready to state our first result regarding the estimation of $\tau_i^* = \tau\{g(X_i)\}$, with g the prognostic score defined in Assumption 3.

Theorem 2 *Suppose that Assumptions 1–6 hold, $m \asymp n$, and that*

$$d(\log^{1/2} n + d^{1/2}\|\theta^*\|_1) \;\leq\; \frac{c_1 n}{\log n},$$

for some large enough constant $c_1 > 0$. Then for a value t satisfying

$$t \asymp (\overline{\kappa}_n + 1)\, d \left\{ n(\log^{1/2} n + d^{1/2}\|\theta^*\|_1) \log n \right\}^{1/2},$$

and for a choice of λ with

$$\lambda = \Theta\left\{n^{1/3}(\log n)^2(\log\log n)t^{-1/3}\right\},$$

we have that the estimator defined in (9) satisfies

$$\frac{1}{n}\sum_{i=1}^{n}(\rho_i^*-\hat{\tau}_i)^2 = O_{\mathbb{P}}\left[(\log\log n)(\log n)^2(\overline{\kappa}_n+1)^{2/3}d^{2/3}\left\{\frac{(\log^{1/2} n + d^{1/2}\|\theta^*\|_1)\log n}{n}\right\}^{1/3}\right], \tag{19}$$

where $\rho_i^* = E\{Y(1)|g(X)=g(X_i), Z=1\} - E\{Y(0)|g(X)=g(X_i), Z=0\}$ for $i=1,\ldots,n$. If in addition $Y(0), Y(1) \perp\!\!\!\perp Z\,|\,g(X)$, then (19) holds replacing ρ_i^* with $\tau_i^* = E\{Y(1)-Y(0)|g(X)=g(X_i)\}$ for $i=1,\ldots,n$.

We note that $Y(0), Y(1) \perp\!\!\!\perp Z\,|\,g(X)$ is a strong assumption that might not hold in observational studies, however it might be reasonable in completely randomized trials.

On a related note, Theorem 2 shows that, up to logarithmic factors, our proposed prognostic score based estimator achieves the convergence rate for the mean squared error given by:

$$(\overline{\kappa}_n+1)^{2/3}d^{2/3}\left\{\frac{(1+d^{1/2}\|\theta^*\|_1)}{n}\right\}^{1/3}.$$

Here, the dimension d is allowed to grow with the sample size n, and accordingly, $\|\theta^*\|_1$ may also grow with n. However, in the special case where $d = O(1)$ and $\|\theta^*\|_1 = \mathrm{O}(1)$, the rate simplifies to $(\overline{\kappa}_n+1)^{2/3}/n^{1/3}$. Notably, when the prognostic score is exactly linear—so that $g = \tilde{g}$ and hence $\sigma = \tilde{\sigma}$ —we have $\overline{\kappa}_n = 0$, and the rate further simplifies to $n^{-1/3}$.

It is also possible for $\sigma = \tilde{\sigma}$ to hold even if $g \neq \tilde{g}$, for example, when $g(x) = h(\tilde{g}(x))$ for some strictly increasing function h. Nevertheless, in the worst-case scenario, $\overline{\kappa}_n$ could be large enough to dominate the rate, though we do not expect such behavior in typical applications.

The rate $n^{-1/3}$ is slower than the typical rate $n^{-2/3}$ achieved in one-dimensional nonparametric regression using the fused lasso under a bounded variation assumption. However, a key distinction in our setting is that the design points are not directly observed but instead we rely on their estimates. This introduces an additional layer of complexity and variability that affects the convergence rate. We do not claim that our procedure is minimax optimal; in fact, we conjecture that it is not. Nonetheless, our method remains computationally efficient and provably consistent, making it a practical and scalable choice, as we demonstrate in Section 4.

We conclude this section with a remark that can be thought as a straightforward generalization of Theorem 2.

Remark 3 *Notice that the rate $n^{-1/3}$ does not depend on the dimension d of the covariates as it is the case of other nonparametric estimators, see Gao and Han (2020). In fact our results are not directly comparable with Gao and Han (2020) as the authors there consider different classes of functions. A main driver behind the rate $n^{-1/3}$ is Assumption 4. If instead $t^* = \max\{t_0, t_1\}$ is allowed to grow, then the upper bound in Theorem 2 should be inflated by a factor $(t^*)^{2/3}$. Hence, similar to the discussion above this would lead to the rate $(t^*)^{2/3}n^{-1/3}$.*

3.2 Main result for propensity score based estimator

We now study the statistical properties of the estimator defined in Section 2.3. Since the assumptions required to arrive at our main result here are similar to those in Section 3.1, here we only present the conclusion of our result and the assumptions are given in Section C.

Theorem 4 *Under Assumptions 1, and 7–11, $dn^{1/2} \geq C_{\min}$, $n \asymp m$, there exists $t > 0$ such that*

$$t \asymp \max\left\{\frac{dn^{1/2} \cdot \log^{1/2} n \, \log^{1/2}(nd)}{C_{\min}}, \log n\right\} \tag{20}$$

and choice of λ satisfying

$$\lambda \asymp n^{1/3}(\log n)^2(\log\log n)t^{-1/3},$$

such that the estimator $\hat{\tau}$ defined in (16) satisfies

$$\frac{1}{n}\sum_{i=1}^{n}(\rho_i^* - \hat{\tau}_i)^2 = O_{\mathbb{P}}\left\{\frac{d^{2/3}(\log n)^3(\log\log n)}{C_{\min}^{2/3} n^{1/3}}\right\}, \tag{21}$$

where $\rho_i^* = E\{Y(1) \,|\, e(X) = e(X_i), Z = 1\} - E\{Y(0) \,|\, e(X) = e(X_i), Z = 0\}$ for $i = 1, \ldots, n$. If in addition $Y(0), Y(1) \perp\!\!\!\perp Z \,|\, e(X)$, then (25) holds replacing ρ_i^* with $\tau_i^* = E\{Y(1) - Y(0) \,|\, e(X) = e(X_i)\}$ for $i = 1, \ldots, n$.

Importantly, Theorem 4 implies that the estimator $\hat{\tau}$ defined in (16) can consistently estimate the subgroup treatment effects τ^* under general conditions. One of such conditions is that $Y(0), Y(1) \perp\!\!\!\perp Z \,|\, e(X)$, which in the language of Rosenbaum and Rubin (1983) means that treatment is strongly ignorable given $e(\cdot)$. As Theorem 3 in Rosenbaum and Rubin (1983) showed, $Y(0), Y(1) \perp\!\!\!\perp Z \,|\, e(X)$ holds under overlapping (Assumption 1) and unconfoundedness which can be writen as $Y(0), Y(1) \perp\!\!\!\perp Z \,|\, e(X)$. When these conditions are violated, Theorem 4 shows that $\hat{\tau}$ can still approximate ρ^* under Assumptions 1, and 7–11.

4. Experiments

We will now validate with experiments the proposed methods in this paper. Throughout this section, we refer to the procedure in Section 2.2 as *Causal Fused Lasso 1 (CFL1)*, and the procedure in Section 2.3 as *Causal Fused Lasso 2 (CFL2)*. For both estimators, we select the tuning parameter λ by minimizing the Bayesian Information Criterion (BIC) over a grid of candidate values, as described in Section 2.2. For each λ, we compute the fused lasso estimator and evaluate BIC using the residual sum of squares and the estimated degrees of freedom following the approach of Tibshirani et al. (2012).

We benchmark our methods against several widely used baselines. These include causal random forests Procedure 1 (WA1) and Procedure 2 (WA2) from Wager and Athey (2018), the robust generalized random forest (GRF) from Section 6.2 of Athey et al. (2019), and the estimator of Abadie et al. (2018) (ACW). We also include two flexible, nonparametric methods: Bayesian Additive Regression Trees (BART) from Chipman et al. (2010a), which models the outcome as a sum of regression trees and estimates individual treatment effects as the difference in posterior mean outcomes under treatment and control; and the Augmented Inverse Probability Weighting (AIPW) estimator from Glynn and Quinn (2010), which combines outcome regression and propensity score weighting and enjoys double robustness.

To further enhance our evaluation and directly address recent developments in interpretable causal inference, we include a family of matching-based estimators motivated by the "almost exact matching" framework. These include: Genetic Matching (GM) from Diamond and Sekhon (2013), MALTS from Parikh et al. (2022), Lasso Coefficient Matching (LCM) from Lanners et al. (2023), ADD-MALTS from Katta et al. (2024), and Adaptive Hyperbox Matching (AHB) from Morucci et al. (2020). Each of these approaches defines a strategy to identify comparable units, using learned distance metrics or adaptive rules, and estimates the treatment effect for each unit by imputing one or both missing potential outcomes from matched units.

In addition, we consider two interpretable subset-based matching estimators: FLAME from Wang et al. (2021) and DAME from Liu et al. (2018). These algorithms construct matched groups by sequentially selecting subsets of covariates that optimize a trade-off between covariate balance and predictive accuracy of the outcome, resulting in interpretable, rule-based matching schemes. Unlike instance-level matching approaches such as MALTS or AHB that rely on learned distance metrics, FLAME and DAME perform combinatorial matching on covariate subsets to identify groups where units match exactly on a carefully chosen set of features. Because these methods are designed for categorical covariates, we adapt them to our continuous covariate setting by discretizing each feature into quantile-based bins prior to matching. While this preprocessing step introduces approximation error, it enables a meaningful comparison with these almost-exact matching approaches in our synthetic scenarios.

We note that CFL1 and CFL2 induce interpretable, data-driven subgroups via total variation regularization applied to estimated prognostic or propensity scores. This leads to piecewise-constant treatment effect estimates across individuals, derived not from pairwise similarity or nearest-neighbor heuristics, but from optimization principles that robustly segment units based on heterogeneity. In doing so, CFL1 and CFL2 attain the interpretability of subgroup-based methods like FLAME and DAME, while avoiding some of the limitations associated with distance-based matching, such as sensitivity to poor overlap or dependence on learned distance metrics.

4.1 Simulated and Semi-Synthetic Experiments

We assess the performance of our proposed methods across eight distinct scenarios that encompass both completely synthetic and semi-synthetic designs. Scenarios 1–6 are fully synthetic, with both covariates and outcomes generated from known models. Scenarios 7 and 8 are semi-synthetic, using covariates from real datasets (the National JTPA and Project STAR studies, respectively) and simulated outcomes as in Abadie et al. (2018). For Scenarios 1–4, we consider varying values of the sample size $n \in \{800, 1600\}$ and the covariate dimension $d \in \{2, 10\}$. For Scenarios 5 and 6, we set $n = 4000$ and $d = 10$. For each combination, we generate a dataset $\{(Z_i, X_i, Y_i)\}_{i=1}^n$ according to the corresponding generative model. In Scenarios 7 and 8, which include semi-synthetic simulations, the values

Table 1: Performance evaluations (median ± standard error) over 50 Monte Carlo simulations for synthetic scenarios with varying (n, d). **Bold** indicates the best method, and *italic* indicates the second-best.

Method	(n, d)	Scenario 1	Scenario 2	Scenario 3	Scenario 4
CFL1	(800, 2)	**0.004 ± 0.0012**	0.195 ± 0.055	0.181 ± 0.037	**0.301 ± 0.065**
	(1600, 2)	**0.003 ± 0.0009**	0.108 ± 0.032	0.136 ± 0.028	**0.183 ± 0.044**
	(800, 10)	**0.005 ± 0.0013**	**0.503 ± 0.083**	0.412 ± 0.082	**0.450 ± 0.086**
	(1600, 10)	**0.003 ± 0.0010**	**0.319 ± 0.069**	0.293 ± 0.063	**0.277 ± 0.068**
CFL2	(800, 2)	0.011 ± 0.0023	*	**0.074 ± 0.017**	*
	(1600, 2)	0.004 ± 0.0011	*	0.051 ± 0.014	*
	(800, 10)	0.016 ± 0.0038	*	**0.146 ± 0.033**	*
	(1600, 10)	0.005 ± 0.0016	*	**0.109 ± 0.027**	*
GRF	(800, 2)	0.013 ± 0.0031	**0.152 ± 0.048**	0.143 ± 0.031	1.788 ± 0.318
	(1600, 2)	0.011 ± 0.0024	**0.063 ± 0.018**	0.106 ± 0.024	0.771 ± 0.177
	(800, 10)	0.010 ± 0.0032	0.565 ± 0.092	0.408 ± 0.078	3.261 ± 0.557
	(1600, 10)	0.006 ± 0.0021	0.350 ± 0.077	0.359 ± 0.070	1.291 ± 0.308
AIPW	(800, 2)	0.015 ± 0.0040	0.235 ± 0.058	0.109 ± 0.027	1.918 ± 0.362
	(1600, 2)	0.010 ± 0.0025	0.132 ± 0.036	**0.049 ± 0.015**	1.089 ± 0.291
	(800, 10)	0.023 ± 0.0052	0.672 ± 0.112	0.342 ± 0.059	3.988 ± 0.611
	(1600, 10)	0.019 ± 0.0041	0.498 ± 0.096	0.286 ± 0.051	3.134 ± 0.552
MALTS	(800, 2)	0.018 ± 0.0042	0.248 ± 0.057	0.098 ± 0.023	2.156 ± 0.391
	(1600, 2)	0.011 ± 0.0031	0.112 ± 0.028	***0.050 ± 0.014***	1.141 ± 0.283
	(800, 10)	0.030 ± 0.0061	0.618 ± 0.102	0.238 ± 0.051	3.274 ± 0.553
	(1600, 10)	0.024 ± 0.0054	0.401 ± 0.087	0.212 ± 0.046	2.148 ± 0.472
AHB	(800, 2)	0.010 ± 0.0028	0.231 ± 0.051	0.117 ± 0.025	1.503 ± 0.297
	(1600, 2)	**0.003 ± 0.0009**	0.136 ± 0.031	0.081 ± 0.017	0.908 ± 0.239
	(800, 10)	0.013 ± 0.0034	0.492 ± 0.083	0.234 ± 0.048	2.182 ± 0.459
	(1600, 10)	0.005 ± 0.0016	0.331 ± 0.071	0.191 ± 0.041	1.523 ± 0.382

of n and d are determined by the underlying real datasets or specific design choices. In all cases, we evaluate performance using the mean squared error (MSE),

$$\frac{1}{n}\sum_{i=1}^{n}\left(\tau_i^* - \hat{\tau}_i\right)^2,$$

where $\tau_i^* = \mathbb{E}[Y_i \mid X_i, Z_i = 1] - \mathbb{E}[Y_i \mid X_i, Z_i = 0]$, and $\hat{\tau}_i$ is the estimate from a given method. MSEs are averaged over 50 Monte Carlo replications, and we report associated standard errors. Descriptions of each scenario follow.

Scenarios 1–4, which are completely synthetic, are described in Section F. The first two scenarios come from Wager and Athey (2018) and both consist of $\tau_i^* = 0$ for all $i \in \{1, \dots, n\}$. In Scenario 3, we define the treatment effect as $\tau_i^* = \mathbf{1}_{\{e(X_i)>0.6\}}$, where $e(x) = \Phi(\beta^\top x)$ is the propensity score and Φ denotes the standard normal CDF. The vector $\beta \in \mathbb{R}^d$

is fixed with $\beta_j = 1$ for $j \leq \lfloor d/2 \rfloor$ and $\beta_j = -1$ otherwise. Furthermore, Scenario 4 is the model described in (10).

Scenario 5. This fully synthetic scenario comes from Abadie et al. (2018). Setting $d = 10$ and $n = 4000$ the data is generated as: $Y_i = \ 1 + \beta^\top X_i + \epsilon_i$, $X_i \overset{\text{ind}}{\sim} \mathcal{N}(0, \mathbf{I}_{d\times d})$ and $\epsilon_i \overset{\text{ind}}{\sim} \mathcal{N}(0, 100 - d)$, where $\beta = (1, \ldots, 1)^\top \in \mathbb{R}^d$. Moreover, the treatment indicators $Z_i \in \{0, 1\}$ are assigned such that $\sum_i Z_i = \lceil n/2 \rceil$. Clearly, the vector of treatment effects satisfies $\tau^* = 0$.

Scenario 6. For our final fully synthetic model we set $d = 10$, $n = 4000$, and generate data as

$$\begin{array}{rcl} Y_i & = & (1 - Z_i)Y_i(0) + Z_i Y_i(1), \\ Z_i & \sim & \text{Binom}(1, \frac{1}{2}), \\ Y_i(l) & \sim & \mathcal{N}(f_l(X_i), 1), \ \ \forall l \in \{0, 1\}, \\ f_0(x) & = & x^\top \beta, \ \ \forall x \in [0, 1]^d, \\ f_1(x) & = & f_0(x) + \mathbf{1}_{\{x^\top \beta > 1\}} + \mathbf{1}_{\{x^\top \beta < 0.2\}}, \ \ \forall x \in [0, 1]^d, \\ X_i & \overset{\text{ind}}{\sim} & U[0, 1]^d, \ \ \forall i \{1, \ldots, n\}, \end{array}$$

where $\beta \in \mathbb{R}^p$ with $\beta_j = 1$ if $j \in \{1, \ldots, \lfloor p/2 \rfloor\}$, and $\beta_j = -1$ otherwise. Notice that in this case the treatment effect for unit i is $\tau_i^* = \mathbf{1}_{\{X_i^\top \beta > 1\}} + \mathbf{1}_{\{X_i^\top \beta < 0.2\}}$.

The final two scenarios are semi-synthetic designs, where covariates are taken from real datasets, while the outcomes are simulated according to a known data-generating process. This allows us to evaluate the estimators in realistic covariate spaces while preserving ground-truth treatment effects.

Scenario 7. We use the setting of the National JTPA Study used in Abadie et al. (2018). This consists of a National JTPA Study evaluating an employment and training program commissioned by the U.S. Department of Labor in the late 1980s. Other authors that have also analyzed this data include Orr (1996) and Bloom et al. (1997). In the JTPA study, based on a randomized assignment, subjects were a assigned into one of two groups. In the treatment group the subjects had access to JTPA services that included one of three possibilities: on-the-job training/job search assistance, classroom training, and other services. In contrast, subjects in the control group were not given access to the JTPA services. The raw data consists of 2530 units $(n_{obs} = 2530)$, 1681 of which are treated observations and 849 are untreated observations. With these measurements, we generate simulated data as in Abadie et al. (2018) where the treatment effect is zero across all units. The details are given in Appendix G.1.

Scenario 8. We also consider an example used in Abadie et al. (2018). Specifically, we use the Project STAR class-size study, see for instance Krueger (1999). In this data, 3,764 students who entered the study in kindergarten were assigned to small classes or to regular-size classes (without a teacher's aide). The outcome variable is standardized end-of-the-year kindergarten math test scores. As for covariates, some of these include race, eligibility for the free lunch program, and school attended. With the original data we proceed as in Abadie et al. (2018) and simulate data in a setting where the treatment effects are all zero. The details are given in Appendix G.2.

The results of our experiments in Scenarios 1–4 for the top six competitors are reported in Table 1. The full comparison with all methods is deferred to the appendix (see Table 4). There, we can see that for Scenario 1 the best methods are our proposed estimators, which is reasonable since the treatment effect is zero across units. AHB matches our performance at the configuration $(n, d) = (1600, 2)$ in this scenario. In Scenario 2, we do not compare CFL2 since such method is not suitable for experimental designs where the propensity score takes on a constant value. GRF achieves the best performance in low dimension ($d = 2$), closely followed by CFL1. However, in higher dimension ($d = 10$), CFL1 outperforms GRF, highlighting its robustness in more complex covariate settings.

In Scenario 3, we see that CFL2 generally outperforms the competitors. At the configuration $(n, d) = (1600, 2)$, however, AIPW achieves the best performance, closely followed by MALTS, both of which slightly outperform CFL2. Notice that in Scenario 3 the treatment effect is a function of the propensity score, where the propensity score belongs to the family of probit models. This does not seem to be a problem for our estimator which provides accurate estimation despite relying on logistic regression in the first stage. Moreover, in Scenario 4, we see that CFL1 outperforms the competitors. Again, since the propensity score is constant, we do not benchmark CFL2.

Table 2 summarizes results for Scenarios 5–8. To facilitate comparison across scenarios, we standardized the outcome variable Y in each case by subtracting the mean and dividing by the standard deviation before applying each estimator. In Scenarios 5 and 6, which are fully synthetic and involve a larger sample size of 4000 observations, CFL1 achieves the best performance, outperforming all benchmark methods by a substantial margin. These results underscore the flexibility and robustness of CFL1, particularly in large-sample settings and under varied treatment effect structures, including the presence of high noise (Scenario 5) and sharp discontinuities (Scenario 6).

In the semi-synthetic Scenarios 7 and 8, which use covariates from real datasets, CFL1 remains highly competitive. In Scenario 7 (JTPA), ACW attains the lowest MSE, followed closely by CFL1 and WA2. While ACW performs best in Scenario 7, this is likely due to the structure of the real-world covariates aligning well with its stratification procedure, rather than the constancy of the treatment effect alone. In Scenarios 1–4, despite having constant or piecewise-constant treatment effects, ACW is consistently outperformed by our methods, particularly CFL1 (see Table 4 in Appendix A). This suggests that flexible subgroup discovery and modeling heterogeneity, as in CFL1 and CFL2, are advantageous even when effects are simple.

In Scenario 8 (Project STAR), which involves a high-dimensional covariate space and a null treatment effect, CFL1 again outperforms all competitors, including ACW. These findings demonstrate that CFL1 performs reliably not only with realistic covariate distributions but also under more complex or high-dimensional conditions where accurate estimation requires stronger regularization and global modeling capabilities. Overall, our results highlight the strong empirical performance of both CFL1 and CFL2 across diverse experimental setups, with one of the two—CFL1 or CFL2—consistently ranking among the top two methods in every scenario.

4.2 National Supported Work data

4.2.1 Randomized example

To illustrate the behavior of our estimators, we use the data from LaLonde (1986); Dehejia and Wahba (1999, 2002). This dataset consists of a 445 sub-sample from the National

Table 2: Performance evaluations (median ± standard error) over 50 Monte Carlo simulations for synthetic (Scenarios 5–6) and semi-synthetic (Scenarios 7–8) setups. **Bold** indicates the best method, and *italic* indicates the second-best. For Scenarios 5–6, $(n, d) = (4000, 10)$; for Scenario 7, $(n, d) = (3764, 79)$; and for Scenario 8, $(n, d) = (2530, 18)$.

Method	Scenario 5	Scenario 6	Scenario 7	Scenario 8
CFL1	**0.078 ± 0.017**	**0.074 ± 0.015**	***0.042 ± 0.008***	**0.308 ± 0.071**
WA1	0.489 ± 0.077	0.423 ± 0.072	0.103 ± 0.020	0.592 ± 0.089
WA2	0.352 ± 0.066	0.295 ± 0.057	0.044 ± 0.010	0.435 ± 0.075
GRF	0.319 ± 0.058	0.204 ± 0.041	0.045 ± 0.012	0.382 ± 0.068
ACW	0.462 ± 0.070	***0.131 ± 0.030***	**0.015 ± 0.004**	0.501 ± 0.086
BART	***0.248 ± 0.044***	0.166 ± 0.037	0.057 ± 0.011	***0.329 ± 0.069***
AIPW	0.278 ± 0.051	0.149 ± 0.035	0.044 ± 0.010	0.399 ± 0.073
GM	0.376 ± 0.061	0.231 ± 0.048	0.061 ± 0.013	0.449 ± 0.078
MALTS	0.341 ± 0.059	0.188 ± 0.039	0.065 ± 0.013	0.371 ± 0.072
LCM	0.302 ± 0.056	0.226 ± 0.043	0.071 ± 0.014	0.401 ± 0.079
ADD-MALTS	0.265 ± 0.050	0.132 ± 0.031	0.068 ± 0.013	0.343 ± 0.071
AHB	0.312 ± 0.055	0.161 ± 0.034	0.066 ± 0.012	0.361 ± 0.073
FLAME	0.407 ± 0.066	0.273 ± 0.049	0.073 ± 0.014	0.371 ± 0.075
DAME	0.331 ± 0.057	0.211 ± 0.045	0.060 ± 0.012	0.355 ± 0.072

Supported Work Demonstration (NSW). The NSW was a program implemented in the mid-1970s in which the treatment group consisted of randomly selected subjects to gain 12 to 18 months of work experience and 260 subjects in a control group. The response variable is the post-study earnings in 1978. The predictor variables include age, education, indicator of Black and Hispanic for race, marital status, high-school degree indicator, earnings in 1974, and earnings in 1975.

To construct our estimator, we first estimate the prognostic score using the data from the control group and running a linear regression model. With the prognostic scores, we then compute an ordering and run the fused lasso leading to our CFL1 estimator. This is depicted in Figure 2. There, we also see the estimates based on the method ACW from Abadie et al. (2018). Both CFL1 and ACW estimate small positive treatment effects, which is consistent with expectations given the nature of the NSW program—participants in the treatment group would be expected to benefit, in terms of future earnings, from the work experience they received. Interestingly, from an interpretability perspective, CFL1 estimates

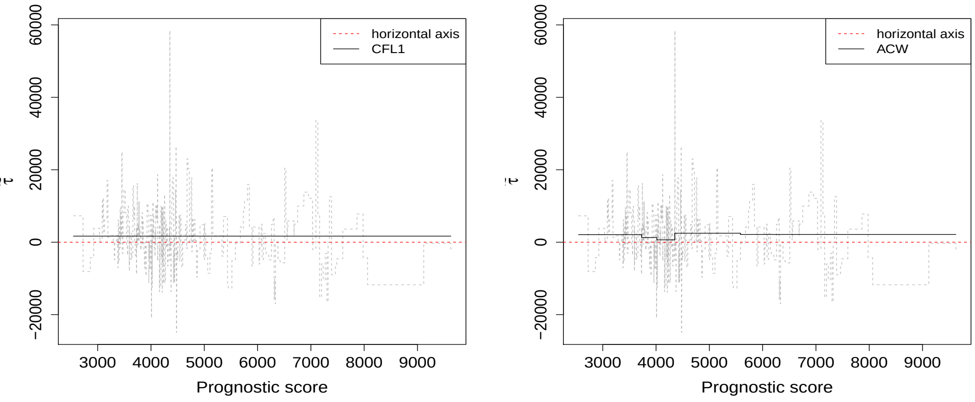

Figure 2: For the NSW example from left to right the two panels show the treatment effect estimates based on causal fused lasso with prognostic score (CFL1) and the ACW method from Abadie et al. (2018).

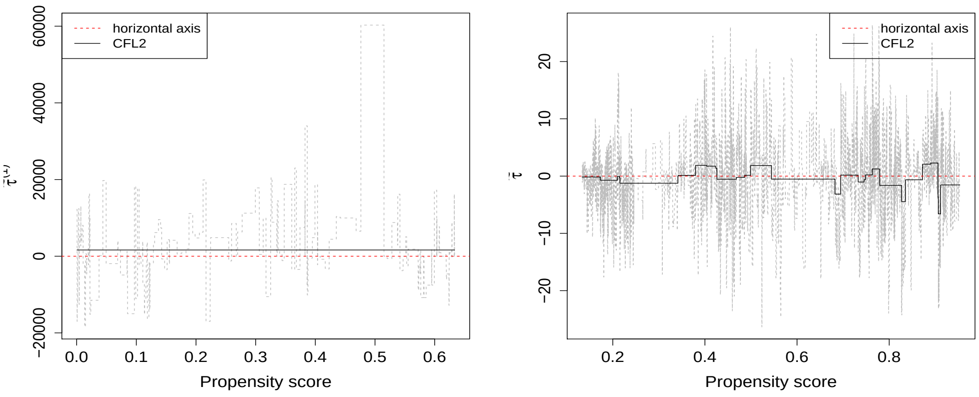

Figure 3: The panel on the left shows the estimated treatment effects on the *treated* based on the CFL2 estimator for the NSW observational data. The right panel then shows the corresponding treatment effect estimates of CFL2 for the NHANES data.

a single group effect, suggesting little to no meaningful heterogeneity in treatment effects. This finding is broadly in line with ACW, which reports slightly varying effects across four prespecified groups—groups that are not derived from the data but fixed in advance. This contrast might highlight a key strength of CFL1: its ability to adaptively detect underlying structure (or confirm its absence) directly from the data, rather than relying on predefined subgroup classifications.

4.2.2 Observational example

For our second example based on the NSW data, we combine the 185 observations in the treatment group of the data from Section 4.2.1 with the largest of the six observational control groups constructed by Lalonde[1]. This results in a total of 16177 samples. Due to the observational nature of the dataset, we run our propensity score based estimator from Section 2.3 by only estimating treatment effects on the treated. Thus, our estimator is the one described in Corollary 6 which we denote as CFL2. As shown in Figure 3, CFL2 estimates a constant positive treatment effect, consistent with our findings in Section 4.2.1. Specifically, CFL2 again estimates a single group effect, suggesting that there may be little to no heterogeneity in treatment effects—though a small positive effect is still detected.

4.3 National Health and Nutrition Examination Survey

In our final example, we use data from the 2007–2008 National Health and Nutrition Examination Survey (NHANES). The data consist of 2330 children and their participation in the National School Lunch or the School Breakfast programs in order to assess the effectiveness of such meal programs in increasing body mass index (BMI). In the study 1284 randomly selected children participated in the meal programs while 1046 did not. The predictor variables are age, gender, age of adult respondent, and categorical variables such as Black race, Hispanic race, whether the family of the child is above 200% of the federal poverty level, participation in Special Supplemental Nutrition program, Participation in food stamp program, childhood food security, any type of insurance, and gender of the adult respondent.

Similarly, as before, we run our propensity score based estimator (CFL2) and show the results in Figure 3. We can see that the sign of estimated treatment effects varies depending on the value of the propensity score. The latter was estimated by logistic regression. Our findings for the treatment effects coincide with several authors who found positive and negative average treatment effects as discussed in Chan et al. (2016). In particular, we find that when the estimated propensity score is low (below 0.34), the treatment effect is predominantly negative. This suggests that individuals who are unlikely to participate in the program may be more prone to experiencing adverse effects. Further inspection of the data reveals that all individuals with propensity scores below 0.34 come from families with incomes above 200% of the federal poverty level.

For propensity scores in the range of 0.34 to 0.68, the estimated treatment effects are generally small or slightly positive, indicating that the meal programs may offer modest benefits to individuals with a moderate likelihood of participation.

In contrast, for subjects with higher propensity scores, the estimated treatment effects display greater heterogeneity. While some subgroups exhibit small positive effects, others show negative ones. Notably, the largest positively affected subgroup falls within the 0.87 to 0.91 propensity score range. However, individuals with propensity scores above 0.91 consistently exhibit negative treatment effects. Examining the data, we find that all individuals with propensity scores above 0.87 fall below the 200% poverty threshold. Among those in the 0.87–0.91 range, only 13% experienced childhood food insecurity, whereas 66% of those above 0.91 did.

[1] Dataset is available here http://users.nber.org/ rdehejia/nswdata2.html

Overall, this analysis highlights the heterogeneous nature of the data and the value of our method in identifying subgroups with differing treatment responses.

4.4 Right Heart Catheterization (RHC) Study

We now consider a clinical example drawn from an observational study examining the effects of right heart catheterization (RHC) on short-term survival outcomes in critically ill patients admitted to an intensive care unit (ICU). RHC is a diagnostic procedure used to assess cardiac function by directly measuring pulmonary pressures and cardiac output. While it can inform treatment decisions, it also poses non-negligible procedural risks, and its clinical benefit has been the subject of ongoing debate. Due to the ethical and logistical challenges of conducting randomized trials in this setting, observational data has been widely used to evaluate the causal effect of RHC on mortality outcomes.

The data consists of $2,707$ patients, with $1,103$ receiving RHC treatment within the first 24 hours of admission ($Z = 1$), and $1,604$ not receiving the procedure ($Z = 0$). The binary outcome Y indicates whether the patient died within 180 days of hospital admission. Each patient is characterized by a set of 72 covariates, which include continuous measurements, binary indicators, and dummy variables derived from categorical attributes. These variables capture demographic, clinical, and treatment-related characteristics relevant to patient prognosis.

We apply the propensity score-based fused lasso estimator CFL2 as defined in Section 2.3 (Equation (16)), focusing on the estimation of treatment effects for the treated population. Following standard preprocessing steps from prior clinical studies, we first exclude all patient records containing missing covariate values. Among the remaining 2,707 individuals, we drop any covariates with no variability and transform all categorical variables into dummy variables using one-hot encoding. This results in a covariate matrix with 72 features, including continuous, binary, and dummy-encoded categorical variables. We then fit a logistic regression model to estimate propensity scores using the full set of encoded covariates as predictors.

Figure 4 displays the estimated treatment effects across the range of estimated propensity scores using the CFL2 estimator. We observe that for most strata of the population, identified by similar estimated propensity scores, the estimated effect of RHC is either negative or close to zero. This suggests that RHC may offer limited or no survival benefit for the majority of patients, and may even be associated with worse short-term outcomes in some subgroups. These findings support the view that the use of RHC should be carefully evaluated on a patient-specific basis. The estimated step function highlights the ability of CFL2 to adaptively identify subgroups with differing treatment responses without imposing pre-specified strata. This interpretation is in line with prior work such as Abadie et al. (2018), which considers population partitions into a small number of subgroups with distinct average treatment effects. Our estimator adaptively detects such subgroup heterogeneity from data without requiring these partitions to be specified in advance. The presence of clear jumps in the estimated treatment effect across strata (as shown in Figure 4) aligns with this interpretation and illustrates the ability of CFL2 to flexibly model heterogeneity in observational settings.

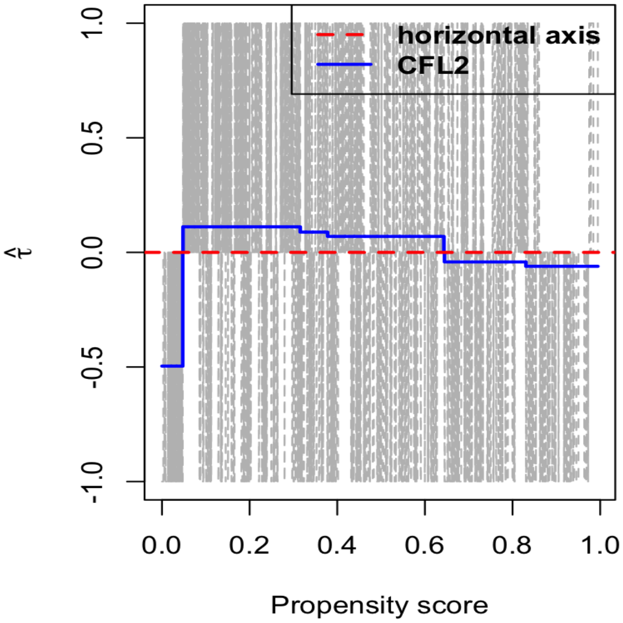

Figure 4: Estimated treatment effects across the range of estimated propensity scores using CFL2. The stepwise structure of the estimated effects highlights heterogeneous responses to RHC treatment.

5. Conclusion

In this paper, we studied two methods for estimating heterogeneous treatment effects. The first approach, based on the prognostic score, is designed for randomized experiments. It involves constructing a prognostic score and then applying the fused lasso, using as input a noisy estimate of the treatment effect derived from matching, with the prognostic score serving as the covariate.

The second approach, which relies on the propensity score, is suitable for observational studies. It follows the same general structure as the first method, but substitutes the propensity score in place of the prognostic score.

A key strength of both methods lies in their simplicity and their usefulness as exploratory tools. For each, we provide theoretical guarantees in the form of finite-sample bounds on the mean squared error for estimating heterogeneous treatment effects. However, as with much of the existing literature on the fused lasso, our methods do not currently offer confidence bands with theoretical guarantees. One promising direction for addressing this limitation is to explore residual bootstrap techniques (Efron, 1992), potentially incorporating ideas from the nonparametric approach in Padilla et al. (2024). We leave this extension for future research.

On a related note, since both of our methods accommodate treatment effects that may exhibit discontinuities as a function of the score, it is natural to consider applications where such discontinuities arise organically. As one reviewer suggested, this setting might be particularly relevant in the social sciences (Wong, 2010), where individuals may become eligible for various forms of assistance or educational programs upon exceeding specific thresholds—such as standardized test scores or income cutoffs.

Finally, another promising direction is to extend the results of Theorems 2 and 4 to out-of-sample settings. Specifically, for Theorem 2, suppose X_{n+1} is drawn independently from the same distribution as $X_1, \ldots, X_n$. We conjecture that it is possible to derive an upper bound on

$$\mathbb{E}((\rho^*_{n+1} - \hat{\tau}_{n+1})^2) \tag{22}$$

where $\rho^*_{n+1} = E\{Y(1)|g(X) = g(X_{n+1}), Z = 1\} - E\{Y(0)|g(X) = g(X_{n+1}), Z = 0\}$, and $\hat{\tau}_{n+1}$ is constructed via interpolation6y using the trained estimators $\hat{\tau}_1, \ldots, \hat{\tau}_n$. We expect that the upper bound for the quantity in (22) will match, up to logarithmic factors, the upper bound in (19). However, a formal proof of this result is nontrivial and is left for future work. We also conjecture that an analogous result may hold for Theorem 4, replacing the prognostic score with the propensity score.

Appendix A. Additional numerical results

In this appendix, we present an additional synthetic experiment, labeled *Scenario 9.* This data-generating process is designed to challenge estimators with both strong nonlinearity in the outcome functions and treatment effect heterogeneity that interacts with the covariates in a complex, nonlinear fashion. In particular, it combines nonlinear transformations, sinusoidal interactions, and a nontrivial treatment assignment mechanism driven by a noisy logistic function of covariates.

Table 3: Performance evaluations (median ± standard error) over 50 Monte Carlo simulations for the additional nonlinear synthetic Scenario 9. $(n, d) = (4000, 10)$. **Bold** indicates the best method, and *italic* indicates the second-best.

Method	Scenario 9
CFL1	0.267 ± 0.046
CFL2	***0.189 ± 0.039***
WA1	0.395 ± 0.063
WA2	0.319 ± 0.052
GRF	0.248 ± 0.045
ACW	0.401 ± 0.066
BART	**0.174 ± 0.036**
AIPW	0.305 ± 0.050
GM	0.362 ± 0.058
MALTS	0.331 ± 0.054
LCM	0.352 ± 0.057
ADD-MALTS	0.201 ± 0.041
AHB	0.369 ± 0.059
FLAME	0.402 ± 0.063
DAME	0.384 ± 0.061

Scenario 9. This synthetic is inspired by nonlinear regression examples commonly used in benchmarking flexible estimators. The data is generated as follows, for $(n, d) = (4000, 10)$:

$$
\begin{aligned}
&x_{i,1}, \dots, x_{i,10} \overset{\text{iid}}{\sim} \mathcal{U}(0,1), \quad \epsilon_{i,(0)}, \epsilon_{i,(1)}, \epsilon_{i,(\text{treat})} \overset{\text{iid}}{\sim} \mathcal{N}(0,1),\\
&Y_i(0) = 10 \sin(\pi x_{i,1} x_{i,2}) + 20(x_{i,3} - 0.5)^2 + 10 x_{i,4} + 5 x_{i,5} + \epsilon_{i,(0)},\\
&Y_i(1) = Y_i(0) + x_{i,3} \cos(\pi x_{i,1} x_{i,2}) + \epsilon_{i,(1)},\\
&Z_i = \mathbf{1}_{\left\{\text{expit}(x_{i,1} + x_{i,2} - 0.5 + \epsilon_{i,(\text{treat})}) > 0.5\right\}},\\
&Y_i = (1 - Z_i) Y_i(0) + Z_i Y_i(1),
\end{aligned}
$$

where $\text{expit}(u) = 1/(1 + e^{-u})$ denotes the logistic sigmoid function.

This setup introduces modeling difficulties not only through the nonlinearity of the outcome regression, but also by encoding treatment effects that depend on both the covariates and their interactions. As shown in Table 3, even in this challenging setting, our proposed estimator CFL2 performs competitively and achieves the second-best performance overall, closely trailing BART. Notably, CFL2 outperforms all other benchmark methods by a clear margin. This result highlights the flexibility and generalization strength of our proposed method under complex nonlinear conditions.

For completeness, Table 4 in the appendix summarizes the performance of all estimators across Scenarios 1–4.

Table 4: Performance evaluation (median ± standard error) for synthetic scenarios with varying (n, d). **Bold** indicates the best, and *italic* indicates the second-best.

Method	(n, d)	Scenario 1	Scenario 2	Scenario 3	Scenario 4
CFL1	(800, 2)	**0.004 ± 0.0012**	0.195 ± 0.055	0.181 ± 0.037	**0.301 ± 0.065**
	(1600, 2)	**0.003 ± 0.0009**	0.108 ± 0.032	0.136 ± 0.028	**0.183 ± 0.044**
	(800, 10)	**0.005 ± 0.0013**	**0.503 ± 0.083**	0.412 ± 0.082	**0.450 ± 0.086**
	(1600, 10)	**0.003 ± 0.0010**	**0.319 ± 0.069**	0.293 ± 0.063	**0.277 ± 0.068**
CFL2	(800, 2)	0.011 ± 0.0023	*	**0.074 ± 0.017**	*
	(1600, 2)	0.004 ± 0.0011	*	0.051 ± 0.014	*
	(800, 10)	0.016 ± 0.0038	*	**0.146 ± 0.033**	*
	(1600, 10)	0.005 ± 0.0016	*	**0.109 ± 0.027**	*
WA1	(800, 2)	0.045 ± 0.011	0.495 ± 0.104	0.226 ± 0.044	4.923 ± 0.819
	(1600, 2)	0.029 ± 0.009	0.199 ± 0.057	0.164 ± 0.031	3.228 ± 0.607
	(800, 10)	0.067 ± 0.013	0.773 ± 0.139	0.550 ± 0.089	6.528 ± 0.908
	(1600, 10)	0.068 ± 0.012	0.534 ± 0.127	0.499 ± 0.086	6.146 ± 0.875
WA2	(800, 2)	0.012 ± 0.0040	0.264 ± 0.069	0.143 ± 0.029	2.922 ± 0.486
	(1600, 2)	0.010 ± 0.0026	0.164 ± 0.043	0.106 ± 0.023	2.049 ± 0.417
	(800, 10)	0.007 ± 0.0021	0.794 ± 0.147	0.362 ± 0.063	5.796 ± 0.768
	(1600, 10)	0.003 ± 0.0015	0.723 ± 0.116	0.316 ± 0.059	5.322 ± 0.736
GRF	(800, 2)	0.013 ± 0.0031	**0.152 ± 0.048**	0.143 ± 0.031	1.788 ± 0.318
	(1600, 2)	0.011 ± 0.0024	**0.063 ± 0.018**	0.106 ± 0.024	0.771 ± 0.177
	(800, 10)	0.010 ± 0.0032	0.565 ± 0.092	0.408 ± 0.078	3.261 ± 0.557
	(1600, 10)	0.006 ± 0.0021	0.350 ± 0.077	0.359 ± 0.070	1.291 ± 0.308
ACW	(800, 2)	0.017 ± 0.0034	0.402 ± 0.088	0.213 ± 0.043	2.203 ± 0.378
	(1600, 2)	0.014 ± 0.0025	0.297 ± 0.067	0.181 ± 0.037	1.654 ± 0.324
	(800, 10)	0.029 ± 0.0056	0.638 ± 0.096	0.489 ± 0.081	4.394 ± 0.672
	(1600, 10)	0.024 ± 0.0049	0.412 ± 0.084	0.443 ± 0.075	3.612 ± 0.528
BART	(800, 2)	0.009 ± 0.0026	0.184 ± 0.052	0.102 ± 0.026	1.842 ± 0.334
	(1600, 2)	0.006 ± 0.0017	0.095 ± 0.027	0.075 ± 0.019	1.021 ± 0.266
	(800, 10)	0.021 ± 0.0048	0.592 ± 0.096	0.351 ± 0.068	3.482 ± 0.562
	(1600, 10)	0.015 ± 0.0035	0.412 ± 0.085	0.306 ± 0.060	2.693 ± 0.504
AIPW	(800, 2)	0.015 ± 0.0040	0.235 ± 0.058	0.109 ± 0.027	1.918 ± 0.362
	(1600, 2)	0.010 ± 0.0025	0.132 ± 0.036	**0.049 ± 0.015**	1.089 ± 0.291
	(800, 10)	0.023 ± 0.0052	0.672 ± 0.112	0.342 ± 0.059	3.988 ± 0.611
	(1600, 10)	0.019 ± 0.0041	0.498 ± 0.096	0.286 ± 0.051	3.134 ± 0.552
GM	(800, 2)	0.020 ± 0.0051	0.300 ± 0.064	0.202 ± 0.041	2.421 ± 0.394
	(1600, 2)	0.005 ± 0.0013	0.187 ± 0.048	0.072 ± 0.020	1.618 ± 0.351
	(800, 10)	0.035 ± 0.0067	0.689 ± 0.103	0.489 ± 0.075	5.088 ± 0.779
	(1600, 10)	0.0062 ± 0.0015	0.522 ± 0.091	0.088 ± 0.023	4.523 ± 0.721
MALTS	(800, 2)	0.018 ± 0.0042	0.248 ± 0.057	0.098 ± 0.023	2.156 ± 0.391
	(1600, 2)	0.011 ± 0.0031	0.112 ± 0.028	***0.050 ± 0.014***	1.141 ± 0.283
	(800, 10)	0.030 ± 0.0061	0.618 ± 0.102	0.238 ± 0.051	3.274 ± 0.553
	(1600, 10)	0.024 ± 0.0054	0.401 ± 0.087	0.212 ± 0.046	2.148 ± 0.472
LCM	(800, 2)	0.014 ± 0.0038	0.248 ± 0.051	0.129 ± 0.028	1.697 ± 0.323
	(1600, 2)	0.0042 ± 0.0012	0.157 ± 0.035	0.069 ± 0.018	0.996 ± 0.262
	(800, 10)	0.019 ± 0.0043	0.562 ± 0.089	0.278 ± 0.054	2.643 ± 0.497
	(1600, 10)	0.0055 ± 0.0014	0.384 ± 0.077	0.081 ± 0.021	1.839 ± 0.421
ADD-MALTS	(800, 2)	0.015 ± 0.0040	0.276 ± 0.059	0.122 ± 0.026	1.823 ± 0.332
	(1600, 2)	0.0045 ± 0.0011	0.169 ± 0.036	0.063 ± 0.017	1.184 ± 0.271
	(800, 10)	0.017 ± 0.0038	0.588 ± 0.095	0.256 ± 0.051	2.392 ± 0.474
	(1600, 10)	0.0049 ± 0.0013	0.392 ± 0.081	0.072 ± 0.019	1.601 ± 0.409
AHB	(800, 2)	0.010 ± 0.0028	0.231 ± 0.051	0.117 ± 0.025	1.503 ± 0.297
	(1600, 2)	**0.003 ± 0.0009**	0.136 ± 0.031	0.081 ± 0.017	0.908 ± 0.239
	(800, 10)	0.013 ± 0.0034	0.492 ± 0.083	0.234 ± 0.048	2.182 ± 0.459
	(1600, 10)	0.005 ± 0.0016	0.331 ± 0.071	0.191 ± 0.041	1.523 ± 0.382
FLAME	(800, 2)	0.041 ± 0.009	0.298 ± 0.069	0.319 ± 0.062	4.121 ± 0.754
	(1600, 2)	0.030 ± 0.008	0.219 ± 0.052	0.271 ± 0.056	3.589 ± 0.645
	(800, 10)	0.064 ± 0.012	0.743 ± 0.135	0.613 ± 0.089	6.621 ± 0.922
	(1600, 10)	0.060 ± 0.011	0.701 ± 0.127	0.538 ± 0.082	6.179 ± 0.832
DAME	(800, 2)	0.043 ± 0.010	0.334 ± 0.074	0.307 ± 0.058	4.308 ± 0.737
	(1600, 2)	0.038 ± 0.009	0.293 ± 0.069	0.278 ± 0.054	3.945 ± 0.692
	(800, 10)	0.072 ± 0.014	0.803 ± 0.141	0.622 ± 0.092	6.734 ± 0.941
	(1600, 10)	0.075 ± 0.013	0.762 ± 0.134	0.577 ± 0.088	6.502 ± 0.910

Appendix B. Possible extensions

A natural extension of the estimators described in Sections 2.2 and 2.3 is to consider the case where the number of covariates can be large, perhaps $d >> n$, but only a small number of them plays a role in the prognostic score (propensity score). In the case of the prognostic score based estimator, it is reasonable to estimate g with lasso regression (Tibshirani, 1996). The resulting procedure would be the same as in Section 2.2, except that we would define $\hat{g}(x) = x^\top \hat{\theta}$ where

$$\hat{\theta} \,=\, \underset{\theta \in \mathbb{R}^d}{\arg\min} \left\{ \frac{1}{m} \sum_{i=1}^{m} \left(Y_i' - X_i'^\top \theta \right)^2 + \nu \sum_{j=1}^{d} |\theta_j| \right\},$$

for a tuning parameter $\nu > 0$. Similarly, we can modify the propensity score estimator, replacing $\hat{e}$ by ℓ_1-regularized logistic regression in the spirit of Ravikumar et al. (2010).

A simpler modification of the estimators from Sections 2.2–2.3 can be obtained by adding a sparsity penalty in the objective function. This is similar to the definition of the fused lasso in Tibshirani et al. (2005). The resulting estimator would be reasonable if there is a belief that most of the treatment effects are zero. It would basically amount to apply soft-thresholding to the estimators from (9)–(16), see for instance Wang et al. (2016).

While the definition of our estimators naturally extends to high-dimensional settings, the more challenging task lies in analyzing their statistical properties. Our current theoretical results are limited to scenarios where the fused lasso is combined with parametric estimation of the prognostic or propensity score, which are tailored for low-dimensional covariate spaces. Extending the theory to high-dimensional contexts remains an important direction for future work.

Appendix C. Main result for propensity score based estimator

We now study the statistical properties of the estimator defined in Section 2.3. As in Section 3.2 we start by stating required assumptions.

Assumption 7 (Sub-Gaussian errors) *Define $V(z,x) = E\{Y|Z = z, e(X) = e(x)\}$ and $\epsilon_i = Y_i - V(Z_i, X_i)$ for $i = 1, \ldots, n$, with $e(\cdot)$ as in Assumption 1. Then the vector $(\epsilon_1, \ldots, \epsilon_n)^\top$ has independent coordinates that are mean zero sub-Gaussian(v) for some constant $v > 0$.*

Assumption 7 parallels of Assumption 3 when we replace the prognostic score with the propensity score.

Assumption 8 *The functions $f_1(s) = E\{Y|Z = 1,\, e(X) = s\}$, and $f_0(s) = E\{Y|Z = 0,\, e(X) = s\}$ for $s \in [e_{\min}, e_{\max}]$ are bounded and have bounded variation.*

As for the distribution of the covariates, we allow for more generality than in Section 3.2. Specifically, we allow for general multivariate sub-Gaussian distributions.

Assumption 9 (Distribution of covariates) *The random vector $X \in \mathbb{R}^d$ is centered ($E(X) = 0$) sub-Gaussian(C).*

We refer the reader to Vershynin (2010) which contains important concentrations results regarding multivariate sub-Gaussian distributions.

Assumption 10 *The propensity score staisfies $e(X) := F(X^\top \theta^*)$ for some $\theta^* \in \mathbb{R}^d$, where $F(x) = \exp(x)/\{1+\exp(x)\}$ as before. Furthermore $e(X)$ is a continuous random variable with pdf $h(\cdot)$ bounded by above ($\|h\|_\infty = h_{\max}$ for some positive constant $h_{\max}$), and there exist constants a_1 and a_2 such that*

$$a_1 t \;\leq\; \mathbb{P}(|e(X) - b| \leq t) \;\leq\; a_2 t$$

for all b in the support of $e(X)$ and $t \in (0, t_0)$, where $t_0 > 0$ is a constant.

Assumption 11 (Dependency condition) *Let $\Lambda_{\min}(\cdot)$ and $\Lambda_{\max}(\cdot)$ be the minimum and maximum eigenvalue functions, respectively. We assume that there exist positive $C_{\min}$ and $D_{\max}$ such that*

$$\Lambda_{\min}\left(E\left[\{\eta(X^\top\theta^*)XX^\top\}\right]\right) \;>\; C_{\min} > c_1 \|X\|_{\psi_2}^2 \left(\frac{d \log m}{m}\right)^{1/2},$$

and

$$\Lambda_{\max}\{E(XX^\top)\} \;<\; D_{\max},$$

with $\eta(t) = F(t)\{1 - F(t)\}$ with F as in Assumption 10, and where $c_1 > 0$ is a constant. We also require that

$$\frac{C_{\min}^2}{D_{\max}} > c_2 \frac{d \log m}{\sqrt{m}}, \tag{23}$$

for a large enough constant $c_2 > 0$.

Assumption 11 is basically the Dependency condition in the analysis of high-dimensional logistic regression from Ravikumar et al. (2010). As the authors there assert, this condition prevents the covariates from becoming overly dependent.

We are now in position to present the main result regarding our propensity scored based estimator.

Theorem 5 *Under Assumptions 1, and 7–11, $dn^{1/2} \geq C_{\min}$, $n \asymp m$, there exists $t > 0$ such that*

$$t \;\asymp\; \max\left\{\frac{dn}{C_{\min}} \frac{\log^{1/2} m \, \log^{1/2}(nd)}{m^{1/2}}, \log n\right\} \tag{24}$$

and choice of λ satisfying

$$\lambda \;\asymp\; n^{1/3} (\log n)^2 (\log\log n) t^{-1/3},$$

such that the estimator $\hat{\tau}$ defined in (16) satisfies

$$\frac{1}{n}\sum_{i=1}^{n} (\rho_i^* - \hat{\tau}_i)^2 \;=\; O_{\mathbb{P}}\left\{\frac{d^{2/3} (\log n)^3 (\log\log n)}{C_{\min}^{2/3} n^{1/3}}\right\}, \tag{25}$$

where $\rho_i^* = E\{Y(1) \,|\, e(X) = e(X_i), Z = 1\} - E\{Y(0) \,|\, e(X) = e(X_i), Z = 0\}$ for $i = 1, \ldots, n$. If in addition $Y(0), Y(1) \perp\!\!\!\perp Z \,|\, e(X)$, then (25) holds replacing ρ_i^* with $\tau_i^* = E\{Y(1) - Y(0) \,|\, e(X) = e(X_i)\}$ for $i = 1, \ldots, n$.

Importantly, Theorem 4 implies that the estimator $\hat{\tau}$ defined in (16) can consistently estimate the subgroup treatment effects τ^* under general conditions. One of such conditions is that $Y(0), Y(1) \perp\!\!\!\perp Z \,|\, e(X)$, which in the language of Rosenbaum and Rubin (1983) means that treatment is strongly ignorable given $e(\cdot)$. As Theorem 3 in Rosenbaum and Rubin (1983) showed, $Y(0), Y(1) \perp\!\!\!\perp Z \,|\, e(X)$ holds under overlapping (Assumption 1) and unconfoundedness which can be written as $Y(0), Y(1) \perp\!\!\!\perp Z \,|\, X$. When these conditions are violated, Theorem 4 shows that $\hat{\tau}$ can still approximate ρ^* under Assumptions 1, and 7–11.

Furthermore, as in Remark 3, Theorem 4 can be relaxed. Specifically, we can replace Assumption 8 with

$$t^* \,:=\, \max\{\mathrm{TV}(f_0, n), \mathrm{TV}(f_1, n)\}.$$

Then the upper bound in Theorem 4 needs to be inflated by $(t^*)^{2/3}$.

We conclude this section with immediate consequence of the proof of Theorem 4 concerning heterogenous treatment effects of the treated units.

Corollary 6 (Treatment effects of the treated) *Suppose that the conditions of Theorem 4 for (25) to hold are met. Let $\hat{\tau}$ be the propensity score estimator from Section 2.3 with a slight modification. After the matching is done and the signal $Y - \tilde{Y}$ is calculated, we only run the the fused lasso estimator, with the ordering based on the estimated propensity score, on the treated units. Then*

$$\frac{1}{n} \sum_{i\,:\,Z_i=1} (\rho_i^* - \hat{\tau}_i)^2 \,=\, O_{\mathbb{P}} \left\{ \frac{d^{2/3} (\log n)^3 (\log\log n)}{C_{\min}^{2/3} n^{1/3}} \right\}, \tag{26}$$

where $\rho_i^* \,=\, E\{Y(1) \,|\, e(X) \,=\, e(X_i), Z \,=\, 1\} - E\{Y(0) \,|\, e(X) \,=\, e(X_i), Z \,=\, 0\}$ for $i = 1, \ldots, n$. If in addition $Y(0) \perp\!\!\!\perp Z \,|\, e(X)$, then (26) holds replacing ρ_i^* with $\tau_i^* = E\{Y(1) - Y(0) \,|\, e(X) = e(X_i),\, Z = 1\}$, for $i = 1, \ldots, n$.

Appendix D. Proof of Theorem 4

Throughout this section we write

$$\hat{L}(\theta) = -\sum_{i=1}^{m} \frac{1}{m} \left[Z_i' \log F(X_i'^{\top}\theta) + (1 - Z_i') \log\{1 - F(X_i'^{\top}\theta)\} \right],$$

and

$$\delta \,:=\, \frac{d}{C_{\min}} \frac{(\log^{1/2} m) \log^{1/2}(nd)}{m^{1/2}}.$$

We also define the first order matrix $\Delta^{(1)} \in \mathbb{R}^{(n-1)\times n}$, such that for any $b \in \mathbb{R}^n$, the following holds:

$$\|\Delta^{(1)} b\|_1 \,=\, \sum_{i=1}^{n-1} |(\Delta^{(1)} b)_i| \,=\, \sum_{i=1}^{n-1} |b_i - b_{i+1}|.$$

Hence, with this notation, the estimator defined in (16) becomes

$$\hat{\tau} \,=\, \underset{b \in \mathbb{R}^n}{\arg\min} \left\{ \frac{1}{2} \sum_{i=1}^{n} (Y_i - \widetilde{Y}_i + (-1)^{Z_i} b_i)^2 \,+\, \lambda \|\Delta^{(1)} \hat{P} b\|_1 \right\}. \tag{27}$$

D.1 Total variation auxiliary lemmas

Lemma 7 *The estimator $\hat{\tau}$ defined in (16) satisfies*

$$\|\hat{\tau}\|_\infty \;=\; O_{\mathbb{P}}\left(\max\{\|f_0\|_\infty, \|f_1\|_\infty\} + \lambda + \log^{1/2} n\right).$$

Proof We beging by introducing some notation. For a vector $x \in \mathbb{R}^s$, a vector $\text{sign}(x) \in \mathbb{R}^s$ is defined as

$$(\text{sign}(x))_i \;=\; \begin{cases} 1 & \text{if } \; x_i > 0 \\ -1 & \text{if } \; x_i < 0 \\ \in [0,1] & \text{otherwise.} \end{cases}$$

To proceed, we first condition on X and Z. Then, using Equation (27) along with the KKT conditions, we obtain:

$$Y - \widetilde{Y} + (-1)^Z \circ \hat{\tau} + \lambda \hat{P}^T (\Delta^{(1)})^\top \text{sign}(\Delta^{(1)} \hat{P} \hat{\tau}) \;=\; 0, \tag{28}$$

where $\circ$ is the Hadamard product. Next, notice that (28) implies that

$$\begin{aligned} \|\hat{\tau}\|_\infty \;&\leq\; 2\|Y\|_\infty + 2\lambda \\ &\leq\; 2\max\{\|f_0\|_\infty, \|f_1\|_\infty\} + 2\lambda + 2\|\epsilon\|_\infty. \end{aligned}$$

The claim then follows since $\|\epsilon\|_\infty \;=\; O_{\mathbb{P}}(\log^{1/2} n)$, by Sub-Gaussian tail inequality, and integrating over X and Z. ■

Lemma 8 *Assumptions 1 and 7 imply that*

$$\left|\frac{1}{n^{1/2}} \sum_{i=1}^n (-1)^{Z_i+1} \epsilon_{N(i)}\right|^2 \;=\; O_{\mathbb{P}}(\log^3 n).$$

Proof Let $S_i = \{j \in \{1,\ldots,n\} \,:\, N(j) = i\}$, we have

$$\left|\frac{1}{n^{1/2}} \sum_{i=1}^n (-1)^{Z_i+1} \epsilon_{N(i)}\right|^2 \;=\; \left|\frac{1}{n^{1/2}} \sum_{i=1}^n \left\{ \sum_{j \in S_i} (-1)^{Z_j+1} \right\} \epsilon_i \right|^2,$$

and

$$\frac{1}{n} \sum_{i=1}^n \left\{ \sum_{j \in S_i} (-1)^{Z_j+1} \right\}^2 \;\leq\; \left(\max_{i=1,\ldots,n} |S_i| \right)^2.$$

On the other hand, since $\hat{e}(X_1), \ldots, \hat{e}(X_n)$ are independent and identically distributed, we have that

$$(|S_1|, \ldots, |S_n|) \;\sim\; \text{Multinomial}\left(n; \frac{1}{n}, \ldots, \frac{1}{n}\right).$$

Therefore, by Chernoff's inequality and union bound,

$$\left(\max_{i=1,\ldots,n} |S_i| \right)^2 \;=\; O_{\mathbb{P}}(\log^2 n),$$

and so the claim follows by the sub-Gaussian tail inequality.

■

Lemma 9 *Let $\tau_i^* = E\{Y_i|e(X_i), Z_i = 1\} - E\{Y_i|e(X_i), Z_i = 0\}$ be the treatment effect for unit i. Suppose that $\hat{e}$ is independent of $\{(Y_i, X_i, Z_i)\}_{i=1}^n$ and satisfies that for $\hat{\sigma}$ as defined in (14) we write*

$$\mathrm{tv}_1 \,:=\, \max\left\{\sum_{i=1}^{n-1}\left|f_0\{e(X_{\hat{\sigma}(i)})\} - f_0\{e(X_{\hat{\sigma}(i+1)})\}\right|,\, \sum_{i=1}^{n-1}\left|f_1\{e(X_{\hat{\sigma}(i)})\} - f_1\{e(X_{\hat{\sigma}(i+1)})\}\right|\right\}, \tag{29}$$

and

$$\mathrm{tv}_2 \,:=\, \max\left\{\sum_{i=1}^{n}|f_0\{e(X_i)\} - f_0\{e(X_{N(i)})\}|,\, \sum_{i=1}^{n}|f_1\{e(X_i)\} - f_1\{e(X_{N(i)})\}|\right\}, \tag{30}$$

and assume that these random sequences satisfy $\max\{\mathrm{tv}_1, \mathrm{tv}_2\} = O_{\mathbb{P}}(t)$ for a deterministic t that can depend on n and can diverge. Then, if Assumptions 1, 7 and 8 hold, and

$$\lambda \,=\, \Theta\left\{n^{1/3}(\log n)^2(\log\log n)t^{-1/3}\right\}, \tag{31}$$

we have that

$$\frac{1}{n}\sum_{i=1}^{n}(\tau_i^* - \hat{\tau}_i)^2 \,=\, O_{\mathbb{P}}\left\{\frac{\log^3 n}{n} + (\log n)^2(\log\log n)\left(\frac{t}{n}\right)^{2/3} + \frac{t\log^{1/2} n}{n} + \frac{t^2}{n^2}\right\}.$$

Proof

Let $R = \mathrm{row}(\Delta^{(1)})$, the row space of $\Delta^{(1)}$. Also write P_R and $P_{R^\perp}$ for the orthogonal projection matrices onto R and its orthogonal complement $R^\perp$, respectively. Then let

$$\tilde{\tau} \,=\, \underset{\tau\in R^n}{\arg\min}\left\{\frac{1}{2}\|P_R U - \tau\|^2 + \lambda\|\Delta^{(1)}\hat{P}\tau\|_1\right\},$$

where $U_i \,=\, (-1)^{Z_i+1}\left\{Y_i - Y_{N(i)}\right\}$, for $i = 1, \ldots, n$. Hence, $\hat{\tau} \,=\, P_{R^\perp}U + \tilde{\tau}$, see Proof of Theorem 3 in Wang et al. (2016).

Therefore,

$$\|P_{R^\perp}(\hat{\tau} - \tau^*)\|^2 \,=\, \left|\frac{1}{n^{1/2}}\sum_{i=1}^{n}(U_i - \tau_i^*)\right|^2.$$

We now study the quantity

$$\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(U_i - \tau_i^*)\right|^2.$$

We begin by noticing that for each $i = 1, \ldots, n$, the observed outcome can be written as

$$Y_i = f_1(e(X_i))Z_i + f_0(e(X_i))(1 - Z_i) + \epsilon_i,$$

where $e(X_i)$ is the true propensity score, and the functions f_0, f_1 are defined as

$$f_0(s) := \mathbb{E}(Y \mid Z = 0, e(X) = s), \quad f_1(s) := \mathbb{E}(Y \mid Z = 1, e(X) = s).$$

The error term $\epsilon_i = Y_i - \mathbb{E}(Y \mid Z_i, e(X_i))$ is mean-zero and sub-Gaussian by Assumption 3. In addition, we have

$$\tau_i^* = f_1(e(X_i)) - f_0(e(X_i)),$$

by the definition of τ_i^*, f_0, and f_1. Expanding the definition of U_i and subtracting τ_i^* yields

$$\begin{aligned} U_i - \tau_i^* = Z_i \left[-f_0(e(X_{N(i)})) + f_0(e(X_i))\right] + (1 - Z_i) \left[-f_1(e(X_i)) + f_1(e(X_{N(i)}))\right] \\ + (-1)^{Z_i+1}(\epsilon_i - \epsilon_{N(i)}). \end{aligned}$$

Consequently, the quantity of interest becomes

$$\begin{aligned} \left|\frac{1}{\sqrt{n}}\sum_{i=1}^n (U_i - \tau_i^*)\right|^2 = \Bigg|\frac{1}{\sqrt{n}}\sum_{i=1}^n \Big[& Z_i \left\{-f_0(e(X_{N(i)})) + f_0(e(X_i))\right\} \\ & + (1 - Z_i) \left\{-f_1(e(X_i)) + f_1(e(X_{N(i)}))\right\} \\ & +(-1)^{Z_i+1}(\epsilon_i - \epsilon_{N(i)})\Big]\Bigg|^2. \end{aligned}$$

To bound this expression, define

$$\begin{aligned} A &:= \frac{1}{\sqrt{n}}\sum_{i=1}^n \left[Z_i \left\{-f_0(e(X_{N(i)})) + f_0(e(X_i))\right\} + (1 - Z_i) \left\{-f_1(e(X_i)) + f_1(e(X_{N(i)}))\right\}\right], \\ B &:= \frac{1}{\sqrt{n}}\sum_{i=1}^n (-1)^{Z_i+1}\epsilon_i, \quad C := \frac{1}{\sqrt{n}}\sum_{i=1}^n (-1)^{Z_i+1}\epsilon_{N(i)}. \end{aligned}$$

Then, applying the inequality $(a + b + c)^2 \leq 2a^2 + 4b^2 + 4c^2$, we obtain

$$\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n (U_i - \tau_i^*)\right|^2 \leq 2|A|^2 + 4|B|^2 + 4|C|^2.$$

This leads to the bound

$$\begin{aligned} \left|\frac{1}{\sqrt{n}}\sum_{i=1}^n (U_i - \tau_i^*)\right|^2 \leq 2\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \left(Z_i[f_0(e(X_{N(i)})) - f_0(e(X_i))] + (1 - Z_i)[f_1(e(X_i)) - f_1(e(X_{N(i)}))]\right)\right|^2 \\ + 4\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n (-1)^{Z_i+1}\epsilon_i\right|^2 + 4\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n (-1)^{Z_i+1}\epsilon_{N(i)}\right|^2. \end{aligned}$$

Therefore,

$$\begin{array}{rcl}\|P_{R^{\perp}}(\hat{\tau}-\tau^*)\|^2 & = & \left|\frac{1}{n^{1/2}}\sum_{i=1}^{n}(U_i-\tau_i^*)\right|^2 \\ & \leq & 2\left|\frac{1}{n^{1/2}}\sum_{i=1}^{n}\left(Z_i[f_0\{e(X_i)\}-f_0\{e(X_{N(i)})\}]+(1-Z_i)[-f_1\{e(X_i)\}+f_1\{e(X_{N(i)})\}]\right)\right|^2 \\ & & +4\left|\frac{1}{n^{1/2}}\sum_{i=1}^{n}(-1)^{Z_i+1}\epsilon_i\right|^2+4\left|\frac{1}{n^{1/2}}\sum_{i=1}^{n}(-1)^{Z_i+1}\epsilon_{N(i)}\right|^2 \\ & \leq & \frac{4}{n}\left[\sum_{i=1}^{n}|f_0\{e(X_i)\}-f_0\{e(X_{N(i)})\}|\right]^2+\frac{4}{n}\left[\sum_{i=1}^{n}|f_1\{e(X_i)\}-f_1\{e(X_{N(i)})\}|\right]^2 \\ & & +4\left|\frac{1}{n^{1/2}}\sum_{i=1}^{n}(-1)^{Z_i+1}\epsilon_i\right|^2+4\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(-1)^{Z_i+1}\epsilon_{N(i)}\right|^2.\end{array} \tag{32}$$

However, by the sub-Gaussian tail inequality,

$$\left|\frac{1}{n^{1/2}}\sum_{i=1}^{n}(-1)^{Z_i+1}\epsilon_i\right|^2 = o_{\mathbb{P}}\left(\log n\right). \tag{33}$$

Therefore, combining (32), (33), (30) and Lemma 8,

$$\|P_{R^{\perp}}(\hat{\tau}-\tau^*)\|^2 = O_{\mathbb{P}}\left(\log^3 n+\frac{t^2}{n}\right). \tag{34}$$

Next, we proceed to bound $\|P_R(\hat{\tau}-\tau^*)\|^2$. Notice that by the optimality of $\tilde{\tau}$

$$\frac{1}{2}\|P_R(\tau^*-\tilde{\tau})\|^2 \leq \tilde{\epsilon}^{\top}P_R(\tilde{\tau}-\tau^*)+\lambda\left\{\|\Delta^{(1)}\hat{P}\tau^*\|_1-\|\Delta^{(1)}\hat{P}\tilde{\tau}\|_1\right\}, \tag{35}$$

where

$$\begin{array}{rcl}\tilde{\epsilon}_i & := & (-1)^{Z_i+1}(Y_i-Y_{N(i)})-\tau_i^* \\ & = & Z_i[f_0\{e(X_i)\}-f_0\{e(X_{N(i)})\}]+(1-Z_i)[-f_1(e(X_i))+f_1\{e(X_{N(i)})\}]+(-1)^{Z_i+1}(\epsilon_i-\epsilon_{N(i)}),\end{array}$$

for $i=1,\ldots,n$.

Then by Hölder's inequality, Lemma 7, and the inequality $\|P_R v\|_\infty \leq 2\|v\|_\infty$, there exists a constant $\tilde{C}>0$ such that, with probability approaching one,

$$\begin{array}{rcl}\tilde{\epsilon}^{\top}P_R(\tilde{\tau}-\tau^*) & \leq & \sum_{i=1}^{n}(-1)^{Z_i+1}(\epsilon_i-\epsilon_{N(i)})\{P_R(\tilde{\tau}-\tau^*)\}_i+ \\ & & +2\|\tilde{\tau}-\tau^*\|_\infty\sum_{i=1}^{n}|f_1\{e(X_i)\}-f_1\{e(X_{N(i)})\}|+2\|\tilde{\tau}-\tau^*\|_\infty\sum_{i=1}^{n}|f_0\{e(X_i)\}-f_0\{e(X_{N(i)})\} \\ & \leq & \sum_{i=1}^{n}(-1)^{Z_i+1}(\epsilon_i-\epsilon_{N(i)})\{P_R(\tilde{\tau}-\tau^*)\}_i+A,\end{array} \tag{36}$$

where

$$A := \tilde{C}\,\left(\max\{\|f_0\|_\infty, \|f_1\|_\infty\} + \lambda + \log^{1/2} n\right) t,$$

for some positive constant $\tilde{C}$.

Next, suppose that $\|P_R(\tau^* - \hat{\tau})\|^2/4 \leq A$. Then, due to our choice of λ, (35), and (36), there exists $\tilde{C}_2 > 0$ such that

$$\begin{aligned} \frac{1}{n}\|P_R(\tau^* - \hat{\tau})\|^2 \;\; &\leq \;\; \frac{4\tilde{C}\,\left(\max\{\|f_0\|_\infty, \|f_1\|_\infty\} + \lambda + \log^{1/2} n\right) t}{n} \\ &\leq \;\; \frac{4\tilde{C}\,\max\{\|f_0\|_\infty, \|f_1\|_\infty\} t}{n} + \\ & \quad\;\; \frac{4\,\tilde{C}_2\,\tilde{C}(\log\log n)(\log n)^2\, t^{2/3}}{n^{2/3}} + \\ & \quad\;\; 4\tilde{C}\,\frac{t\log^{1/2} n}{n}, \end{aligned} \tag{37}$$

with probability approaching one.

On the contrary, if $\|P_R(\tau^* - \hat{\tau})\|^2/4 > A$, then (35) and (36) imply

$$\begin{aligned} \frac{1}{4}\|P_R(\tau^* - \tilde{\tau})\|^2 \;\; &\leq \;\; \sum_{i=1}^{n}(-1)^{Z_i+1}(\epsilon_i - \epsilon_{N(i)})\{P_R(\tilde{\tau} - \tau^*)\}_i + \\ & \quad\;\; \lambda\left\{\|\Delta^{(1)}\hat{P}\tau^*\|_1 - \|\Delta^{(1)}\hat{P}\tilde{\tau}\|_1\right\}, \end{aligned} \tag{38}$$

Next, we proceed to bound the first term in the right hand size of the previous inequality. Towards that end, with the notation from the proof of Lemma 8, we exploit the argument in the proof of Lemma 9 from Wang et al. (2016). First, by Lemma 9, Theorem 10, and Corollary 12 from Wang et al. (2016), we have that

$$\sup_{u \in \mathrm{row}(\Delta^{(1)})\,:\,\|\Delta^{(1)}\hat{P}u\|_1 \leq 1,} \frac{\sum_{i=1}^{n}(-1)^{Z_i+1}\epsilon_i\, u_i}{\|u\|^{1/2}} \;=\; O_{\mathbb{P}}\left\{n^{1/4}(\log\log n)^{1/2}\right\}, \tag{39}$$

which follows due to the independence and sub-Gaussian assumption of the errors $\{\epsilon_i\}_{i=1}^n$.

On the other hand, conditioning on X and $\hat{e}$, we define the random vectors $\epsilon^{(1)}, \ldots, \epsilon^{(M)} \in \mathbb{R}^n$, constructed as follows. First, let

$$M \;=\; \max_{1 \leq i \leq n} |\{j \in \{1, \ldots, n\} \,:\, N(j) = i\}|\,.$$

Then

$$\epsilon_i^{(1)} \;=\; \begin{cases} (-1)^{Z_i+1}\epsilon_{N(1)} & \text{if } \; i = 1, \\ (-1)^{Z_i+1}\epsilon_{N(i)} & \text{if } \; N(j) \neq N(i) \; \text{ for all } \; j < i \\ 0 & \text{otherwise,} \end{cases}$$

and

$$S^{(1)} \;=\; \{i \,:\, N(j) \neq N(i) \;\; \forall j < i \quad\} \cup \{1\}.$$

And for $l > 1$, we iteratively construct

$$\epsilon_i^{(l)} \;=\; \begin{cases} (-1)^{Z_i+1}\epsilon_{N(i)} & \text{if } \; i \notin \cup_{m=1}^{l-1} S^{(m)}, \; \text{ and } \; N(j) \neq N(i) \; \text{ for all } \; j < i \; \text{ and } \; j \notin \cup_{m=1}^{l-1} S^{(m)} \\ 0 & \text{otherwise.} \end{cases}$$

Notice that, by construction, the components of each $\epsilon^{(l)}$ are independent and subGaussian(v). Hence, by triangle inequality,

$$\sup_{u\in \mathrm{row}(\Delta^{(1)})\,:\,\|\Delta^{(1)}\hat{P}u\|_1\leq 1,} \frac{\sum_{i=1}^n (-1)^{Z_i+1}\epsilon_{N(i)}\, u_i}{\|u\|^{1/2}} \;\;\leq\;\; \sum_{j=1}^{M} \sup_{u\in \mathrm{row}(\Delta^{(1)})\,:\,\|\Delta^{(1)}\hat{P}u\|_1\leq 1,} \frac{u^\top \epsilon^{(j)}}{\|u\|^{1/2}}.$$

Then as in the proof of Theorem 10 in Wang et al. (2016), which exploits Lemma 3.5 from Van de Geer (1990), we have that

$$\mathbb{P}\left(\sup_{u\in \mathrm{row}(\Delta^{(1)})\,:\,\|\Delta^{(1)}\hat{P}u\|_1\leq 1} \frac{u^\top \epsilon^{(j)}}{\|u\|^{1/2}} \geq c_1 L n^{1/4}(\log\log n)^{1/2} \,\middle|\, X, \hat{e} \right) \;\leq\; \exp\left(-\frac{c_0 L^2 \log\log n}{v}\right),$$

for $j=1,\ldots,M$ and for some constant $c_0, c_1 > 0$, and for any constant $L > L_0$, where L_0 is a fixed constant.

Therefore, by a union bound,

$$\begin{aligned} &\mathbb{P}\left(\sup_{u\in \mathrm{row}(\Delta^{(1)})\,:\,\|\Delta^{(1)}\hat{P}u\|_1\leq 1,} \frac{\sum_{i=1}^n (-1)^{Z_i+1}\epsilon_{N(i)}\, u_i}{\|u\|^{1/2}} \geq c_1 L n^{1/4}(\log\log n)^{1/2} M \,\middle|\, X, \hat{e} \right) \\ &\leq\; M \exp\left(-\frac{c_0 L^2 \log\log n}{v}\right). \end{aligned} \tag{40}$$

Hence, combining (39) with (40), integrating over X and $\hat{e}$, and proceeding as in the proof of Lemma 8, we arrive at

$$\sup_{u\in \mathrm{row}(\Delta^{(1)})\,:\,\|\Delta^{(1)}\hat{P}u\|_1\leq 1} \frac{\sum_{i=1}^n (-1)^{Z_i+1}(\epsilon_i - \epsilon_{N(i)}) u_i}{\|u\|^{1/2}} \;=\; O_{\mathbb{P}}\left(K\right), \tag{41}$$

where

$$K = n^{1/4}(\log\log n)^{1/2}\log n.$$

Next, we notice that due to (38) , (41), the proof of Lemma 9 in Wang et al. (2016), and since $\tilde{\tau} = P_R\hat{\tau}$, we have that

$$\|P_R(\tau^* - \hat{\tau})\|^2 \;\;=\;\; O_{\mathbb{P}}\left\{\lambda\, t + K^4\left(\tfrac{1}{\lambda}\right)^2\right\}, \tag{42}$$

and

$$\lambda\, t + K^4\left(\tfrac{1}{\lambda}\right)^2 \;\;=\;\; O\big\{n^{1/3}(\log n)^2(\log\log n)\, t^{2/3}\big\}.$$

Hence, combining (34), (37) and (42), we obtain that

$$\|\tau^* - \hat{\tau}\|^2 \;\;=\;\; O_{\mathbb{P}}\left\{ \log^3 n + n^{1/3}(\log n)^2(\log\log n)\, t^{2/3} + t\log^{1/2} n + \frac{t^2}{n}\right\},$$

and the claim follows.

■

D.2 Propensity score auxiliary lemmas

Lemma 10 *Under Assumption 9, for all $\zeta > 0$,*

$$\mathbb{P}\left[\left|\nabla \hat{L}(\theta^*)_j\right| \leq \sigma_0 \left\{\frac{2\log(3/\zeta)}{m}\right\}^{1/2}\right] \geq 1-\zeta,$$

for all $j = 1, \ldots, d$ and for a positive constant σ_0.

Proof First, we observe that

$$\nabla \hat{L}(\theta^*) \;=\; -\frac{1}{m}\sum_{i=1}^{m} X_i' \left\{Z_i' - F(X_i'^{\top}\theta^*)\right\}. \tag{43}$$

Furthermore, notice that

$$\begin{array}{rcl} L(\theta) & = & -E\left(E\left[Z\log F(X^{\top}\theta) + (1-Z)\log\{1-F(X^{\top}\theta)\}|X\right]\right) \\ & = & -E\left[e(X)\log F(X^{\top}\theta) + (1-e(X))\log\{1-F(X^{\top}\theta)\}\right], \end{array}$$

and define

$$H(x,\theta) = e(x)\log F(x^{\top}\theta) + \{1-e(x)\}\log\{1-F(x^{\top}\theta)\}.$$

Then

$$\nabla_\theta H(x,\theta) \;=\; e(x)x - F(x^{\top}\theta)x,$$

and $|(\nabla_\theta H(x,\theta))_j| \leq 2|x_j|$ for all $j \in \{1,\ldots,d\}$. Hence, by the dominated convergence theorem,

$$\begin{array}{rcl} \nabla L(\theta) & = & -E\left\{\nabla_\theta H(X,\theta)\right\} \\ & = & -E\left(e(X)X - F(X^{\top}\theta)X\right) \\ & = & -E\left[E\left\{ZX - F(X^{\top}\theta)X|X\right\}\right] \\ & = & -E\left\{ZX - F(X^{\top}\theta)X\right\}. \end{array}$$

Therefore,

$$E\left\{ZX - F(X^{\top}\theta^*)X\right\} = 0. \tag{44}$$

Also,

$$\left|X_{i,j}'\left\{Z_i' - F(X_i'^{\top}\theta^*)\right\}\right| \leq |X_{i,j}'|.$$

Hence, by Assumption 9, (43), (44) and Corollary 2.6 from Boucheron et al. (2013), we obtain the conclusion. ■

Lemma 11 *Suppose that Assumption 11 holds and $d \leq 4\|X\|_{\psi_2}^2 m^{1/2}/\log^{1/2} m$. Then*

$$\mathbb{P}\left\{\Lambda_{\min}(Q^m) \leq C_{\min}/2\right\} \leq 2\exp\left(-\frac{cd}{16}\log^2 m + d\log 9\right),$$

where

$$Q^m = \frac{1}{m}\sum_{i=1}^{m}\eta(X_i'^{\top}\theta^*)X_i'X_i'^{\top},$$

and c is a positive constant. Similarly,

$$\mathbb{P}\left(\Lambda_{\max}\left\{\frac{1}{m}\sum_{i=1}^{m} X_i' X_i'^{\top}\right\} \geq 3D_{\max}/2\right) \leq\; 2\exp\left(-\frac{cd}{16}\log^2 m + d\log 9\right).$$

Proof Proceeding as in the proof of Lemma 5 in Ravikumar et al. (2010), we obtain that

$$\Lambda_{\min}\left(Q^m\right) \geq C_{\min} - \|Q - Q^m\|_2,$$

where $\|\cdot\|_2$ denotes the spectral norm and with

$$Q \;=\; E\left\{\eta(X^{\top}\theta^*)XX^{\top}\right\}.$$

To bound the quantity $\|Q - Q^m\|_2$ we let $v \in \mathbb{R}^d$ with $\|v\| = 1$, and notice that

$$v^T\left(Q^m - Q\right)v = \frac{1}{m}\sum_{i=1}^{m}\left\{\left[\{\eta(X_i'^{\top}\theta^*)\}^{1/2}v^T X_i'\right]^2 \;-\; E\left(\left[\{\eta(X_i'^{\top}\theta^*)\}^{1/2}v^T X_i'\right]^2\right)\right\},$$

and by Proposition 5.16 in Vershynin (2010)

$$\mathbb{P}\left\{\left|v^T\left(Q^m - Q\right)v\right| \geq r\right\} \;\leq\; 2\exp\left(-c\min\left\{\frac{r^2 m}{16\|X\|_{\psi_2}^4}, \frac{rm}{4\|X\|_{\psi_2}^2}\right\}\right),$$

for all $r > 0$, and for an absolute constant $c > 0$. Hence, taking $r = \|X\|_{\psi_2}^2 d^{1/2}\log^{1/2} m/m^{1/2}$, and with the same entropy based argument from the proof of Lemma 5 in Wang et al. (2017), we arrive at

$$\mathbb{P}\left\{\|Q^m - Q\|_2 \geq \|X\|_{\psi_2}^2\left(\frac{d\log m}{m}\right)^{1/2}\right\} \;\leq\; 2\exp\left(-\frac{cd}{16}\log^2 m + d\log 9\right).$$

Finally,

$$\begin{aligned}
&\mathbb{P}\left\{\Lambda_{\max}\left(\frac{1}{m}\sum_{i=1}^{m} X_i' X_i'^{\top}\right) \geq 3D_{\max}/2\right\} \\
&\leq\; \mathbb{P}\left\{\left\|\frac{1}{m}\sum_{i=1}^{m} X_i' X_i'^{\top} - E\left(\frac{1}{m}\sum_{i=1}^{m} X_i' X_i'^{\top}\right)\right\|_2 \geq \|X\|_{\psi_2}^2\left(\frac{d\log m}{m}\right)^{1/2}\right\}
\end{aligned}$$

and the proof concludes with the same argument from above.

■

Lemma 12 *Suppose that Assumption 9–11 hold. Then for a positive constant C_1, the estimator $\hat{\theta}$ defined in (13) satisfies*

$$\|\hat{\theta} - \theta^*\| \;\leq\; \frac{C_1}{C_{\min}}\left(\frac{d\log m}{m}\right)^{1/2},$$

with probability approaching one.

Proof For $u \in \mathbb{R}^d$ let

$$G(u) \;=\; \hat{L}(\theta^* + u) - \hat{L}(\theta^*).$$

Clearly, $G(0) = 0$, and $G(\hat{u}) \leq 0$ where $\hat{u} = \hat{\theta} - \theta^*$. Let

$$B := \frac{8\sigma_0}{C_{\min}} \left(\frac{12 d \log m}{m} \right)^{1/2}. \tag{45}$$

We proceed to show that $G(u) > 0$ for all $\|u\| = B$, which implies, by convexity, that $\|\hat{u}\| \leq B$. Towards that end, notice that, by Taylor's theorem, we have

$$G(u) = \nabla \hat{L}(\theta^*)^\top u + u^\top \nabla^2 \hat{L}(\theta^* + \alpha u) u,$$

for some $\alpha \in [0,1]$. Also,

$$|\nabla \hat{L}(\theta^*)^\top u| \leq \|\nabla \hat{L}(\theta^*)\|_\infty \|u\|_1 \leq d^{1/2} \|\nabla \hat{L}(\theta^*)\|_\infty \|u\|.$$

Hence, by Lemma 10 and a union bound,

$$G(u) \geq -d^{1/2} \|u\| \sigma_0 \left(\frac{12 \log m}{m} \right)^{1/2} + u^T \nabla^2 \hat{L}(\theta^* + \alpha u) u, \tag{46}$$

with probability at least $1 - 1/m$.

Furthermore,

$$\begin{aligned} \nabla^2 \hat{L}(\theta^* + \alpha u) \;&=\; \frac{1}{m} \sum_{i=1}^m F'\{X_i'^\top (\theta^* + \alpha u)\} X_i' X_i'^\top \\ &=: \; \mathcal{A}_1. \end{aligned}$$

Also,

$$\begin{aligned} q^* \;&:=\; \Lambda_{\min}(\mathcal{A}_1) \\ &\geq\; \min_{\alpha \in [0,1]} \Lambda_{\min} \left[\frac{1}{m} \sum_{i=1}^m \eta\{X_i'^\top (\theta^* + \alpha u)\}\, X_i' X_i'^\top \right] \\ &\geq\; \min_{\alpha \in [0,1]} \Lambda_{\min} \left\{ \frac{1}{m} \sum_{i=1}^m \eta(X_i'^\top \theta^*)\, X_i' X_i'^\top \right\} - \\ &\quad\; \max_{\alpha \in [0,1]} \left\| \frac{1}{m} \sum_{i=1}^m \eta'\{X_i'^\top (\theta^* + \alpha u)\} (u^T X_i') X_i' X_i'^\top \right\|_2 \\ &=\; \Lambda_{\min}(Q^m) - \max_{\alpha \in [0,1]} \|A(\alpha)\|_2, \end{aligned} \tag{47}$$

where

$$A(\alpha) = \frac{1}{m} \sum_{i=1}^m \eta'\{X_i'^\top (\theta^* + \alpha u)\} (u^T X_i') X_i' X_i'^\top.$$

Now for $\alpha \in [0,1]$, $v \in \mathbb{R}^d$ with $\|v\| = 1$, we have

$$\begin{aligned} v^T A(\alpha) v &= \frac{1}{m}\sum_{i=1}^{m} \eta'\{X_i'(\theta^* + \alpha u)\}\left(u^\top X_i'\right)\left(X_i'^\top v\right)^2 \\ &\leq \|u\| \left\{\max_{i=1,\ldots,m} \left|(u/\|u\|)^\top X_i'\right|\right\} \|\eta'\|_\infty \frac{1}{m}\sum_{i=1}^{m}\left(X_i'^\top v\right)^2 \\ &\leq \|u\| \left\{\max_{i=1,\ldots,m} \left|(u/\|u\|)^\top X_i'\right|\right\} \left\|\frac{1}{m}\sum_{i=1}^{m} X_i' X_i'^\top\right\|_2 \\ &\leq c_2 d^{1/2} \log^{1/2} m \|u\|\, D_{\max}. \end{aligned} \tag{48}$$

with probability approaching one, and where $c_2 > 0$ is a constant. Here, we have used the fact the random variables $\{(u/\|u\|)^\top X_i'\}_{i=1}^m$ are sub-Gaussian, and the second claim in Lemma 11. Therefore, with probability approaching one,

$$\begin{aligned} G(u) &\geq -d^{1/2}\|u\|\sigma_0 \left(\frac{12\log m}{m}\right)^{1/2} + \frac{C_{\min}\|u\|^2}{2} - c_2 (d\log m)^{1/2}\|u\|^3\, D_{\max} \\ &\geq -d^{1/2}\|u\|\sigma_0 \left(\frac{12\log m}{m}\right)^{1/2} + \frac{C_{\min}\|u\|^2}{4} \\ &> 0, \end{aligned}$$

where the first inequality follows from (46)– (48), the second from (45) and (23), and the third since

$$\|u\| = B > \frac{4\sigma_0}{C_{\min}}\left(\frac{12 d \log m}{m}\right)^{1/2}.$$

■

Lemma 13 *Under Assumptions 1 and 9–11, there exists a constant $C_0 > 0$ such that*

$$\sup_{i=1,\ldots,n} |\hat{e}(X_i) - e(X_i)| \leq C_0 \frac{d}{C_{\min}} \left[\frac{(\log m)\{\log(nd)\}}{m}\right]^{1/2},$$

with probability approaching one, provided that $d = O(m)$.

Proof We have that

$$\begin{aligned} \max_{i=1,\ldots,n} |\hat{e}(X_i) - e(X_i)| &= \max_{i=1,\ldots,n} \left|F(X_i^\top \hat{\theta}) - F(X_i^\top \theta^*)\right| \\ &\leq \max_{i=1,\ldots,n} \|F'\|_\infty \left|X_i^\top \hat{\theta} - X_i^\top \theta^*\right| \\ &\leq \|\hat{\theta} - \theta^*\| \max_{i=1,\ldots,n,} \|X_i\| \\ &\leq d^{1/2}\, \|\hat{\theta} - \theta^*\| \max_{i=1,\ldots,n,\ j=1,\ldots,d} |X_{i,j}| \\ &\leq d^{1/2}\, \|\hat{\theta} - \theta^*\| \max_{i=1,\ldots,n,\ j=1,\ldots,d} |X_{i,j} - (\mathbb{E}(X))_j| \\ &\quad + d^{1/2}\, \|\hat{\theta} - \theta^*\|\, \|E(X)\|_\infty, \end{aligned} \tag{49}$$

and the claim follows by Lemma 12.

■

D.3 Lemma combining both stages

Lemma 14 *There exists a positive constant C_1 such that the event*

$$\max_{i=1,\ldots,n} |i - \sigma^{-1}\{\hat{\sigma}(i)\}| \leq C_1 \max\{\log n, n\delta\},$$

holds with probability approaching one.

Proof First, by Lemma 13, we will assume that the event

$$\sup_{i=1,\ldots,n} |\hat{e}(X_i) - e(X_i)| \leq C_0\delta, \tag{50}$$

holds.

Then for each $i \in \{1,\ldots,n\}$ define

$$m_i = \left|\left\{k \in \{1,\ldots,n\}\backslash\{i\} \,:\, e(X_k) \in \left(e(X_i) - 2C_0\tilde{\delta}, e(X_i) + 2C_0\tilde{\delta}\right)\right\}\right|.$$

where $\tilde{\delta} = \max\{\delta\,,\, C_1\, n^{-1}\log n\}$ for a positive constant C_1 to be chosen later.

Then by Assumption 10,

$$m_i \,\sim\, \text{Binomial}\left(n-1, \int_0^1 \int_{\max\{0,t-2C_0\tilde{\delta}\}}^{\min\{1,t+2C_0\tilde{\delta}\}} h(s)\,h(t)\,ds\,dt\right).$$

Hence,

$$E(m_i) \leq 4C_0 h_{\max}^2 n\tilde{\delta}.$$

Therefore, by a union bound, and Chernoff's inequality,

$$\mathbb{P}\left(\max_{i=1,\ldots,n} m_i \,\geq\, 12C_0\, h_{\max}^2 n\tilde{\delta}\right) \quad\leq\quad \exp\left(-C_1\log n/4 + \log n\right) \,\to 0, \qquad \text{as } n \to \infty, \tag{51}$$

provide that C_1 is chosen large enough.

On the other hand, by (50), we have

- If $e(X_{i'}) < e(X_i) - 2C_0\tilde{\delta}$, then $\hat{e}(X_{i'}) < \hat{e}(X_i)$.
- If $e(X_{i'}) > e(X_i) + 2C_0\tilde{\delta}$, then $\hat{e}(X_{i'}) > \hat{e}(X_i)$.

Therefore,

$$\max_{i=1,\ldots,n} |i - \sigma^{-1}\{\hat{\sigma}(i)\}| \,\leq\, \max_{i=1,\ldots,n} m_i,$$

and the claim follows combining (50) and (51). ■

Lemma 15 *With the notation from before,*

$$\sum_{i=1}^{n-1} |f_1\{e(X_{\hat{\sigma}(i)})\} - f_1\{e(X_{\hat{\sigma}(i+1)})\}| \,=\, O_{\mathbb{P}}\left(\max\{\log n, n\delta\}\right),$$

and

$$\sum_{i=1}^{n-1} |f_0\{e(X_{\hat{\sigma}(i)})\} - f_0\{e(X_{\hat{\sigma}(i+1)})\}| \,=\, O_{\mathbb{P}}\left(\max\{\log n, n\delta\}\right).$$

Proof Suppose that the event

$$\max_{i=1,\ldots,n} |i - \sigma^{-1}\{\hat{\sigma}(i)\}| \leq C_1 \max\{\log n, n\delta\}, \tag{52}$$

holds for some positive constant C_1, see Lemma 14.

Then

$$\begin{array}{rcl} \sum_{i=1}^{n-1} |f_1\{e(X_{\hat{\sigma}(i)})\} - f_1\{e(X_{\hat{\sigma}(i+1)})\}| & \leq & \sum_{i=1}^{n-1} \sum_{j=\max\{1, i-C_1\max\{\log n, n\delta\}\}}^{\min\{n, i+C_1\max\{\log n, n\delta\}\}} |f_1\{e(X_{\sigma(j)})\} - f_1\{e(X_{\sigma(j+1)})\}| \\ & \leq & 2C_1 \max\{\log n, n\delta\} \sum_{i=1}^{n-1} |f_1\{e(X_{\sigma(i)})\} - f_1\{e(X_{\sigma(i+1)})\}| \end{array}$$

and the claim follows.

■

Lemma 16 *There exists a positive constant* c_1 *such that the event*

$$\max\left\{ \max_{i=1,\ldots,n} \min_{Z_j \neq Z_i} |e(X_i) - e(X_j)|\,,\; \max_{i=1,\ldots,n} \min_{Z_j = Z_i} |e(X_i) - e(X_j)| \right\} \leq \frac{c_1 \log n}{n}, \tag{53}$$

happens with probability approaching one.

Proof Let $c_1 > 0$ be a constant to be chosen later. Fix $i \in \{1, \ldots, n\}$ and define

$$\Lambda_i = \{j \geq 1 \,:\, Z_{i+k} \neq Z_i \text{ for } k \in \{1, \ldots, j\}\},$$

and $m_i = \max\{a \,:\, a \in \Lambda_i\}$.

Then we have that

$$\begin{array}{rcl} \mathbb{P}\left(m_i \geq c_1 \log n\right) & = & \mathbb{P}\left(m_i \geq c_1 \log n | Z_i = 1\right) \mathbb{P}(Z_i = 1) + \mathbb{P}\left(m_i \geq c_1 \log n | Z_i = 1\right) \mathbb{P}(Z_i = 0) \\ & \leq & \mathbb{P}\left(m_i \geq c_1 \log n | Z_i = 1\right) e_{\max} + \mathbb{P}\left(m_i \geq c_1 \log n | Z_i = 0\right) (1 - e_{\min}) \\ & \leq & (1 - e_{\min})^{c_1 \log n} e_{\max} + e_{\max}^{c_1 \log n}(1 - e_{\min}) \\ & \leq & \max\{e_{\max}, 1 - e_{\min}\}^{c_1 \log n} \\ & \leq & \exp\{c_1(\max\{\log(e_{\max}), \log(1 - e_{\min})\}) \log n\}. \end{array}$$

Hence, by union bound,

$$\mathbb{P}\left(\max_{i=1,\ldots,n} m_i \geq c_1 \log n\right) \quad \leq \quad \exp\{c_1(\max\{\log(e_{\max}), \log(1 - e_{\min})\}) \log n + \log n\},$$

and so we set $c_1 = -2/\max\{\log(e_{\max}), \log(1 - e_{\min})\}$. The claim follows by Assumption 10 and the argument in the proof of Proposition 30 in Von Luxburg et al. (2014) implying that, with high probability, the distance of of each $e(X_i)$ to its r nearest neighbor is of order $O(\log n/n)$ for $r \asymp \log n$.

■

Lemma 17 *For any $i \in \{1,\ldots,n\}$ let*

$$\xi_i \,=\, \left|\{j \in \{1,\ldots,n\} \,:\quad \text{and} \quad e(X_i) \le e(X_j) \le e(X_{N(i)}) \;\; \text{or} \;\; e(X_{N(i)}) \le e(X_j) \le e(X_i)\}\right|.$$

Then, for some constant $C_2 > 0$,

$$\max_{i=1,\ldots,n} \xi_i \,\le\, C_2 \max\{\log n, n\delta\},$$

with probability approaching one.

Proof First, assume that the event (50) holds. Next, by Lemma 16, with high probability, for all $i \in \{1,\ldots,n\}$ there exits $j(i) \in \{1,\ldots,n\}$ such that $Z_{j(i)} \neq Z_i$ and $|e(X_i) - e(X_{j(i)})| \le (c_1 \log n)/n$. Under such event,

$$\begin{array}{lll} |e(X_i) - e(X_{N(i)})| & \le & |e(X_i) - \hat{e}(X_i)| + |\hat{e}(X_i) - \hat{e}(X_{N(i)})| + |\hat{e}(X_{N(i)}) - e(X_{N(i)})| \\ & \le & |e(X_i) - \hat{e}(X_i)| + |\hat{e}(X_i) - \hat{e}(X_{j(i)})| + |\hat{e}(X_{N(i)}) - e(X_{N(i)})| \\ & \le & 2C_0\delta + |\hat{e}(X_i) - \hat{e}(X_{j(i)})| \\ & \le & 4C_0\delta + |e(X_i) - e(X_{j(i)})| \\ & \le & 4C_0\delta + \frac{c_1 \log n}{n}, \end{array}$$

where the second inequality follows from the definition of $N(i)$, the third and fourth from Lemma 13, and the last from the construction of $j(i)$. Therefore,

$$\xi_i \,\le\, \left|\left\{ j \,:\, |e(X_i) - e(X_j)| \le 4C_0\delta + \frac{c_1 \log n}{n} \right\}\right|,$$

and the claim follows in the same way that we bounded the counts $\{m_i\}$ in the proof of Lemma 14. ■

Lemma 18 *With the notation from before,*

$$\sum_{i=1}^{n} |f_l\{e(X_i)\} - f_l\{e(X_{N(i)})\}| = O_{\mathbb{P}}\left(\max\{\log n, n\delta\}\right),$$

for $l \in \{0,1\}$.

Proof By Lemma 17 and the triangle inequality, we have that, with probability approaching one,

$$\begin{array}{lll} \displaystyle\sum_{i=1}^{n} |f_l\{e(X_i)\} - f_l\{e(X_{N(i)})\}| & \le & \displaystyle\sum_{i=1}^{n-1} \sum_{j=\max\{1, i - C_2\max\{\log n, n\delta\}\}}^{\min\{n, i + C_2\max\{\log n, n\delta\}\}} |f_l\{e(X_{\sigma(j)})\} - f_l\{e(X_{\sigma(j+1)})\}| \\ & \le & [C_2 \max\{\log n, n\delta\}] \left[\displaystyle\sum_{j=1}^{n-1} |f_l\{e(X_{\sigma(j)})\} - f_l\{e(X_{\sigma(j+1)})\}|\right] \end{array}$$

and the claim follows.

■

D.4 Putting the pieces together

The claim in Theorem 4 follows immediately from Lemmas 15, 18 and 9.

Appendix E. Proof of Theorem 2

E.1 Notation

Throughout this section we define the function $\hat{L} : \mathbb{R}^d \to \mathbb{R}$ as

$$\hat{L}(\theta) \,=\, \frac{1}{m}\sum_{i=1}^{m}\left(Y_i' - X_i'^{\top}\theta\right)^2, \qquad \theta \in \mathbb{R}^d,$$

and set

$$\delta = \frac{d}{C_{\min}}\left\{\frac{\left(\log^{1/2} m + d^{1/2}\|\theta^*\|\right)\log m}{m}\right\}^{1/2}.$$

Furthermore, we consider the orderings σ, $\tilde{\sigma}$, and $\hat{\sigma}$ satisfying

$$g\{X_{\sigma(1)}\} < \,\ldots\, < g\{X_{\sigma(n)}\}, \tag{54}$$

$$\tilde{g}\{X_{\tilde{\sigma}(1)}\} < \,\ldots\, < \tilde{g}\{X_{\tilde{\sigma}(n)}\}, \tag{55}$$

and

$$\hat{g}\{X_{\hat{\sigma}(1)}\} < \,\ldots\, < \hat{g}\{X_{\hat{\sigma}(n)}\}. \tag{56}$$

E.2 Auxiliary lemmas

Lemma 19 *Under Assumptions 2-6, we have that for some $C_1 > 0$,*

$$\max_{j=1,\ldots,d}\left|\left\{\nabla\hat{L}(\theta^*)\right\}_j\right| \,\le\, C_1\left\{\frac{(\log^{1/2} m + d^{1/2}\|\theta^*\|)\log m}{m}\right\}^{1/2},$$

with probability approaching one.

Proof First notice that by the optmiality of θ^* we have that

$$\nabla L(\theta^*) \,=\, E\left\{X(X^{\top}\theta^* - Y)|Z = 0\right\}.$$

Hence,

$$0 \,=\, E\left\{\nabla\hat{L}(\theta^*)\right\} \,=\, E\left\{\frac{1}{m}\sum_{i=1}^{m} X_i'(X_i'^{\top}\theta^* - Y_i')\right\}.$$

Furthermore, defining $\tilde{\epsilon}_i = Y_i' - f_0(X_i')$, we have

$$\begin{aligned} |X_{i,j}'(X_i'^{\top}\theta^* - Y_i')| \;&\le\; |X_{i,j}'|\,(\|f_0\|_\infty + \|\tilde{\epsilon}\|_\infty + \|X_i'\|_\infty\|\theta^*\|_1) \\ &\le\; \|X_i'\|_\infty\left(\|f_0\|_\infty + \|\tilde{\epsilon}\|_\infty + \|X_i'\|_\infty d^{1/2}\|\theta^*\|\right) \end{aligned} \tag{57}$$

where the first inequality follows by Hölder's inequality, and the second by the relation between ℓ_1 and ℓ_2 norms.

However, by Assumption 3, it follows that

$$\Omega = \{\|\tilde{\epsilon}\|_\infty \leq 3\sigma \log^{1/2} m\},$$

holds with probability at least $1 - \frac{1}{m^2}$. Hence, by Assumption 6, (57), and Hoeffding's inequality

$$\begin{aligned}&\left|\tfrac{1}{m}\textstyle\sum_{i=1}^m X'_{i,j}(X_i'^{\top}\theta^* - Y_i')\right|\\ &\leq \left\{\frac{4\max\{\|a\|_\infty, \|b\|_\infty\}\left(\|f_0\|_\infty + 3\sigma\log^{1/2} m + \max\{\|a\|_\infty, \|b\|_\infty\} d^{1/2}\|\theta^*\|\right)\log\left(m^2\right)}{2m}\right\}^{1/2},\end{aligned}$$

with proability $1 - \frac{4}{m^2}$.

■

Lemma 20 *For any $i \in \{1, \ldots, n\}$ let*

$$\xi_i = \left|\{j \in \{1, \ldots, n\} : \quad \textit{and} \quad \tilde{g}(X_i) \leq \tilde{g}(X_j) \leq \tilde{g}(X_{N(i)}) \ \textit{ or } \ \tilde{g}(X_{N(i)}) \leq \tilde{g}(X_j) \leq \tilde{g}(X_i)\}\right|.$$

Then, for some constant $C_2 > 0$,

$$\max_{i=1,\ldots,n} \xi_i \leq C_2 \max\{\log n, n\delta\},$$

with probability approaching one.

Proof The claim follows as the proof of Lemma 17, exploiting Lemma 21. ■

Lemma 21 *Under Assumptions 2–6, we have that for some constant $\tilde{C} > 0$,*

$$\max_{i=1,\ldots,n} |\hat{g}(X_i) - \tilde{g}(X_i)| \leq \frac{\tilde{C}d}{C_{\min}}\left\{\frac{\left(\log^{1/2} m + d^{1/2}\|\theta^*\|\right)\log m}{m}\right\}^{1/2},$$

with probability approaching one.

Proof We define the function

$$G(u) = \hat{L}(\theta^* + u) - \hat{L}(\theta^*),$$

and observe that $G(0) = 0$, and $G(\hat{u}) < 0$ where $\hat{u} = \hat{\theta} - \theta^*$. Let

$$B = \frac{c_1}{C_{\min}}\left\{\frac{d\left(\log^{1/2} m + d^{1/2}\|\theta^*\|\right)\log m}{m}\right\}^{1/2},$$

for some $c_1 > 0$, and take $u \in \mathbb{R}^d$ such that $\|u\| = B$. Then, with probability approaching one, by Lemma 19 and the proof of Lemma 11, we have that

$$\begin{aligned} G(u) &= \frac{1}{m}\sum_{i=1}^{m}\left\{X_i'^{\top}(\theta^* + u) - Y_i'\right\}^2 - \frac{1}{m}\sum_{i=1}^{m}\left(X_i'^{\top}\theta^* - Y_i'\right)^2 \\ &= u^{\top}\left(\frac{1}{m}\sum_{i=1}^{m} X_i' X_i'^{\top}\right) u + 2u^{\top}\nabla\hat{L}(\theta^*) \\ &\geq u^{\top}\left(\frac{1}{m}\sum_{i=1}^{m} X_i' X_i'^{\top}\right) u - 2\|u\|_1\|\nabla\hat{L}(\theta^*)\|_{\infty} \\ &\geq \tfrac{C_{\min}\|u\|^2}{2} - 2c_2\|u\| d^{1/2}\left\{\tfrac{(\log^{1/2} m + d^{1/2}\|\theta^*\|)\log m}{m}\right\}^{1/2} \\ &> 0 \end{aligned}$$

for some constant $c_2 > 0$, and where the last inequality follows from the choice of B with a large enough c_1.

Therefore,

$$\|\hat{\theta} - \theta^*\| \;\leq\; \frac{4c_2}{C_{\min}}\left\{\frac{d\left(\log^{1/2} m + d^{1/2}\|\theta^*\|\right)\log m}{m}\right\}^{1/2}, \tag{58}$$

with probability approaching one. Furthermore,

$$\begin{aligned} \max_{i=1,\ldots,n} |\hat{g}(X_i) - \tilde{g}(X_i)| &= \max_{i=1,\ldots,n}\left|X_i^{\top}\hat{\theta} - X_i^{\top}\theta^*\right| \\ &\leq \|\hat{\theta} - \theta^*\| \max_{i=1,\ldots,n,} \|X_i\| \\ &\leq d^{1/2}\,\|\hat{\theta} - \theta^*\| \max_{i=1,\ldots,n,\ j=1,\ldots,d} |X_{i,j}|, \end{aligned} \tag{59}$$

and the conclusion follows from (58) and the fact the $X_i's$ have compact support. ■

E.3 Proof of Theorem 2

The theorem follows as a Theorem 9, proceeding as in the proof of Theorem 4 by using the lemmas below.

Lemma 22 *Let $\tilde{\sigma}$ and $\hat{\sigma}$ as defined in (54)–(56). There exists a positive constant C_2 such that the event*

$$\max_{i=1,\ldots,n} |i - \tilde{\sigma}^{-1}\{\hat{\sigma}(i)\}| \leq C_2 \max\{\log n, n\delta\},$$

holds with probability approaching one.

Proof The claim follows as the proof of Lemma 14, exploiting Lemma 21. ■

Lemma 23 *With the notation from (56),*

$$\sum_{i=1}^{n-1} |f_1\{g(X_{\hat{\sigma}(i)})\} - f_1\{g(X_{\hat{\sigma}(i+1)})\}| \;=\; O_{\mathbb{P}}\left\{\max\{\log n, n\delta\}\left(\overline{\kappa}_n + 1\right)\right\},$$

and

$$\sum_{i=1}^{n-1} |f_0\{g(X_{\hat{\sigma}(i)})\} - f_0\{g(X_{\hat{\sigma}(i+1)})\}| \;=\; O_{\mathbb{P}}\left\{\max\{\log n, n\delta\}\;\left(\overline{\kappa}_n + 1\right)\right\}.$$

Proof By Lemma 22 and the triangle inequality, we have that, with probability approaching one,

$$\begin{aligned}\sum_{i=1}^{n} |f_l\{g(X_i)\} - f_l\{g(X_{N(i)})\}| \;&\leq\; \sum_{i=1}^{n-1} \sum_{j=\max\{1, i-C_2\max\{\log n, n\delta\}\}}^{\min\{n, i+C_2\max\{\log n, n\delta\}\}} |f_l\{g(X_{\tilde{\sigma}(j)})\} - f_l\{g(X_{\tilde{\sigma}(j+1)})\}| \\ &\leq\; [C_2\max\{\log n, n\delta\}]\left[\sum_{j=1}^{n-1} |f_l\{g(X_{\tilde{\sigma}(j)})\} - f_l\{g(X_{\tilde{\sigma}(j+1)})\}|\right].\end{aligned}$$

Furthermore,

$$\begin{aligned}&\sum_{i=1}^{n-1} |f_1\{g(X_{\tilde{\sigma}(i)})\} - f_1\{g(X_{\tilde{\sigma}(i+1)})\}| \;\leq \sum_{i=1}^{n-1} \sum_{j=\min\{\sigma^{-1}(\tilde{\sigma}(i)), \sigma^{-1}(\tilde{\sigma}(i+1))\}}^{\max\{\sigma^{-1}(\tilde{\sigma}(i)), \sigma^{-1}(\tilde{\sigma}(i+1))\}} |f_1\{g(X_{\sigma(j)})\} - f_1\{g(X_{\sigma(j+1)})\}| \\ &= \sum_{j=1}^{n} \left|\left\{i \,:\, j \in [\min\{\sigma^{-1}(\tilde{\sigma}(i)), \sigma^{-1}(\tilde{\sigma}(i+1))\}, \max\{\sigma^{-1}(\tilde{\sigma}(i)), \sigma^{-1}(\tilde{\sigma}(i+1))\}]\right\}\right| \cdot \\ &\quad |f_1\{g(X_{\sigma(j)})\} - f_1\{g(X_{\sigma(j+1)})\}|.\end{aligned} \tag{60}$$

However, if $j \in [\min\{\sigma^{-1}(\tilde{\sigma}(i)), \sigma^{-1}(\tilde{\sigma}(i{+}1))\}, \max\{\sigma^{-1}(\tilde{\sigma}(i)), \sigma^{-1}(\tilde{\sigma}(i{+}1))\}]$, then $\sigma^{-1}(\sigma(j))$ is between $\min\{\sigma^{-1}(\tilde{\sigma}(i)), \sigma^{-1}(\tilde{\sigma}(i{+}1))\}$ and $\max\{\sigma^{-1}(\tilde{\sigma}(i)), \sigma^{-1}(\tilde{\sigma}(i{+}1))\}$, $\tilde{\sigma}^{-1}(\tilde{\sigma}(i)) = i$, and $\tilde{\sigma}^{-1}(\tilde{\sigma}(i)) = i + 1$. Therefore, either $\tilde{\sigma}(i) \in \mathcal{K}_{\sigma(j)}$ or $\tilde{\sigma}(i+1) \in \mathcal{K}_{\sigma(j)}$. Hence,

$$\begin{aligned}\sum_{i=1}^{n-1} |f_1\{g(X_{\tilde{\sigma}(i)})\} - f_1\{g(X_{\tilde{\sigma}(i+1)})\}| \;&\leq\; \sum_{j=1}^{n} \left(1 + \kappa_{\sigma(j)}\right) |f_1\{g(X_{\sigma(j)})\} - f_1\{g(X_{\sigma(j+1)})\}| \\ &\leq\; (1+\kappa) \sum_{j=1}^{n} |f_1(g(X_{\sigma(j)})) - f_1(g(X_{\sigma(j+1)})), \\ &=\; O_{\mathbb{P}}(\overline{\kappa}_n + 1)\end{aligned}$$

where the last inequality follows from Assumption 5. The proof for f_0 proceeds with the same argument.

■

Lemma 24 *For any $i \in \{1, \ldots, n\}$ let*

$$\xi_i \;=\; \left|\{j \in \{1, \ldots, n\} \,:\quad \textit{and} \quad \tilde{g}(X_i) \leq \tilde{g}(X_j) \leq \tilde{g}(X_{N(i)}) \;\; \textit{or} \;\; \tilde{g}(X_{N(i)}) \leq \tilde{g}(X_j) \leq \tilde{g}(X_i)\}\right|.$$

Then, for some constant $C_2 > 0$,

$$\max_{i=1,\ldots,n} \xi_i \;\leq\; C_2 \max\{\log n, n\delta\},$$

with probability approaching one.

Proof This follows as the proof of Lemma 17. ■

Lemma 25 *With the notation from (56),*

$$\sum_{i=1}^{n} |f_l\{g(X_i)\} - f_l\{g(X_{N(i)})\}| = O_{\mathbb{P}}\left\{\max\{\log n, n\delta\}\; (\overline{\kappa}_n + 1)\right\},$$

for $l \in \{0,1\}$.

Proof By Lemma 22 and the triangle inequality, we have that, with probability approaching one,

$$\begin{aligned}\sum_{i=1}^{n} |f_l\{g(X_i)\} - f_l\{g(X_{N(i)})\}| \;&\leq\; \sum_{i=1}^{n-1} \sum_{j=\max\{1,i-C_2\max\{\log n,n\delta\}\}}^{\min\{n,i+C_2\max\{\log n,n\delta\}\}} |f_l\{g(X_{\tilde{\sigma}(j)})\} - f_l\{g(X_{\tilde{\sigma}(j+1)})\}| \\ &\leq\; C_2 \max\{\log n, n\delta\} \sum_{j=1}^{n-1} |f_l\{g(X_{\tilde{\sigma}(j)})\} - f_l\{g(X_{\tilde{\sigma}(j+1)})\}|\end{aligned}$$

and the claim follows as in Lemma 23. ■

Appendix F. Details for comparisons with Wager and Athey (2018), and Athey et al. (2019)

Scenario 1. This is the first model considered in Wager and Athey (2018) (see Equation 27 there). The data satisfies

$$\begin{array}{lcl} Y_i & = & (1-Z_i)Y_i(0) + Z_iY_i(1), \\ Z_i & \sim & \mathrm{Binom}(1, e(X_i)), \\ Y_i(0) & \sim & \mathcal{N}(2X_i^\top \boldsymbol{e}_1 - 1, 1), \\ Y_i(1) & \sim & \mathcal{N}(2X_i^\top \boldsymbol{e}_1 - 1, 1), \\ e(x) & = & \frac{1}{4}(1 + \beta_{2,4}(x_1)),\;\; \forall x \in [0,1]^d, \\ X_i & \overset{\text{ind}}{\sim} & U[0,1]^d,\;\; \forall i\{1,\ldots,n\}, \end{array}$$

where $\boldsymbol{e}_1 = (1, 0, \ldots, 1)^\top$, and $\beta_{2,4}$ is β-density with shape parameters 2 and 4. Notice that in this case $\tau_i^* = 0$ for all $i \in \{1,\ldots,n\}$.

Scenario 2. Our second scenario also comes from Wager and Athey (2018) (see Equation 28 there).

$$\begin{array}{lcl} Y_i & = & m(X_i) + (Z_i - e(X_i))\tau(X_i) + \epsilon_i, \\ Z_i & \sim & \text{Binom}(1, e(X_i)), \\ m(x) & = & e(x)\tau(x), \;\; \forall x \in [0,1]^d, \\ e(x) & = & 0.5, \;\; \forall x \in [0,1]^d, \\ \tau(x) & = & \varsigma(x_1)\varsigma(x_2), \;\; \forall x \in [0,1]^d, \\ \varsigma(u) & = & 1 + \frac{1}{1+\exp\{-20(u-\frac{1}{3})\}}, \;\; \forall u \in [0,1], \\ X_i & \overset{\text{ind}}{\sim} & U[0,1]^d, \;\; \forall i \{1,\ldots,n\}. \end{array}$$

Hence, once again $\tau_i^* = 0$ for all $i \in \{1,\ldots,n\}$.

Scenario 3. Here we generate the measurements as

$$\begin{array}{lcl} Y_i & = & (1-Z_i)Y_i(0) + Z_iY_i(1), \\ Z_i & \sim & \text{Binom}(1, e(X_i)), \\ Y_i(l) & \sim & \mathcal{N}(f_l(e(X_i)), 1), \quad \forall l \in \{0,1\}, \\ e(x) & = & \Phi(\beta^\top x), \quad \forall x \in [0,1]^d, \\ f_0(s) & = & s^2, \quad \forall s \in [0,1], \\ f_1(s) & = & s^2 + \mathbf{1}_{\{s>0.6\}}, \quad \forall s \in [0,1], \\ X_i & \overset{\text{iid}}{\sim} & U[0,1]^d, \quad \forall i \in \{1,\ldots,n\}, \end{array}$$

where $\beta \in \mathbb{R}^p$ is defined by $\beta_j = 1$ for $j \in \{1,\ldots,\lfloor p/2 \rfloor\}$, and $\beta_j = -1$ otherwise. Furthermore, Φ denotes the cumulative distribution function of the standard normal distribution. Clearly, in this case $\tau_i^* = \mathbf{1}_{\{e(X_i)>0.6\}}$ for all $i \in \{1,\ldots,n\}$.

Scenario 4. This is the model described in (10).

Appendix G. Details of comparisons with Abadie et al. (2018)

G.1 National JTPA Study

We follow the experimental setting in Abadie et al. (2018). Specifically, let the JTPA measurements be $\{(z_i^{obs}, x_i^{obs}, y_i^{obs})\}_{i=1}^{2530}$, where $y_i^{obs} \in \mathbb{R}$ corresponds to the outcome (earnings), $x_i^{obs} \in \mathbb{R}^d$ to the covariates, and $z_i^{obs} \in \{0,1\}$ to the treatment indicator. Then, to generate simulated outcomes, construct a parameter θ as in Abadie et al. (2018):

$$\theta \;=\; \underset{\beta \in \mathbb{R}^d}{\arg\min} \sum_{i=1}^{n_{obs}} \left\{ \frac{\left(y_i^{obs}\right)^\lambda - 1}{\lambda} - x_i^{obs\top}\beta \right\}^2 \mathbf{1}\left\{ z_i^{obs} = 0, y_i^{obs} > 0 \right\},$$

where $\lambda = 0.3667272$.

Furthermore, the variance of the errors is computed as:

$$\sigma^2 \;=\; \frac{1}{n_{obs} - d - 1} \sum_{i=1}^{n_{obs}} \left\{ \frac{\left(y_i^{obs}\right)^\lambda - 1}{\lambda} - x_i^{obs\top}\theta \right\}^2 \mathbf{1}\left\{ z_i^{obs} = 0, y_i^{obs} > 0 \right\}.$$

A third parameter of interest is:

$$\gamma \triangleq \underset{\beta \in \mathbb{R}^d}{\arg\max} \sum_{i=1}^{n_{obs}} \mathbf{1}\left\{ z_i^{obs} = 0 \right\} \log \left\{ \left(\frac{e^{\beta^\top x_i^{obs}}}{1 + e^{\beta^\top x_i^{obs}}} \right)^{\mathbf{1}\left\{ y_i^{obs} > 0 \right\}} \left(1 - \frac{e^{\beta^\top x_i^{obs}}}{1 + e^{\beta^\top x_i^{obs}}} \right)^{1 - \mathbf{1}\left\{ y_i^{obs} > 0 \right\}} \right\},$$

the result of fitting a logistic regression model to predict whether a unit in the experiment's control group will have positive earnings.

Next, simulation data is generated as:

$$
\begin{array}{lcl}
Y_i & = & \left(\max\left\{0, 1+\lambda\left(X_i^\top\theta+\epsilon_i\right)\right\}\times \mathbf{1}\left\{U_i>0\right\}\right)^{1/\lambda}, \\
U_i & \overset{i.i.d.}{\sim} & \text{Bernoulli}\left(\frac{e^{X_i^\top\gamma}}{1+e^{X_i^\top\gamma}}\right), \\
X_i & \overset{iid}{\sim} & \mathbb{P}\left(X=x_j^{obs};\{x_j^{obs}\}_{j=1}^{n_{obs}}\right)=\frac{1}{n_{obs}},\quad j\in\{1,\ldots,n_{obs}\}, \qquad \text{(Empirical Distribution)} \\
\epsilon_i & \overset{i.i.d.}{\sim} & \mathcal{N}\left(0,\sigma^2\right).
\end{array}
$$

Here, the treatment effect is zero. Furthermore, the treatment indicators for the simulations are such that $\sum_i Z_i = 1681$ for the training set, and $\mathbb{P}\left(Z_i = 1\right) = \frac{1681}{2530}$ for the test set.

G.2 Project STAR

With the observations $\{(z_i^{obs}, x_i^{obs}, y_i^{obs})\}_{i=1}^{3764}$ for this study, where $y_i^{obs} \in \mathbb{R}$ is the outcome variable, $x_i^{obs} \in \mathbb{R}^d$ the vector of covariates, and $z_i^{obs} \in \{0, 1\}$ the treatment assignment, we generate data following Abadie et al. (2018). Thus, measurements arise from the model

$$
\begin{array}{lcl}
Y_i & = & X_i^\top\beta_0+\epsilon_i \\
X_i & \overset{\text{ind}}{\sim} & \mathbb{P}\left(X=x_j^{obs};\{x_j^{obs}\}_{j=1}^{n_{obs}}\right)=\frac{1}{n_{obs}},\quad j\in\{1,\ldots,n_{obs}\}\ , \qquad \text{(Empirical Distribution)} \\
\epsilon_i & \overset{\text{ind}}{\sim} & \mathcal{N}\left(0,\sigma^2\right),
\end{array}
$$

where

$$
\beta_0 \;=\; \underset{\beta\in\mathbb{R}^d}{\arg\min}\sum_{i=1}^{n_{obs}}\left(y_i^{obs}-x_i^{obs\top}\beta\right)^2\mathbf{1}\left\{z_i^{obs}=0\right\},
$$

and the variance for the errors is computed as:

$$
\sigma^2 = \frac{1}{n_{obs}-p-1}\sum_{i=1}^{n_{obs}}\left(y_i^{obs}-x_i^{obs\top}\beta_0\right)^2\mathbf{1}\left\{z_i^{obs}=0\right\}.
$$

Finally, in this scenario, the treatment indicators for the simulations are such that $\sum_i W_i = \lceil n/2 \rceil$.

Appendix H. National Supported Work data

References

Alberto Abadie, Matthew M Chingos, and Martin R West. Endogenous stratification in randomized experiments. *Review of Economics and Statistics*, 100(4):567–580, 2018.

Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion.* Princeton university press, 2008.

Susan F Assmann, Stuart J Pocock, Laura E Enos, and Linda E Kasten. Subgroup analysis and other (mis) uses of baseline data in clinical trials. *The Lancet*, 355(9209):1064–1069, 2000.

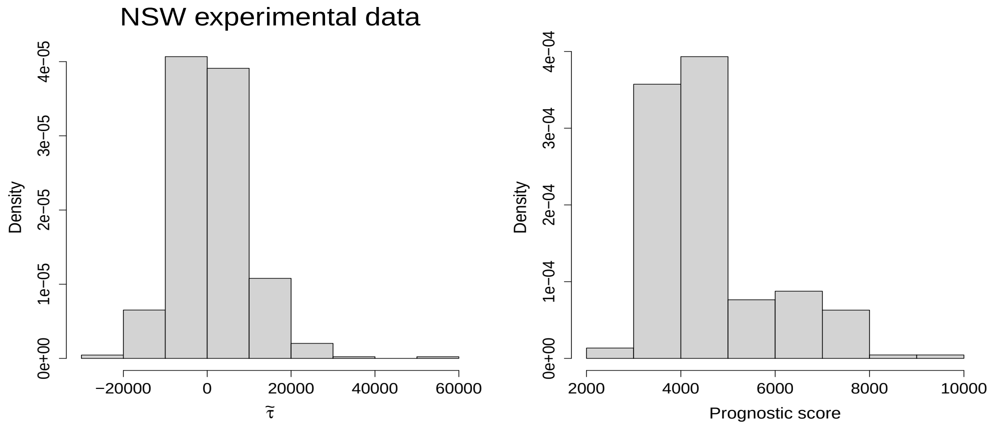

Figure 5: Data and estimates for the NSW example described in Section 4.2.1. From left to right the two panels show a histogram of $\tilde{\tau}^{(1)}$'s (the scores obtained after matching) and the estimated prognostic scores.

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

Susan Athey, Julie Tibshirani, Stefan Wager, et al. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.

Alvaro Barbero and Suvrit Sra. Modular proximal optimization for multidimensional total-variation regularization. *arXiv preprint arXiv:1411.0589*, 2014.

Howard S Bloom, Larry L Orr, Stephen H Bell, George Cave, Fred Doolittle, Winston Lin, and Johannes M Bos. The benefits and costs of jtpa title ii-a programs: Key findings from the national job training partnership act study. *Journal of human resources*, pages 549–576, 1997.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

L Breiman, JH Friedman, RA Olshen, and CJ Stone. Classification and regression trees (cart). *Wadsworth, Monterey, CA, USA*, 1984.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Alberto Caron, Gianluca Baio, and Ioanna Manolopoulou. Estimating individual treatment effects using non-parametric regression models: A review. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 185(3):1115–1149, 2022.

Kwun Chuen Gary Chan, Sheung Chi Phillip Yam, and Zheng Zhang. Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 78 (3):673, 2016.

Sabyasachi Chatterjee and Subhajit Goswami. New risk bounds for 2d total variation denoising. *arXiv preprint arXiv:1902.01215*, 2019.

Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266 – 298, 2010a. doi: 10.1214/09-AOAS285. URL `https://doi.org/10.1214/09-AOAS285`.

Hugh A Chipman, Edward I George, Robert E McCulloch, et al. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010b.

David I Cook, Val J Gebski, and Anthony C Keech. Subgroup analysis in clinical trials. *Medical Journal of Australia*, 180(6):289, 2004.

Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. Nonparametric tests for treatment effect heterogeneity. *The Review of Economics and Statistics*, 90(3): 389–405, 2008.

Rajeev H Dehejia and Sadek Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association*, 94(448):1053–1062, 1999.

Rajeev H Dehejia and Sadek Wahba. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and statistics*, 84(1):151–161, 2002.

Alexis Diamond and Jasjeet S Sekhon. Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics*, 95(3):932–945, 2013.

Peng Ding, Avi Feller, and Luke Miratrix. Randomization inference for treatment effect variation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78 (3):655–671, 2016.

Peng Ding, Fan Li, et al. Causal inference: A missing data perspective. *Statistical Science*, 33(2):214–237, 2018.

David L Donoho and Iain M Johnstone. Minimax estimation via wavelet shrinkage. *The annals of Statistics*, 26(3):879–921, 1998.

Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution*, pages 569–593. Springer, 1992.

Zijun Gao and Yanjun Han. Minimax optimal nonparametric estimation of heterogeneous treatment effects. *arXiv preprint arXiv:2002.06471*, 2020.

Adam N Glynn and Kevin M Quinn. An introduction to the augmented inverse propensity weighted estimator. *Political analysis*, 18(1):36–56, 2010.

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

Adityanand Guntuboyina, Donovan Lieu, Sabyasachi Chatterjee, Bodhisattva Sen, et al. Adaptive risk bounds in univariate total variation denoising and trend filtering. *The Annals of Statistics*, 48(1):205–229, 2020.

László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A distribution-free theory of nonparametric regression.* Springer Science & Business Media, 2006.

P Richard Hahn, Jared S Murray, Carlos M Carvalho, et al. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. *Bayesian Analysis*, 2020.

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

James J Heckman, Hedibert F Lopes, and Rémi Piatek. Treatment effects: A bayesian perspective. *Econometric reviews*, 33(1-4):36–67, 2014.

Jennifer Hill and Yu-Sung Su. Assessing lack of common support in causal inference using bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children's cognitive outcomes. *The Annals of Applied Statistics*, pages 1386–1420, 2013.

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

Jan-Christian Hütter and Philippe Rigollet. Optimal rates for total variation denoising. In *Conference on Learning Theory*, pages 1115–1146, 2016.

Kosuke Imai, Marc Ratkovic, et al. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470, 2013.

Guido W Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *The review of Economics and Statistics*, 86(1):4–29, 2004.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences.* Cambridge university press, 2015.

Nicholas A Johnson. A dynamic programming algorithm for the fused lasso and l 0-segmentation. *Journal of Computational and Graphical Statistics*, 22(2):246–260, 2013.

Srikar Katta, Harsh Parikh, Cynthia Rudin, and Alexander Volfovsky. Interpretable causal inference for analyzing wearable, sensor, and distributional data. In *International Conference on Artificial Intelligence and Statistics*, pages 3340–3348. PMLR, 2024.

Alan B Krueger. Experimental estimates of education production functions. *The quarterly journal of economics*, 114(2):497–532, 1999.

Ravi Kumar and Sergei Vassilvitskii. Generalized distances between rankings. In *Proceedings of the 19th international conference on World wide web*, pages 571–580, 2010.

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

Robert J LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, pages 604–620, 1986.

Quinn Lanners, Harsh Parikh, Alexander Volfovsky, Cynthia Rudin, and David Page. Variable importance matching for causal inference. In *Uncertainty in Artificial Intelligence*, pages 1174–1184. PMLR, 2023.

Yameng Liu, Aw Dieng, Sudeepa Roy, Cynthia Rudin, and Alexander Volfovsky. Interpretable almost matching exactly for causal inference. *arXiv preprint arXiv:1806.06802*, 2018.

Enno Mammen, Sara van de Geer, et al. Locally adaptive regression splines. *The Annals of Statistics*, 25(1):387–413, 1997.

Marco Morucci, Vittorio Orlandi, Sudeepa Roy, Cynthia Rudin, and Alexander Volfovsky. Adaptive hyper-box matching for interpretable individualized treatment effect estimation. In *Conference on Uncertainty in Artificial Intelligence*, pages 1089–1098. PMLR, 2020.

Marco Morucci, Cynthia Rudin, and Alexander Volfovsky. Matched machine learning: A generalized framework for treatment effect inference with learned metrics. *arXiv preprint arXiv:2304.01316*, 2023.

Jersey Neyman. Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10:1–51, 1923.

Larry L Orr. *Does training for the disadvantaged work?: Evidence from the National JTPA study*. The Urban Insitute, 1996.

Francesco Ortelli and Sara van de Geer. Synthesis and analysis in total variation regularization. *arXiv preprint arXiv:1901.06418*, 2019.

Carlos Misael Madrid Padilla, Oscar Hernan Madrid Padilla, Yik Lun Kei, Zhi Zhang, and Yanzhen Chen. Confidence interval construction and conditional variance estimation with dense relu networks. *arXiv preprint arXiv:2412.20355*, 2024.

Oscar Hernan Madrid Padilla, James Sharpnack, and James G Scott. The dfs fused lasso: Linear-time denoising over general graphs. *The Journal of Machine Learning Research*, 18(1):6410–6445, 2018.

Oscar Hernan Madrid Padilla, James Sharpnack, Yanzhen Chen, and Daniela M Witten. Adaptive non-parametric regression with the k-nn fused lasso. *Biometrika*, 2020.

Harsh Parikh, Cynthia Rudin, and Alexander Volfovsky. Malts: Matching after learning to stretch. *Journal of Machine Learning Research*, 23(240):1–42, 2022.

Stuart J Pocock, Susan E Assmann, Laura E Enos, and Linda E Kasten. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practiceand problems. *Statistics in medicine*, 21(19):2917–2930, 2002.

Pradeep Ravikumar, Martin J Wainwright, John D Lafferty, et al. High-dimensional ising model selection using ℓ_1-regularized logistic regression. *The Annals of Statistics*, 38(3): 1287–1319, 2010.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Paul R Rosenbaum and Donald B Rubin. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79(387):516–524, 1984.

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

Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268, 1992.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3076–3085. JMLR. org, 2017.

Xiaogang Su, Chih-Ling Tsai, Hansheng Wang, David M Nickerson, and Bogong Li. Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, 10(Feb): 141–158, 2009.

Vasilis Syrgkanis, Victor Lei, Miruna Oprescu, Maggie Hei, Keith Battocchi, and Greg Lewis. Machine learning estimation of heterogeneous treatment effects with instruments. In *Advances in Neural Information Processing Systems*, pages 15167–15176, 2019.

Matt Taddy, Matt Gardner, Liyun Chen, and David Draper. A nonparametric bayesian analysis of heterogenous treatment effects in digital experimentation. *Journal of Business & Economic Statistics*, 34(4):661–672, 2016.

Lu Tian, Ash A Alizadeh, Andrew J Gentles, and Robert Tibshirani. A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109(508):1517–1532, 2014.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58(1):267–288, 1996.

Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

Ryan J Tibshirani, Jonathan Taylor, et al. Degrees of freedom in lasso problems. *The Annals of Statistics*, 40(2):1198–1232, 2012.

Ryan J Tibshirani et al. Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics*, 42(1):285–323, 2014.

Sara Van de Geer. Estimating a regression function. *The Annals of Statistics*, pages 907–924, 1990.

Joaquin Vanschoren. Meta-learning. In *Automated machine learning: methods, systems, challenges*, pages 35–61. Springer International Publishing Cham, 2019.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Ulrike Von Luxburg, Agnes Radl, and Matthias Hein. Hitting and commute times in large random neighborhood graphs. *The Journal of Machine Learning Research*, 15(1):1751–1798, 2014.

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

Grace Wahba. Soft and hard classification by reproducing kernel hilbert space methods. *Proceedings of the National Academy of Sciences*, 99(26):16524–16530, 2002.

Daren Wang, Yi Yu, and Alessandro Rinaldo. Optimal covariance change point detection in high dimension. *arXiv preprint arXiv:1712.09912*, 2017.

Tianyu Wang, Marco Morucci, M Usaid Awan, Yameng Liu, Sudeepa Roy, Cynthia Rudin, and Alexander Volfovsky. Flame: A fast large-scale almost matching exactly approach to causal inference. *Journal of Machine Learning Research*, 22(31):1–41, 2021.

Yu-Xiang Wang, James Sharpnack, Alexander J Smola, and Ryan J Tibshirani. Trend filtering on graphs. *The Journal of Machine Learning Research*, 17(1):3651–3691, 2016.

Herbert I Weisberg and Victor P Pontes. Post hoc subgroups in clinical trials: Anathema or analytics? *Clinical trials*, 12(4):357–364, 2015.

Halbert White. Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society*, pages 1–25, 1982.

Richard J Willke, Zhiyuan Zheng, Prasun Subedi, Rikard Althin, and C Daniel Mullins. From concepts, theory, and evidence of heterogeneity of treatment effects to methodological approaches: a primer. *BMC medical research methodology*, 12:1–12, 2012.

Vivian C Wong. *Addressing theoretical and practical challenges in the regression-discontinuity design.* PhD thesis, Northwestern University, 2010.

Qingyuan Zhao, Dylan S Small, and Ashkan Ertefaie. Selective inference for effect modification via the lasso. *arXiv preprint arXiv:1705.08020*, 2017.