Title

Locally Adaptive Statistical Models with Applications in Quantile Regression and Causal Inference

Author

Ye, Siwei

Publication Date

2025

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Locally Adaptive Statistical Models
With Applications in Quantile Regression and Causal Inference

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Statistics

by

Steven Ye

2025

ABSTRACT OF THE DISSERTATION

Locally Adaptive Statistical Models
With Applications in Quantile Regression and Causal Inference

by

Steven Ye
Doctor of Philosophy in Statistics
University of California, Los Angeles, 2025
Professor Oscar Hernan Madrid Padilla, Chair

Local patterns, such as piecewise constant structures, frequently appear in real-world data. Many applications demand models capable of capturing these localized variations, where data may change abruptly or remain constant across different local regions. Statistical models such as fused lasso and nearest-neighbor algorithms are particularly effective in addressing this localized complexity. The fused lasso encourages sparsity in differences between adjacent observations over a graph structure, making it well-suited for identifying and estimating local patterns. Nearest-neighbor algorithms focus on nearby data points, offering a flexible and locally adaptive framework to capture varying structures within the data.

This dissertation develops innovative statistical models that integrate fused lasso and nearest-neighbor algorithms to address research problems in quantile regression and causal inference, as demonstrated in the following projects. The first project introduces a non-parametric quantile regression framework using the K-nearest-neighbor fused lasso to provide robust and locally adaptive estimation of multivariate functions at different quantile levels. The second project presents a score-based approach for estimating heterogeneous

treatment effects, combining propensity and prognostic scores with nearest-neighbor matching to stratify and estimate treatment effects over a two-dimensional grid. The third project proposes graph-based fused lasso models to estimate heterogeneous treatment effects over graph structures. These projects highlight the versatility and effectiveness of locally adaptive models that build on fused lasso and nearest-neighbor algorithms. Validated through simulation studies and real-world applications, these models offer valuable insight for fields such as economics, medicine, and social science.

The dissertation of Steven Ye is approved.

Arash A. Amini

Yanzhen Chen

Jingyi Li

Qing Zhou

Oscar Hernan Madrid Padilla, Committee Chair

University of California, Los Angeles

2025

To my parents

and all who have supported me

and those I hold dear.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I could not have completed this five-year Ph.D. journey without the support and guidance of many brilliant people. I am deeply grateful to everyone who has supported and encouraged me along the way.

First and foremost, I would like to express my greatest gratitude to my advisor, Dr. Oscar Hernan Madrid Padilla. Oscar's unwavering support, insightful guidance, and meticulous attention to detail throughout my five-year Ph.D. journey were invaluable. Oscar not only taught me how to conduct research, but also demonstrated what it means to be a passionate scientist dedicated to one's field.

I also extend my thanks to my distinguished committee members: Dr. Arash A. Amini, Dr. Jingyi Li, and Dr. Qing Zhou. Their expertise, constructive feedback, and guidance were crucial to the development of my research. I am especially grateful to Dr. Yanzhen Chen for her kind support and warmhearted dedication to my projects over the past few years.

My heartfelt thanks also go to my schoolmates: Jialiang Geng, Jiamin Guo, Kaiwen Jiang, Allen Kei, Luciano Vinas, Dehong Xu, Tianyang Zhao, and Yaxuan Zhu, for their ongoing support and companionship during this journey. I am also very grateful to my dear friends, Xiecheng Shao, Xinchen Xu, Liyang Zhao, Sihao Chen, Xiaoxun Liu, Ke Ning, Daoye Wang, Jingting Zhang, Tao Lan, Yifei Xu, and Yuqi Zhang, for their unwavering friendship and encouragement. Your presence brought joy and positivity to these years.

Lastly, my deepest and sincere thanks go to my parents. Their unconditional support, encouragement, and belief in me, from the day I was born until now, made this achievement possible. I cannot imagine completing this long academic path, culminating in a doctorate degree, without their steadfast guidance and love throughout every moment of my life.

Vita

2015-2019	B.A. in Applied Mathematics and Statistics, UC Berkeley
2018	Actuarial Analyst Intern, Swiss Re America.
2020-2023	Teaching Fellow, UCLA Department of Statistics and Data Science.
2022	Machine Learning and AI Intern, Nokia Bell Labs.
2023	Machine Learning Scientist Intern, Netflix.
2024	Data Scientist Research Intern, Google.

Publications

S. Ye and O.H.M. Padilla. "Non-parametric Quantile Regression via the K-NN Fused Lasso". *Journal of Machine Learning Research*, Vol. 22, No. 111, 1-38, 2021.

S. Ye, Y. Chen, and O.H.M. Padilla. "2D Score Based Estimation of Heterogeneous Treatment Effects". *Journal of Causal Inference*, Vol. 11, No. 1, 1-26, 2023.

A.K. Glazer, H. Luo, S. Devgon, C. Wang, X. Yao, S. Ye, F. McQuarrie, Z. Li, A. Palma, Q. Wan, W. Gu, A. Sen, Z. Wang, G.D. O'Connell, P.B. Stark. "Look Who's Talking: Gender Differences in Academic Job Talks". *ScienceOpen Research*, 2023.

CHAPTER 1

Introduction

1.1 Background

In real-world scenarios, data often exhibit some underlying local patterns that are of significant interest to researchers. Among these patterns, piecewise constant structures are commonly observed. Such local structures appear in situations where a value remains constant within specific time frames or regions before abruptly changing to a new constant value at distinct boundaries. For example, stock prices may remain steady during a trading session, but shift sharply at market open. Mathematically, for one-dimensional data $x = (x_1, \ldots, x_n)$ and a corresponding response $y = f_0(x)$ generated by an underlying function f_0, a piecewise constant structure is characterized by:

$$f_0(x_i) = f_0(x_{i+1}) \quad \text{or} \quad f_0(x_i) \approx f_0(x_{i+1}) \quad \text{for } i = 1, \ldots, n-1.$$

For multi-dimensional data $X = (x_1, \ldots, x_n)$, such local patterns are typically explained using a graph structure denoted as $G = (V, E)$, where $V = \{1, \ldots, n\}$ represents the vertex set and E the edge set. An edge $(i, j) \in E$ indicates a connection between observations x_i and x_j in the graph. Then, a piecewise constant structure over the graph G enforces:

$$f_0(x_i) = f_0(x_j) \quad \text{or} \quad f_0(x_i) \approx f_0(x_j) \quad \text{for } (i, j) \in E.$$

The fused lasso, also referred to as zeroth-order trend filtering or total variation denoising [ROF92, MG97, TSR05, KKB09, WSS16], is a statistical method specifically designed for problems where the data exhibit piecewise constant or similar locally structured patterns.

In a standard non-parametric regression setting:

$$y_i = f_0(x_i) + \epsilon_i,$$

where ϵ_i represents the error term, the objective is to estimate $\hat{\theta}$ that approximates the underlying function $f_0(x_i)$. The fused lasso estimator, denoted by $\hat{\theta}^{FL}$, is formulated as:

$$\hat{\theta}^{FL} = \underset{\theta \in \mathbb{R}^n}{\arg\min} \left\{ \sum_{i=1}^{n} (y_i - \theta_i)^2 + \lambda \sum_{(i,j) \in E} |\theta_i - \theta_j| \right\}.$$

This formulation penalizes the ℓ_1-norm of differences between adjacent observations over an underlying graph structure, thus enforcing the sparsity in these adjacent differences and encouraging local constancy in the estimation. The fused lasso has been extensively studied both computationally and theoretically with a variety of graph structures, including chain graphs [DK01, Joh13, BS18], tree graphs [PSS18, OG18], grid graphs [HR16, SWT16, XST17], as well as general graphs [CD09, Hoe10, CP11, TT11].

Nearest-neighbor algorithms [FH51], on the other hand, offer an alternative approach for localized estimation. These algorithms are manifold adaptive [Kpo11], make them well-suited for capturing complex, non-linear relationships and are especially effective in high-dimensional settings where the underlying structure is difficult to specify. In particular, nearest-neighbor methods are inherently local [Die93], as they focus solely on the locally adjacent neighborhood of each data point, making them adaptable to the local structure of the data. In regression analysis, K-nearest-neighbor (K-NN) regression [Das91] is a popular method that predicts the value of a target variable by averaging the target values of the nearest data points. Specifically, in a regression problem with a response variable y and observations X, if we focus on a given data point $x_i \in X$, the algorithm identifies its K-nearest neighbors, denoted by $\{x_i^K\} \subseteq X \setminus \{x_i\}$, using a predefined distance metric. The value of the function at x_i is then estimated as:

$$\hat{f}(x_i) = \frac{1}{K} \sum_{j \in \{x_i^K\}} y_j.$$

Beyond regression, nearest-neighbor algorithms are widely used in classification [CH67, HKK01, GWB03] and matching [Rub73, Stu10, LDH23].

1.2 Research Objective

This dissertation aims to advance the development of models that integrate fused lasso and nearest-neighbor algorithms, with a particular emphasis on addressing the presence of local patterns in data commonly observed in real-world scenarios. Focusing specifically on quantile regression and causal inference, we seek to develop novel statistical models to highlight both the effectiveness and potential of the new methodologies that incorporate both statistical methods.

1.3 Dissertation Outline

The remaining part of the dissertation is organized as follows:

- In Chapter 2, we present a novel adaptive quantile estimator, named as the quantile K-nearest-neighbor fused lasso. The new estimator is developed within a non-parametric framework, and it incorporates a fused lasso penalty over a K-nearest-neighbor graph. We establish the estimator's consistency under mild assumptions for d-dimensional data, and prove that it achieves the optimal convergence rate of $n^{-1/d}$, up to a logarithmic factor. Moreover, we propose two efficient algorithms for its computation and explore methodologies for model selection. We conduct thorough numerical experiments on simulated and real-world datasets, and we demonstrate the superior performance of the proposed estimator compared to existing state-of-the-art methods.

- In Chapter 3, we introduce a new framework for estimating the conditional average treatment effect (CATE) function by integrating propensity and prognostic scores with a nearest-neighbor matching algorithm and non-parametric regression trees. Unlike traditional methods that rely on multivariate feature sets, this approach estimates the CATE over a two-dimensional grid defined by the two scores, providing a more interpretable analysis. We perform extensive numerical experiments on simulated datasets to demonstrate the advantages of this method compared to existing models. Further-

more, we apply the proposed framework to real-world datasets from clinical and social sciences and provide novel insights.

- In Chapter 4, we develop innovative estimators for heterogeneous treatment effects based on fused lasso over graph structures. We introduce efficient algorithms for computing these estimators and prove that they achieve rapid convergence rates under simple data assumptions. We highlight the superior numerical performance of the proposed methods compared to conventional causal effect estimation methods through simulation studies. We also apply these methods to a case study in criminology and present important findings supported by numerical results.

- In Chapter 5, we summarize the key research contributions presented in the preceding chapters, and we discuss potential future work to build on the projects completed in this dissertation.

CHAPTER 2

Non-parametric Quantile Regression via the K-NN Fused Lasso

2.1 Introduction

Assume that we have n observations, $(x_1, y_1), ..., (x_n, y_n)$, of the pair of random variables (X, Y). The response variable Y is a real-valued vector and X is a multivariate covariate or predictor variable in a metric space $\mathcal{X}$ with metric $d_{\mathcal{X}}$. A standard goal of non-parametric regression is to infer, in some way, the underlying relationship between Y and X. The generative model behind this can be expressed as

$$y_i = f_0(x_i) + \epsilon_i, \text{ for } i = 1, ..., n, \tag{2.1}$$

where f_0 is an unknown function that we want to estimate. While usual regression considers estimation of the conditional mean of the response variable, quantile regression estimates the conditional median (or other desired quantiles) of the response variable. Specifically, given a quantile level $\tau \in (0, 1)$, we can rewrite (2.1) as

$$y_i = \theta_i^* + \epsilon_i, \text{ for } i = 1, ..., n, \tag{2.2}$$

where

$$\theta_i^* = F_{y_i|x_i}^{-1}(\tau). \tag{2.3}$$

Here, θ^* is the vector of τ-quantiles of y, $F_{y_i|x_i}$ represents the cumulative distribution function of y_i given x_i, and $\mathbb{P}(\epsilon_i \leq 0|x_i) = \tau$.

The goal of quantile regression is to estimate θ^* as accurately as possible, and it usually involves an optimization problem in the form

$$\hat{\theta} \in \underset{\theta \in \zeta \subset \mathbb{R}^n}{\arg\min}\ L(\theta), \tag{2.4}$$

where $L(\theta)$ is the loss function is defined by

$$L(\theta) = \sum_{i=1}^{n} \rho_\tau(y_i - \theta_i), \tag{2.5}$$

with $\rho_\tau(t) = (\tau - 1\{t \leq 0\})t$, the asymmetric absolute deviation function [KB78].

In this chapter, we apply total variation denoising for non-parametric quantile regression in a multivariate setup and combine it with the K-NN procedure. We leverage insights gained from recent results for quantile trend filtering [PC22] and K-NN fused lasso [PSC20]. Our proposed estimator, quantile K-NN fused lasso, can adapt to the piecewise linear or piecewise polynomial structure in the true vector properly.

It takes simply two steps to compute our proposed quantile K-NN fused lasso estimator. We first construct a K-nearest-neighbor graph corresponding to the given observations. The second step involves a constrained optimization problem with a lasso-type penalty along the K-NN graph as follow

$$\hat{\theta} \in \underset{\theta \in \zeta \subset \mathbb{R}^n}{\arg\min}\ L(\theta) + \lambda \|\nabla_G \theta\|_1, \tag{2.6}$$

where $\lambda > 0$ is a tuning parameter. The notation ∇_G in the penalty term represents the oriented incidence matrix of the K-NN graph, and we will provide the detailed definition in Section 2.2.

We study the rate of convergence of the estimator defined in (2.6). Towards that end, we denote a loss function $\Delta_n^2 : \mathbb{R}^n \to \mathbb{R}$ by

$$\Delta_n^2(\delta) := \frac{1}{n} \sum_{i=1}^{n} \min\{|\delta_i|, \delta_i^2\}.$$

The function $\Delta_n^2(\cdot)$ appeared in [PC22] and is similar to a Huber loss function [see page 471 in Wai18], which offers a compromise between the least-squares cost and the l_1-norm cost function and is thus less sensitive to outliers in data. We show that under mild conditions,

(2.6) attains a convergence rate of $n^{-1/d}$ for d-dimensional data in terms of Δ_n^2, ignoring the logarithmic factor. The rate is nearly minimax and it matches the mean squared error rates from [PSC20]. However, unlike [PSC20], our result holds under general errors allowing for heavy-tailed distributions. Another notable point in our theoretical analysis, different from previous quantile regression work, is that we only require a bounded variation class of signals and hence guarantee our result to hold under very general models.

2.1.1 Previous Work

Quantile regression, since introduced by [KB78], has become a commonly used class of methods in many applications thanks to its flexibility and robustness for modelling conditional distributions. The study of quantile regression in non-parametric setups can be dated back to [Utr81], [Cox83] and [Eub88], whose works were mainly developed for median regression on one-dimensional data. Later, [KNP94] proposed a more general estimator for any desired quantile τ, quantile smoothing spline, and [HNP98] provided a bivariate version of the estimator. Quantile smoothing spline problems are of a l_1 penalty structure, which led to a more general study on l_1-norm regularized quantile regression by [LZ08]. Other methods for non-parametric quantile regression have also been proposed in the literature. [YJ98], [CX08], and [SWH13] explored local polynomial quantile regression. [BCC19] studied non-parametric series quantile regression, and [Mei06] introduced quantile random forests. Quantile regression with rectified linear unit (ReLU) neural networks is another edge-cutting approach that utilizes some knowledge from the field of deep learning, and the theory is studied in [PTC22].

Our approach exploits the local adaptivity of total variation denoising. [ROF92] first proposed total variation denoising for the application of image processing, and [TSR05] studied fused lasso thoroughly. Later, [KKB09] extended the discussion to the so-called trend filtering on one-dimensional data, and [WSS16] provided a generalization of trend filtering to the setting of estimation on graphs. These problems can be formulated similarly

into an optimization problem of the form

$$\hat{\theta} \in \underset{\theta \in \mathbb{R}^n}{\arg\min} \ \frac{1}{2} \sum_{i=1}^{n} (\theta_i - y_i)^2 + \lambda \|D\theta\|_1, \tag{2.7}$$

where $\lambda > 0$ is a regularization parameter to be chosen carefully and D is a matrix. For instance, trend filtering order k (fused lasso $k = 0$) consists of matrices D that capture the $(k+1)$th order total variation of a given signal, see [Tib14]. A substantial amount of literature focused on theoretical guarantees of these estimators. [MG97] and [Tib14] showed that trend filtering attains nearly minimax rates in mean squared error (MSE) for estimating functions of bounded variation. [HR16] showed a sharp oracle rate of total variation denoising along grid graphs. More recently, [PSC20] incorporated fused lasso with the K-NN procedure and proved that the K-NN fused lasso also achieves nearly minimax convergence rate. In the quantile regression literature, [BC11], [Kat11], [FFB14] studied quantile model selection via l_1 regularization, but most of these works required strict linear assumptions. [PC22] provided proofs for theoretical properties of quantile trend filtering estimator in one dimension. They showed that under minimal assumptions of the data generation mechanism, quantile trend filtering estimator attains a minimax rate for one-dimensional piecewise polynomial regression.

On the computational side, quantile regression is different from l_2 based methods because it requires a non-trivial reformulation of the optimization problem due to non-differentiability of the loss as (2.5). The most well-known algorithm for computing a quantile estimator is due to [Koe05] and it uses an interior point (IP) approach. [PGK17] studied high-dimensional quantile regression problems and obtained estimators by applying the alternating direction method of multipliers [ADMM; BPC11], majorize-minimize [MM; HL00], and coordinate descent [CD; WL08] algorithms for variable selection. For computing trend filtering estimates, [HL17] developed a fast algorithm for quantile fused lasso in $O(n \log n)$ operations. Recently, [BGC20] proposed an ADMM based algorithm for computing kth order quantile trend filtering estimators for one-dimensional data.

2.1.2 Outline

In Section 2.2, we provide the definition of quantile K-nearest-neighbors fused lasso estimator and the constrained version of the problem. Two algorithms to compute the proposed estimators numerically – alternating directions method of multipliers (ADMM), and majorize-minimize (MM), are introduced in Section 2.3, and the discussion on how to select an appropriate penalty parameter in practice is also included. Section 2.4 presents the two theoretical developments regarding the constrained and penalized estimators. The theorems demonstrate that under general assumptions, both estimators converge at a rate of $n^{-1/d}$, up to a logarithmic factor, for estimating d-dimensional data under the loss function Δ_n^2 defined above. Section 2.5 lists the results of numerical experiments on multiple simulated datasets and two real datasets, California housing data and Chicago crime data. The experiments show that the proposed estimator outperform state-of-the-art methods on both simulated and real datasets. Moreover, the comparison on accuracy and computational time among the algorithms introduced in Section 2.3 provides the audience with some insights on choosing a suitable algorithm for specific problems. The proofs of theorems in the chapter are provided in the Appendix.

2.2 Quantile K-NN Fused Lasso

Total variation denoising, as formulated in (2.7), is highly flexible, which can accommodate connected graphs of arbitrary structures that capture local data patterns. However, a common challenge in real-world applications is the absence of a predefined graph structure, making the construction of an appropriate connected graph a crucial step. When a covariate X is observed, a natural approach is to construct a K-NN graph based on the covariate. The K-NN graph effectively preserves local smoothness by ensuring that neighboring data points remain well-connected through a sufficient number of edges. In this section, we introduce the method for constructing an estimator that integrates the K-NN graph into non-parametric quantile regression.

The first step to construct the quantile K-NN fused lasso estimator is to build a K-NN graph G. Specifically, given the observations, G has vertex set $V = \{1, ..., n\}$, and its edge set E_K contains the pair (i, j), for $i \in V$, $j \in V$, and $i \neq j$, if and only if x_i is among the K-nearest neighbors of x_j, with respect to the metric $d_{\mathcal{X}}$, or vice versa. After constructing the K-NN graph, we can formalize an optimization problem for quantile K-NN fused lasso as

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^n}{\arg\min} \sum_{i=1}^{n} \rho_\tau(y_i - \theta_i) + \lambda \|\nabla_G \theta\|_1, \tag{2.8}$$

where $\lambda > 0$ is a tuning parameter, and ∇_G is an oriented incidence matrix of the K-NN graph G. Thus, we define ∇_G as follows: each row of the matrix corresponds to one edge in G; for instance, if the p-th edge in G connects the i-th and j-th observations, then

$$(\nabla_G)_{p,q} = \begin{cases} 1 & \text{if } q = i, \\ -1 & \text{if } q = j, \\ 0 & \text{otherwise.} \end{cases}$$

In this way, the p-th element in $\nabla_G \theta$, $(\nabla_G \theta)_p = \theta_i - \theta_j$. Notice that we choose the ordering of the nodes and edges in ∇_G arbitrarily without loss of generality.

Once $\hat{\theta}$ in (2.8) has been computed, we can predict the value of response corresponding to a new observation $x \in \mathcal{X} \backslash \{x_1, ..., x_n\}$ by the averaged estimated response of the K-nearest neighbors of x in $\{x_1, ..., x_n\}$. Mathematically, we write

$$\hat{y} = \frac{1}{K} \sum_{i=1}^{n} \hat{\theta}_i \cdot \mathbf{1}\{x_i \in \mathcal{N}_K(x)\}, \tag{2.9}$$

where $\mathcal{N}_K(x)$ is the set of K-nearest neighbors of x in the training data. A similar prediction rule was used in [PSC20].

A related estimator to the penalized estimator $\hat{\theta}$ defined in (2.8) is the the constrained estimator $\hat{\theta}_C$, of which the corresponding optimization problem can be written as

$$\begin{aligned} \hat{\theta}_C \quad &= \quad \underset{\theta \in \mathbb{R}^n}{\arg\min} \sum_{i=1}^{n} \rho_\tau(y_i - \theta_i) \\ &\qquad \text{subject to} \quad \|\nabla_G \theta\|_1 \leq C, \end{aligned} \tag{2.10}$$

for some positive constant C.

2.2.1 Comparison with the K-NN Fused Lasso Estimator

Before proceeding to study the properties of our proposed method, we provide some comparisons with its precursor, the K-NN fused lasso estimator from [PSC20]. The latter is defined as

$$\tilde{\theta} = \underset{\theta \in \mathbb{R}^n}{\arg\min} \sum_{i=1}^{n} (y_i - \theta_i)^2 + \lambda \|\nabla_G \theta\|_1 \tag{2.11}$$

where $\lambda > 0$ is a tuning parameter.

In contrasting (2.11) with (2.8) we first highlight that from a practical perspective the latter has some important advantages. Firstly, (2.8) can be used to construct prediction intervals whereas (2.11) can only provide point estimates for the different locations. We illustrate the construction of prediction intervals with a real data example in Section 2.5.2. Secondly, the quantile K-NN fused lasso is by construction expected to be more robust to heavy tails and outliers than its counterpart the K-NN fused lasso. We verify this in our experiments section.

On the computational side, the algorithms for solving the optimization problem in (2.11) cannot be used for finding the estimator in (2.8). Hence, novel algorithms are needed to efficiently compute our proposed estimator. This is the subject of the next section.

Finally, despite the similarity in the definition of the estimators in (2.11) and (2.8), the theory from [PSC20] does not directly translate to analyze our estimator defined in (2.8). Hence, one of the contributions of this chapter is to show that the quantile K-NN fused lasso inherits local adaptivity properties of the K-NN fused lasso in general settings that allow for heavy-tailed error distributions.

2.3 Algorithms and Model Selection

To compute the quantile K-NN fused lasso estimator, the first step is to construct the K-NN graph from the data. The computational complexity of constructing the K-NN graph is of $O(n^2)$, although it is possible to accelerate the procedure to $O(n^t)$ for some $t \in (1, 2)$ using

divide and conquer methods [Ben80, CFS09].

The second step of computation is to solve a constrained optimization problem as (2.8). Here, we introduce three algorithms to solve the problem numerically. Before presenting our algorithms, we stress that both the problems (2.8) and (2.10) are linear programs, therefore we can use any linear programming software to obtain an optimal solution. Noticeably, we can take the advantage of sparsity in the penalty matrix for faster computation. However, a shortcoming of linear programming is that the algorithm can become very time-consuming for large sized problems, especially when n is greater than 5000.

2.3.1 Alternating Directions Method of Multipliers (ADMM)

The alternating directions method of multipliers (ADMM) algorithm [BPC11] is a powerful tool for solving constrained optimization problems.

We first reformulate the optimization problem (2.8) as

$$\begin{aligned} &\underset{\theta\in\mathbb{R}^n, z\in\mathbb{R}^n}{\text{minimize}} \ \sum_{i=1}^{n} \rho_\tau(y_i-\theta_i) + \lambda\|\nabla_G z\|_1 \\ &\text{s.t.} \qquad\qquad z=\theta, \end{aligned} \tag{2.12}$$

and the augmented Lagrangian can then be written as

$$L_R(\theta, z, u) = \sum_{i=1}^{n} \rho_\tau(y_i-\theta_i) + \lambda\|\nabla_G z\|_1 + \frac{R}{2}\|\theta - z + u\|^2,$$

where R is the penalty parameter that controls step size in the update. Thus we can solve (2.12) by iteratively updating the primal and dual

$$\theta \leftarrow \underset{\theta\in\mathbb{R}^n}{\arg\min} \ \left\{ \sum_{i=1}^{n} \rho_\tau(y_i-\theta_i) + \frac{R}{2}\|\theta - z + u\|^2 \right\}, \tag{2.13}$$

$$z \leftarrow \underset{z\in\mathbb{R}^n}{\arg\min} \ \left\{ \frac{1}{2}\|\theta + u - z\|^2 + \frac{\lambda}{R}\|\nabla_G z\|_1 \right\}. \tag{2.14}$$

The primal problem (2.13) can be solved coordinate-wise in closed form as

$$\theta_i = \begin{cases} z_i - u_i + \frac{\tau}{R} & \text{if } y_i - z_i + u_i > \frac{\tau}{R}, \\ z_i - u_i + \frac{\tau - 1}{R} & \text{if } y_i - z_i + u_i < \frac{\tau - 1}{R}, \\ y_i & \text{otherwise;} \end{cases}$$

See Appendix A for the steps to derive the solution. The dual problem (2.14) is a generalized lasso problem that can be solved with the parametric max-flow algorithm from [CD09].

The entire procedure is presented in Algorithm 1. In practice, we can simply choose the penalty parameter R to be $\frac{1}{2}$. We require the procedure to stop if it reaches the maximum iteration or the primal residual $\|\theta^{(k)} - \theta^{(k-1)}\|_2$ is within a tolerance κ, to which we set 10^{-2} in our computation. Actually, the ADMM algorithm converges very quickly with only tens of iterations and hence we find it to be faster than linear programming. Another advantage of ADMM is that the algorithm is not sensitive to the initialization. In Appendix B, we present a simulation study to demonstrate the fast convergence of ADMM under different initializations.

2.3.2 Majorize-Minimize (MM)

We now exploit the majorize-minimize (MM) approach from [HL00] for estimating the conditional median. The main advantage of the MM algorithm versus ADMM is that it is conceptually simple and it is a descent algorithm.

Recall that for $\tau = 0.5$, quantile regression becomes least absolute deviation, or L_1-regression, and the loss function $L(\theta)$ of our problem becomes

$$L(\theta) = \sum_{i=1}^{n} |y_i - \theta_i| + \lambda \|\nabla_G \theta\|_1. \tag{2.15}$$

Next, notice that $L(\theta)$ can be majorized at $\theta^{(k)}$ by $Q(\theta \mid \theta^{(k)})$ given as

$$Q(\theta \mid \theta^{(k)}) = \sum_{i=1}^{n} \frac{(y_i - \theta_i)^2}{|y_i - \theta_i^{(k)}|} + \lambda \sum_{(i,j) \in E_K} \frac{(\theta_i - \theta_j)^2}{|\theta_i^{(k)} - \theta_j^{(k)}|} + \text{const.}, \tag{2.16}$$

Algorithm 1: Alternating Directions Method of Multipliers for quantile K-NN fused lasso

Input: Number of nearest neighbor: K, quantile: τ, penalty parameter: λ, maximum iteration: N_{iter}, tolerance: κ

Data: $X \in \mathbb{R}^{n \times d}$, $y \in \mathbb{R}^n$

Output: $\hat{\theta} \in \mathbb{R}^n$

1. Compute K-NN graph incidence matrix ∇_G from X.

2. Initialize $\theta^{(0)} = y, z^{(0)} = y, u^{(0)} = 0$.

3. For $k = 1, 2, ...,$ until $\|\theta^{(k)} - \theta^{(k-1)}\|_2 \leq \kappa$ or the procedure reaches N_{iter}:

(a) For $i = 1, ..., n$, update

$$\theta_i^{(k)} \leftarrow \arg\min \left\{ \rho_\tau(y_i - \theta_i) + \frac{R}{2}(\theta_i - z_i^{(k-1)} + u_i^{(k-1)})^2 \right\}.$$

(b) Update

$$z^{(k)} \leftarrow \arg\min \left\{ \frac{1}{2}\|\theta^{(k)} + u^{(k-1)} - z\|^2 + \frac{\lambda}{R}\|\nabla_G z\|_1 \right\}.$$

(c) Update

$$u^{(k)} \leftarrow u^{(k-1)} + \theta^{(k)} - z^{(k)}.$$

since it holds that

$$\begin{aligned} Q(\theta^{(k)} \mid \theta^{(k)}) &= L(\theta^{(k)}), \\ Q(\theta \mid \theta^{(k)}) &\geq L(\theta) \text{ for all } \theta. \end{aligned} \tag{2.17}$$

To avoid possible occurrences of zero, we add a perturbation ϵ to the denominator each time. Then, the iterative algorithm optimizes $Q(\theta \mid \theta^{(k)})$ at each iteration. The stopping criterion for MM algorithm remains the same as for ADMM. Because the optimization problem here has a closed-form solution, we can compute the solution directly by solving a linear system (see Step 3c in Algorithm 2). We find the MM algorithm to be faster in running time than linear programming and ADMM for large-size problems and it produces reasonably stable solutions as the others; see the experiments and discussion in Section 2.5.1. A major

drawback of our fast algorithm is that it can only handle median regression at this moment. We leave for future work studying an extension of an MM-based algorithm for estimating general quantiles in the future.

Algorithm 2: Majorize-Minimize for quantile K-NN fused lasso, $\tau = 0.5$

Input: Number of nearest neighbor: K, penalty parameter: λ, maximum iteration: N_{iter}, tolerance: κ

Data: $X \in \mathbb{R}^{n \times d}$, $y \in \mathbb{R}^n$

Output: $\hat{\theta} \in \mathbb{R}^n$

1. Compute K-NN graph incidence matrix ∇_G from X.

2. Initialize $\theta_i^{(0)} = \text{median}(y)$ for $i = 1, ..., n$.

3. For $k = 1, 2, ...,$ until $\|\theta^{(k)} - \theta^{(k-1)}\|_2 \leq \kappa$ or the procedure reaches N_{iter}:

(a) Compute weight matrix $W \in \mathbb{R}^{n \times n}$:

$$W = \text{diag}(1/[|y - \theta^{(k-1)}| + \epsilon]).$$

(b) Compute weight matrix

$$\tilde{W} = \text{diag}(1/[|\theta_i^{(k-1)} - \theta_j^{(k-1)}| + \epsilon]) \text{ if } (i, j) \in E_K.$$

(c) Update

$$\theta^{(k)} \leftarrow [W + \lambda \nabla_G^\top \tilde{W} \nabla_G]^{-1} W y.$$

2.3.3 Model Selection

The choice of the tuning parameter λ in (2.8) is an important practical issue in estimation because it controls the degree of smoothness in the estimator. The value of λ can be chosen through K-fold cross-validation. Alternatively, we can select the regularization parameter based on Bayesian Information Criteria [BIC; Sch78]. The BIC for quantile regression [YM01]

can be computed as

$$\text{BIC}(\tau) = \frac{2}{\sigma}\sum_{i=1}^{n}\rho_\tau(y_i - \hat{\theta}_i) + \nu \log n,$$

where ν denotes the degree of freedom of the estimator and $\sigma > 0$ can be empirically chosen as $\sigma = \frac{1-|1-2\tau|}{2}$. It is also possible to use Schwarz Information Criteria [SIC; KNP94] given by

$$\text{SIC}(\tau) = \log\left[\frac{1}{n}\sum_{i=1}^{n}\rho_\tau(y_i - \hat{\theta}_i)\right] + \frac{1}{2n}\nu \log n.$$

However, BIC is more stable than SIC in practice because when λ is small, SIC may become ill-conditioned as the term inside the logarithm is close to 0.

[TT12] demonstrate that a lasso-type problem has degrees of freedom ν equal to the expected nullity of the penalty matrix after removing the rows that are indexed by the boundary set of a dual solution at y. In our case, we define ν as the number of connected components in the graph ∇_G after removing all edges j's where $|(\nabla_G\hat{\theta})_j|$ is above a threshold γ, typically very small (e.g., 10^{-2}).

2.4 Theoretical Analysis

Before arriving at our main results, we introduce some notation. For a set $A \subset \mathcal{A}$ with $(\mathcal{A}, d_{\mathcal{A}})$ a metric space, we write $B_\epsilon(A) = \{a : \text{exists } a' \in A, \text{ with } d_{\mathcal{A}} \leq \epsilon\}$. The Euclidean norm of a vector $x \in \mathbb{R}$ is denoted by $\|x\|_2 = (x_1^2 + ... + x_d^2)^{1/2}$. The l_1 norm of x is denoted by $\|x\|_1 = |x_1| + ... + |x_d|$. The infinity norm of x is denoted by $\|x\|_\infty = \max_i |x_i|$. In the covariate space $\mathcal{X}$, we consider the Borel sigma algebra, $\mathcal{B}(\mathcal{X})$, induced by the metric $d_{\mathcal{X}}$, and we let μ be a measure on $\mathcal{B}(\mathcal{X})$. We assume that the covariates in the model (2.1) satisfy $x_i \overset{\text{ind}}{\sim} p(x)$. In other word, p is the probability density function associated with the distribution of x_i, with respect to the measure space $(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mu)$. Let $\{a_n\}$ and $\{b_n\} \subset \mathbb{R}$ be two sequences. We write $a_n = O_{\mathbb{P}}(b_n)$ if for every $\epsilon > 0$ there exists $C > 0$ such that $\mathbb{P}(a_n \geq Cb_n) < \epsilon$ for all n. We also write poly(x) as a polynomial function of x, but the functions vary from case to case in the content below. Throughout, we assume the dimension of $\mathcal{X}$ to be greater than 1, since the study of the one-dimensional model, quantile trend filtering, can be found in

[PC22].

We first make some necessary assumptions for analyzing the theoretical guarantees of both the constrained and penalized estimators.

Assumption 1 *(Bounded Variation) We write* $\theta_i^* = F_{y_i|x_i}^{-1}(\tau)$ *for* $i = 1, ..., n$ *and require that* $V^* := \|\nabla_G \theta^*\|_1/[n^{1-1/d}\text{poly}(\log n)]$ *satisfies* $V^* = O_{\mathbb{P}}(1)$*. Here* $F_{y_i|x_i}$ *is the cumulative distribution function of* y_i *given* x_i *for* $i = 1, ..., n$*.*

The first assumption simply requires that θ^*, the vector of τ-quantiles of y, has bounded total variation along the K-NN graph. The scaling $n^{1-1/d}\text{poly}(\log n)$ comes from [PSC20], and we will discuss the details after we present Assumptions 3–5 below.

Assumption 2 *There exists a constant* $L > 0$ *such that for* $\delta \in \mathbb{R}^n$ *satisfying* $\|\delta\|_\infty \leq L$ *we have that*

$$\min_{i=1,...,n} f_{y_i|x_i}(\theta_i^* + \delta_i) \geq \underline{f} \quad a.s.,$$

for some $\underline{f} > 0$*, and where* $f_{y_i|x_i}$ *is the the conditional probability density of* y_i *given* x_i*.*

The assumption that the conditional density of the response variable is bounded by below in a neighborhood is standard in quantile regression analysis. Related conditions appears as D.1 in [BC11] and Condition 2 in [HS94].

The next three assumptions inherit from the study of K-NN graph in [LRH14] and [PSC20].

Assumption 3 *The density of the covariates,* p*, satisfies* $0 < p_{\min} < p(x) < p_{\max}$*, for all* $x \in \mathcal{X}$*, where* $p_{\min}, p_{\max} > 0$ *are fixed constants.*

We only require the distribution of covariates to be bounded above and below by positive constants. In [GKK06] and [Mei06], p is assumed to be the probability density function of the uniform distribution in $[0, 1]^d$.

Assumption 4 *The base measure μ in the metric space $\mathcal{X}$, in which X is defined, satisfies*

$$c_{1,d} r^d \leq \mu\{B_r(x)\} \leq c_{2,d} r^d, \text{ for all } x \in \mathcal{X},$$

for all $0 < r < r_0$, where $r_0, c_{1,d}, c_{2,d}$ are positive constants, and $d \in \mathbb{N}\backslash\{0,1\}$ is the intrinsic dimension of $\mathcal{X}$.

Although $\mathcal{X}$ is not necessarily a Euclidean space, we require in this condition that balls in $\mathcal{X}$ have volume, with respect to some measure μ on the Borel sigma algebra, $\mathcal{B}(\mathcal{X})$, that behaves similarly to the Lebesgue measure of balls in $\mathbb{R}^d$.

Assumption 5 *There exists a homeomorphism $h : \mathcal{X} \to [0,1]^d$, such that*

$$L_{\min} d_{\mathcal{X}}(x, x') \leq \|h(x) - h(x')\|_2 \leq L_{\max} d_{\mathcal{X}}(x, x'),$$

for all $x, x' \in \mathcal{X}$ and for some positive constants $L_{\min}, L_{\max}$, where $d \in \mathbb{N}\backslash\{0,1\}$ is the intrinsic dimension of $\mathcal{X}$.

The existence of a continuous bijection between $\mathcal{X}$ and $[0,1]^d$ ensures that the space has no holes and is topologically equivalent to $[0,1]^d$. Furthermore, Assumption 5 requires $d > 1$. If $d = 1$ then one can simply order the covariates and then run the one dimensional quantile fused lasso studied in [PC22].

On another note, we point that [PSC20] showed, exploiting ideas from [LRH14], that under Assumptions 3–5, $\|\nabla_G \theta^*\|_1 \asymp n^{1-1/d}$ for a K-NN graph G up to a polynomial of $\log n$, with an extra condition on the function $f_0 \circ h^{-1}$ being piecewise Lipschitz. Hence, under such conditions Assumption 1 holds. For completeness, the definition of the class of piecewise Lipschitz functions is provided in Appendix E. It is rather remarkable that [PSC20] also presents an alternative condition on $f_0 \circ h^{-1}$ than piecewise Lipschitz to guarantee the results hold; see Assumption 5 in the same paper.

Now, we are ready to present our first theoretical result on quantile K-NN fused lasso estimates.

Theorem 2.4.1 *Under Assumptions 1–5, by setting* $C = \frac{V}{n^{1-1/d}}$ *in (2.10) for a tuning parameter* V*, where* $V \asymp 1$ *and* $V \geq V^*$*, we have*

$$\Delta_n^2\left(\theta^* - \hat{\theta}_C\right) = O_{\mathbb{P}}\left\{n^{-1/d}\text{poly}(\log n)\right\},$$

for a choice of K *satisfying* $K = \text{poly}(\log n)$*.*

The first theorem shows that quantile K-NN fused lasso attains the optimal rate of $n^{-1/d}$ under the loss $\Delta_n^2(\cdot)$ defined in Section 2.1 for estimating signals in a constrained set.

Theorem 2.4.2 *Under Assumptions 1–5, there exists a choice of* λ *for (2.8) satisfying*

$$\lambda = \begin{cases} \Theta\left\{\log n\right\} & \text{for } d = 2, \\ \Theta\left\{(\log n)^{1/2}\right\} & \text{for } d > 2, \end{cases}$$

such that

$$\Delta_n^2\left(\theta^* - \hat{\theta}\right) = O_{\mathbb{P}}\left\{n^{-1/d}\text{poly}(\log n)\right\},$$

for a choice of K *satisfying* $K = \text{poly}(\log n)$*.*

The second theorem states that, under certain choice of the tuning parameter, the penalized estimator achieves the convergence rate of $n^{-1/d}$, similar to the constrained estimator, ignoring the logarithmic factor. Notice that there is a one-to-one correspondence between (2.8) and (2.10), as they are equivalent optimization problems. Let $\lambda > 0$ as in the statement of Theorem 2.4.2. Then there exists a C that depends on y and λ such that (2.8) and (2.10) have identical solutions. However, the fact that such C depends on y does not imply that for a deterministic C, such as that in Theorem 2.4.1, one can conclude that if the unconstrained estimator attains the rate $n^{-1/d}$ then the constrained estimator attains the same rate. Nevertheless, the fact that we have Theorems 2.4.1–2.4.2 implies that both estimators attain the same rate.

We emphasize that if all or some of Assumptions 1, 3–5 do not hold, then we are not able to characterize the behavior of the K-NN graph. In such case the conclusions of Theorems

2.4.1–2.4.2 would not necessarily hold. The following remark regards the minimax optimality of Theorems 2.4.1–2.4.2.

Remark 1 *Let $f^*(t) = F^{-1}_{Y|X=t}(0.5)$ be the median function and let $\mathcal{C}$ be the class of piecewise Lipschitz functions, see Definition 2.7.12 in Appendix E. It was proven in Proposition 2 of [WNC06] that under the assumption that $\epsilon_i \overset{ind}{\sim} N(0, \sigma^2)$ for some fixed σ, and that $x_i \overset{ind}{\sim} U([0,1]^d)$ for $i = 1, \ldots, n$, it holds that*

$$\inf_{\hat{f} \text{ estimator}} \sup_{f^* \in \mathcal{C},\ \|f^*\|_\infty \leq 1} \mathbb{E}\left[\int_{[0,1]^d} (f^*(t) - \hat{f}(t))^2 dt\right] \geq cn^{-1/d}, \tag{2.18}$$

for some constant $c > 0$. However, by the constraint $\|f^*\|_\infty \leq 1$, the left hand side of (2.18) equals, up to a constant, to

$$\inf_{\hat{f} \text{ estimator}} \sup_{f^* \in \mathcal{C},\ \|f^*\|_\infty \leq 1} \mathbb{E}\left[\int_{[0,1]^d} \min\{|f^*(t) - \hat{f}(t)|, (f^*(t) - \hat{f}(t))^2\} dt\right]. \tag{2.19}$$

Furthermore, a discrete version of

$$\mathbb{E}\left[\int_{[0,1]^d} \min\{|f^*(t) - \hat{f}(t)|, (f^*(t) - \hat{f}(t))^2\} dt\right]$$

is the quantity $\Delta_n^2(\hat{\theta} - \theta^*)$ if $\hat{\theta}_i = \hat{f}(x_i)$ and $\theta_i^* = f^*(x_i)$ for $i = 1, \ldots, n$. Therefore, the rates in Theorems 2.4.1–2.4.2 are nearly minimax in the sense that they match, up to log factors, the lower bound $n^{-1/d}$ on the quantity (2.19) without requiring sub-Gaussian errors, and under more general conditions on the covariates than uniform draws. In contrast, under sub-Gaussian errors, the K-NN fused lasso estimator from [PSC20] attains, up to log factors, the rate $n^{-1/d}$ in terms of mean squared error.

The second remark is related to the manifold adaptivity of our proposed estimator.

Remark 2 *Our proposed method would expect the property of manifold adaptivity, from its connection with the K-nearest-neighbor approach. This allows our method to adapt to the intrinsic dimensionality of the data. While we have not conducted a detailed analysis of this property in our work, it has been rigorously studied in the existing literature on nearest-neighbor methods. For instance, [Kpo11] provides a thorough theoretical proof that K-NN*

regression adapts to local intrinsic dimension, while [PSC20] explores this property for K-NN fused lasso under a sub-Gaussian error assumption.

In our approach, we choose the number of nearest neighbors, K, following the conventional order at a slightly faster rate than $\log n$. The concept of manifold adaptivity highlights the potential for future research to develop a more data-driven strategy for selecting K, tailored to the density of the design points. Additionally, another promising research direction is to extend the theoretical analysis in [PSC20] to more general settings, including applications within a quantile regression framework, as considered in this work.

2.5 Experiments

In this section, we will examine the performance of quantile K-NN fused lasso (QKNN) on various simulated and real datasets. The two benchmark estimators we compare against are K-NN fused lasso [KNN; PSC20] and quantile random forest [QRF; Mei06]. The performance of an estimator is measured by its mean squared error, defined by

$$\mathrm{MSE}(\hat{\theta}) := \frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i^*)^2,$$

where θ^* is the vector of τ-quantiles of the true signal.

For quantile K-NN fused lasso, we use the ADMM algorithm and select the tuning parameter λ based on the BIC criteria described in Section 2.3.3; for K-NN fused lasso, we use the algorithm from [CD09] and the corresponding penalty parameter is chosen to minimize the average mean squared error over 500 Monte Carlo replicates. For quantile random forest, we directly use the R package "quantregForest" with defaulted choice of tree structure and tuning parameters.

Throughout, for both K-NN fused lasso and quantile K-NN fused lasso, we set K to be 5 for sufficient information and efficient computation.

2.5.1 Simulation Study

We generate 500 data sets from models under each scenario described below with sample size between 10^2 and 10^4 and then report the mean squared errors of the three estimators with respect to different quantiles. For each scenario the data are generated as

$$y_i = \theta_i^* + \epsilon_i, \text{ and } \theta_i^* = f_0(x_i), \ i = 1, ..., n,$$

where θ_i^* comes from some underlying functions f_0, and the errors $\{\epsilon_i\}_{i=1}^n$ are independent with $\epsilon_i \sim F_i$ for some distributions F_i, where we select from Gaussian, Cauchy, and t-distributions.

Scenario 1

We generate x_i uniformly from $[0,1]^2$, and define $f_0 : [0,1]^2 \to \{0,1\}$ by

$$f_0(x) = \begin{cases} 1 & \text{if } \ \frac{5}{4}x_{i1} + \frac{3}{4}x_{i2} > 1, \\ 0 & \text{otherwise.} \end{cases}$$

Scenario 2

In this case, we generate $X \in \mathbb{R}^2$ according to the probability density function

$$p(x) = \frac{1}{5}\mathbf{1}_{\{[0,1]^2 \setminus [0.4,0.6]^2\}}(x) + \frac{16}{25}\mathbf{1}_{\{[0.45,0.55]^2\}}(x) + \frac{4}{25}\mathbf{1}_{\{[0.4,0.6]^2 \setminus [0.45,0.55]^2\}}(x).$$

The function $f_0 : \ [0,1]^2 \to \mathbb{R}$ is defined as

$$f_0(x) = \mathbf{1}_{\{\|x - \frac{1}{2}(1,1)^\top\|_2^2 \leq \frac{2}{1000}\}}(x).$$

Scenario 3

Again, x_i are from uniform $[0,1]^2$. The smooth function $f_0 : \ [0,1]^2 \to \mathbb{R}$ is defined as

$$f_0(x_i) = 0.4x_{i1}^2 + 0.6x_{i2}^2.$$

Scenario 4

The function $f_0 : \ [0,1]^d \to \mathbb{R}$ is defined as

$$f_0(x) = \begin{cases} 1 & \text{if } \ \|x - \frac{1}{4}1_d\|_2 < \|x - \frac{3}{4}1_d\|_2, \\ -1 & \text{otherwise,} \end{cases}$$

and the density p is uniform in $[0,1]^d$. The errors are chosen as $(x^\top\beta)\epsilon$, where $\beta = (\frac{1}{d}, ..., \frac{1}{d})^\top$. Here we simulate with $d = 5$.

The scenarios above have been chosen to illustrate the local adaptivity of our proposed approach to discontinuities of the quantile function. Scenario 1 consists of a piecewise constant median function. Scenario 2 is borrowed from [PSC20] and also has a piecewise constant median. However, the covariates in Scenario 2 are not uniformly drawn and are actually highly concentrated in a small region of the domain. Scenario 3 is a smooth function, and Scenario 4 is taken from [PSC20]. In the latter we also have a piecewise constant quantiles, but the boundaries of the different pieces are not axis aligned, and the errors are heteroscedastic.

Figure 2.1 displays the true function and the estimates from both quantile K-NN fused lasso and quantile random forest under Scenarios 1, 2 and 3. Clearly, quantile K-NN fused lasso provides more reasonable estimation in all cases. Quantile random forest estimates are more noisy and the performance is even poorer under Cauchy errors.

The results presented in Table 2.1 indicate that overall, quantile K-NN fused lasso outperforms the competitors in most scenario. As expected, for estimating functions with Gaussian errors like Scenario 1, regular K-NN fused lasso is the best method. For Scenarios 2 and 3, when estimating the conditional median of piecewise continuous or continuous functions with heavy-tail errors (such as Cauchy and t-distributions), quantile K-NN fused lasso achieve the smallest mean square errors over the other two methods.

We also compare the performance of linear programming the two algorithms discussed in Section 2.3, with simulated data from Scenario 3. We obtain almost identical estimators from the three algorithms under the same choice of λ. Regarding computational time, we record the averaged time consumed over 100 simulations for each algorithm. Figure 2.2 demonstrates that majorize-minimize (MM) is the most efficient one among the three algorithms, and linear programming (LP) can be very expensive in operational time for large-size problems.

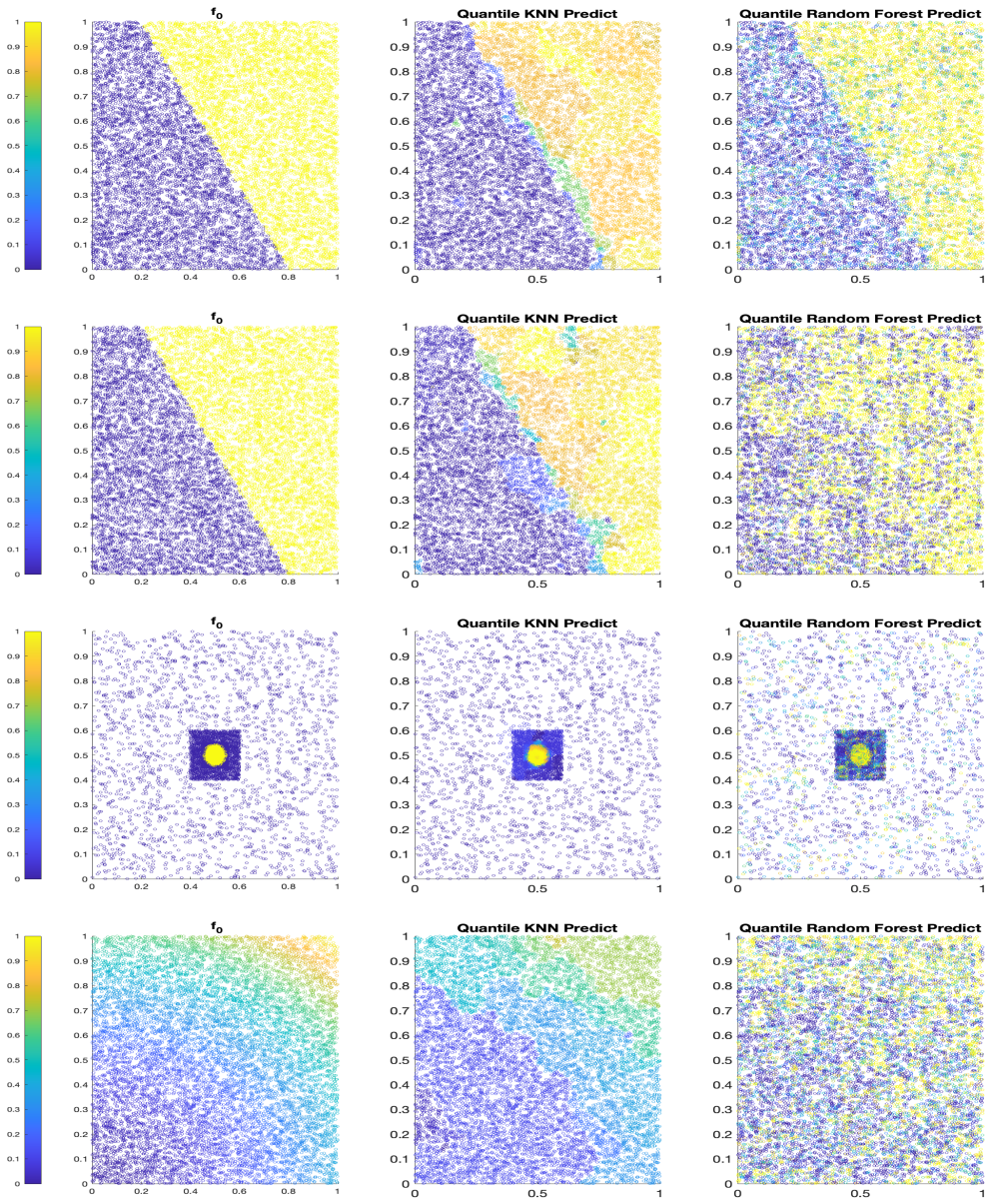

Figure 2.1: Comparison among observations and estimates from Scenario 1 with Gaussian errors (*the first row*), Scenario 1 with Cauchy errors (*the second row*), Scenario 2 (*the third row*), and Scenario 3 (*the fourth row*). *Left column*: the function f_0 evaluated at the observed x_i for $i = 1, \ldots, n$, with $n = 10000$. The horizontal and vertical axis of each panel correspond to the coordinates of x_i. *Middle column*: the corresponding estimate of f_0 obtained via quantile K-NN fused lasso. *Right column*: the estimate of f_0 obtained via quantile random forest.

n	Scenario	ϵ	τ	QKNN	QRF	KNN
100	1	N(0,1)	0.5	0.2364 (0.0628)	0.2242 (0.0467)	**0.1507** (0.0365)
1000	1	N(0,1)	0.5	0.1225 (0.0117)	0.1569 (0.0116)	**0.0872** (0.0239)
5000	1	N(0,1)	0.5	0.0983 (0.0024)	0.1348 (0.0047)	**0.0540** (0.0022)
10000	1	N(0,1)	0.5	0.0370 (0.0019)	0.1279 (0.0028)	**0.0292** (0.0010)
100	1	Cauchy(0,1)	0.5	**0.1776** (0.0746)	3059.63 (54248)	35.4988 (124.19)
1000	1	Cauchy(0,1)	0.5	**0.1440** (0.0817)	26640.10 (179298)	2816.1 (7691.6)
5000	1	Cauchy(0,1)	0.5	**0.1326** (0.0740)	34038.86 (196566)	4628.8 (7989.0)
10000	1	Cauchy(0,1)	0.5	**0.0962** (0.0642)	57748.95 (200327)	4799.0 (7249.6)
100	2	t_3	0.5	**0.1838** (0.0467)	0.5591 (0.3569)	0.2374 (0.0595)
1000	2	t_3	0.5	**0.0780** (0.0135)	0.3935 (0.0875)	0.1428 (0.0536)
5000	2	t_3	0.5	**0.0364** (0.0050)	0.3479 (0.0477)	0.0622 (0.0165)
10000	2	t_3	0.5	**0.0265** (0.0024)	0.3311 (0.0457)	0.0542 (0.0097)
100	3	t_2	0.5	**0.0470** (0.0253)	1.6723 (4.8880)	0.1545 (0.4188)
1000	3	t_2	0.5	**0.0174** (0.0050)	1.5363 (2.9381)	0.0510 (0.0230)
5000	3	t_2	0.5	**0.0075** (0.0015)	1.4207 (2.1827)	0.0414 (0.0106)
10000	3	t_2	0.5	**0.0059** (0.0019)	1.2910 (1.1924)	0.0409 (0.0102)
100	4	t_3	0.9	0.7413 (0.2180)	**0.6948** (0.5340)	*
1000	4	t_3	0.9	**0.3568** (0.1192)	0.5374 (0.2656)	*
5000	4	t_3	0.9	**0.2889** (0.0683)	0.4420 (0.0422)	*
10000	4	t_3	0.9	**0.2409** (0.0173)	0.4344 (0.0548)	*
100	4	t_3	0.1	1.1563 (0.3923)	**0.8877** (0.6858)	*
1000	4	t_3	0.1	**0.4160** (0.1542)	0.6224 (0.2667)	*
5000	4	t_3	0.1	**0.3074** (0.0503)	0.4897 (0.0644)	*
10000	4	t_3	0.1	**0.2597** (0.0301)	0.4536 (0.0494)	*

Table 2.1: Mean squared error $\frac{1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i^*)^2$, averaging over 500 Monte Carlo simulations for the different methods, sample sizes, errors, quantiles considered. The numbers in parentheses indicate the standard Monte Carlo errors over the replications.

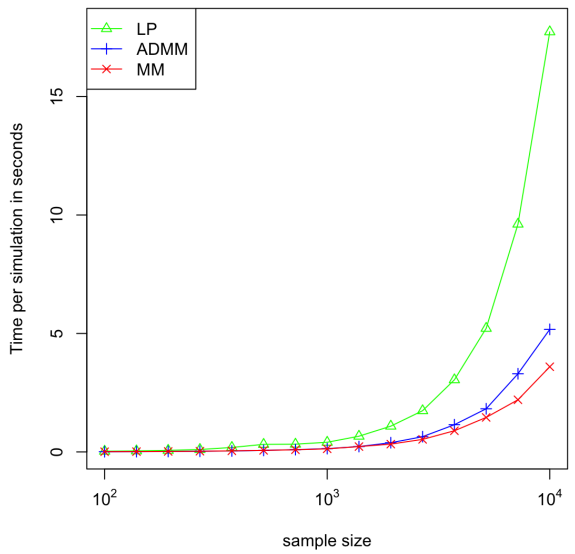

Figure 2.2: A log-scaled plot of time per simulation of LP, ADMM and MM algorithm against problem size n (15 values from 10^2 up to 10^4). For each algorithm, the time to compute the estimate for one simulated data is averaged over 100 Monte Carlo simulations.

2.5.2 Real Data

2.5.2.1 California Housing Data

In this section, we conduct an experiment of predicting house value in California based on median income and average occupancy, similar to the experiment in [PSW16]. The data set, consisting of 20,640 measurements, was originally used in [PB97], and is publicly available from the Carnegie Mellon StatLib data repository (`http://lib.stat.cmu.edu`).

We perform 100 train-test random splits the data, with training sizes 1000, 5000, and 10000. For each split, the data not in the training set is treated as testing data. For median estimation, we compare averaged mean squared prediction errors of the test sets (after taking the log of housing price) from quantile K-NN fused lasso and quantile random forest; besides, we construct 90% and 95% prediction intervals from both methods and report the proportion of true observations in the test set that fall in the intervals. Both evaluations are averaged over 100 repetitions for each method. The tuning parameter λ for quantile K-NN fused lasso

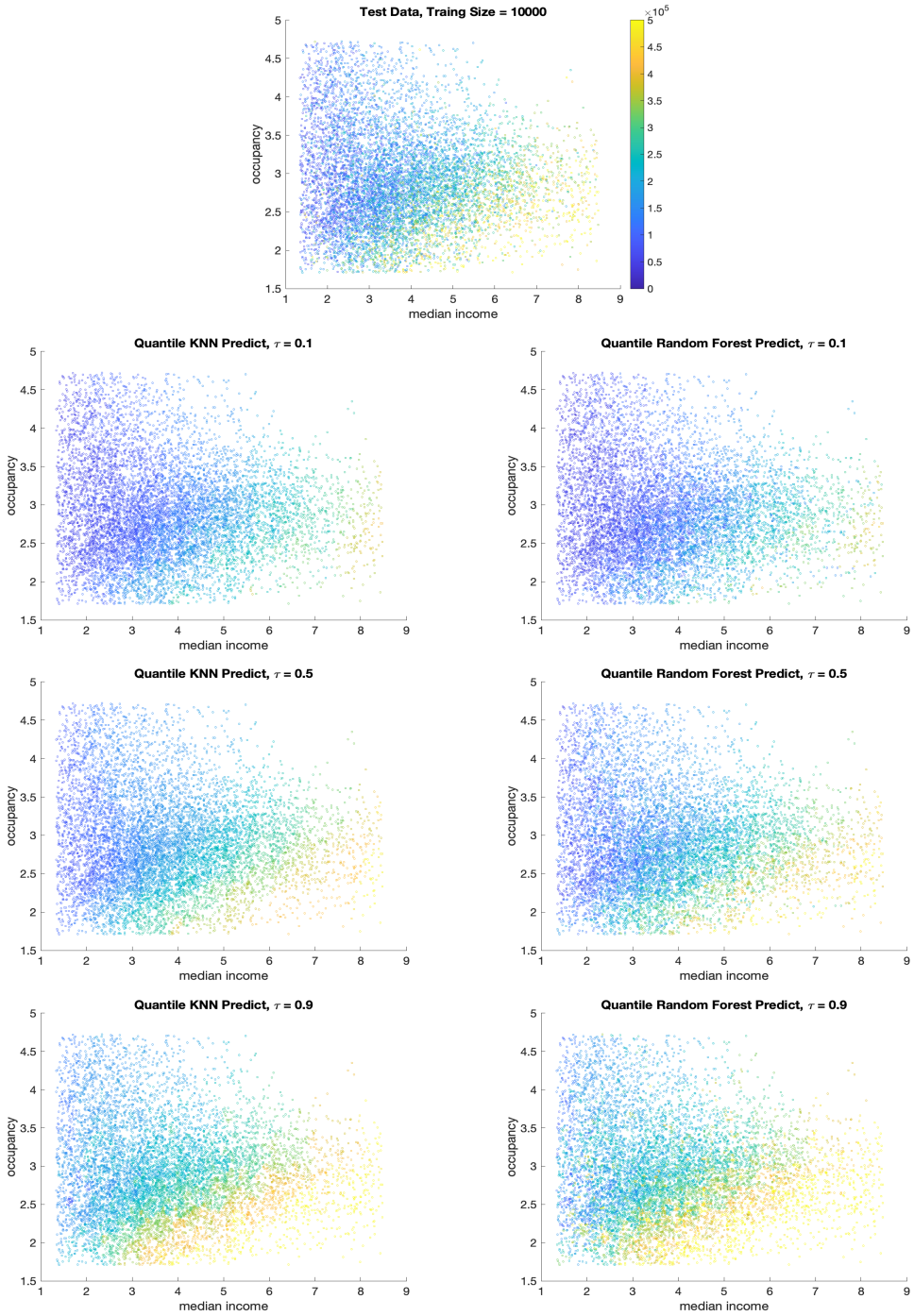

Figure 2.3: Comparison between predictions from two methods for California housing data, with respect to $\tau = 0.1, 0.5, 0.9$. The plot at the top presents the true testing data value, and the color scale is the same among all plots.

is chosen based on the BIC criteria for each training and the parameters for quantile random forest are selected as default.

From the results in Table 2.2, quantile K-NN fused lasso has better performance than quantile random forest in all cases. The result agrees with the nature of piecewise continuity in housing price, that guarantee the advantage of our proposed method over the competitor. When we illustrate the predicted values on a test set from a experiment with training size of 10,000 visually in Figure 2.3, we also observe a piecewise continuous structure in quantile K-NN fused lasso estimates, while estimates from quantile random forest contain more noise, especially for lower and higher quantiles.

Training Size	τ	QKNN	QRF
1000	0.5	**0.1553** (0.0032)	0.1679 (0.0026)
5000	0.5	**0.1485** (0.0016)	0.1646 (0.0016)
10000	0.5	**0.1479** (0.0017)	0.1623 (0.0019)

Training Size	τ	QKNN	QRF
1000	0.9	**0.8405** (0.0141)	0.8315 (0.0086)
5000	0.9	**0.8569** (0.0076)	0.8334 (0.0040)
10000	0.9	**0.8585** (0.0061)	0.8337 (0.0041)
1000	0.95	**0.9163** (0.0127)	0.8887 (0.0084)
5000	0.95	**0.9216** (0.0099)	0.8891 (0.0040)
10000	0.95	**0.9274** (0.0097)	0.8902 (0.0042)

Table 2.2: Average test set prediction error (across the 100 test sets) on California housing data. For median, we report the mean squared errors; for other quantiles ($\tau \in \{0.9, 0.95\}$), we report the averaged proportion of the true data located in the predicted confidence interval. Standard Monte Carlo errors are recorded in parentheses. The number of nearest neighbors, K is chosen as 5 for quantile K-NN fused lasso.

Finally, it is clear from Table 2.2 that both quantile K-NN fused lasso and quantile random forest have coverage probabilities are noticeably lower than the true nominal level. Since the results in Table 2.2 correspond to a real data set, training on a subset and predicting on a different subset of the data, we are not aware of how to improve the coverage for both of the competing methods.

2.5.2.2 Chicago Crime Data

We apply quantile K-NN fused lasso and quantile random forest to a dataset of publicly-available crime report counts in Chicago, Illinois in 2015. We preprocess the dataset in the same way as [TTS18] by merging all observations into a fine-grained 100×100 grid based on latitude and longitude, taking the log of the total counts in each cell, and omitting all cells with zero count. Data points with zero observed crimes were excluded from the dataset, as it was unclear whether they reflected an actual absence of crime or a location outside the jurisdiction of the local police department. The resulting preprocessed data contains a total number of 3756 data points along the grid. Similar to the experiment on California housing data, we perform a train-test split with training size 500, 1000, 1500, and 2000, and model on the median counts according to the position on the X and Y axes of the grid. We then predict the value on test set and report the averaged square errors over 100 test sets. The parameters for both methods are selected in the same way as in the previous experiment.

From the results of Table 2.3, we see that quantile K-NN fused lasso still outperforms quantile random forest in MSE if the training size is at least 1500. We notice that for small sample size, our method suffers. This is presumably due to the fact that the raw data is not smooth enough. Yet, Figure 2.4 shows that quantile K-NN fused lasso capture local patterns more successfully than quantile random forest in most regions.

Training Size	τ	QKNN	QRF
500	0.5	1.2161 (0.0441)	**1.1441** (0.0479)
1000	0.5	1.0151 (0.0374)	**0.9836** (0.0417)
1500	0.5	**0.9138** (0.0343)	0.8959 (0.0380)
2000	0.5	**0.8476** (0.0409)	0.8383 (0.0417)

Table 2.3: Average test set prediction errors (across the 100 test sets) with standard errors on Chicago Crime data.

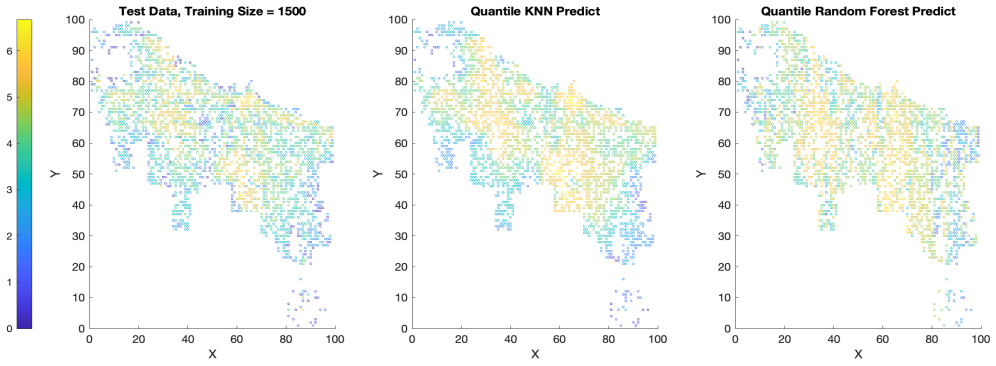

Figure 2.4: *Left*: one test data from Chicago Crime Data, under training size 1500. *Middle*: the estimate obtained via quantile K-NN fused lasso. *Right*: the estimate obtained via quantile random forest.

2.6 Conclusion

In this chapter we have proposed and studied a method for estimating quantile functions that are piecewise constant. Such classes of functions naturally arise in real-life data. Estimating signals accurately with such structure is an important but challenging research topic that arises in many practical contexts. However, there is a lack of existing literature on robust estimation of piecewise constant or piecewise Lipschitz signals in multivariate settings. Motivated by this, we have proposed the quantile K-NN fused lasso estimator.

Our numerical experiments on different real-life and simulated examples lay out empirical

evidence of the superior robustness and local adaptivity of our estimator over other state-of-the-art methods. We have presented two algorithms to efficiently compute the solutions to the corresponding optimization problem behind our method. Our theoretical analysis extends previous work on K-NN fused lasso [PSC20] and quantile trend filtering [PC22]. Specifically, we have shown that the quantile K-NN fused lasso estimator achieves an optimal convergence rate of $n^{-1/d}$ for estimating a d-dimensional piecewise Lipschitz quantile function. This result is guaranteed under mild assumptions on data, and thus our estimator can be applied to more general models rather than only those with sub-Gaussian errors.

2.7 Appendix

2.7.1 Closed-form Solution to the Primal in ADMM Algorithm

In Section 2.3.1, we introduce an ADMM algorithm to compute quantile K-NN estimates. The algorithm requires to solve the primal problem (2.13)

$$\theta \leftarrow \underset{\theta \in \mathbb{R}^n}{\arg\min} \left\{ \sum_{i=1}^{n} \rho_\tau(y_i - \theta_i) + \frac{R}{2} \|\theta - z + u\|^2 \right\}.$$

We can solve the problem coordinate-wisely: for $i = 1, ..., n$, we find the minimizer

$$\theta_i \leftarrow \underset{\theta_i \in \mathbb{R}}{\arg\min} \left\{ \rho_\tau(y_i - \theta_i) + \frac{R}{2}(\theta_i - z_i + u_i)^2 \right\}.$$

By definition,

$$\rho_\tau(y_i - \theta_i) = \begin{cases} \tau(y_i - \theta_i) & \text{if } \ y_i - \theta_i > 0, \\ (\tau - 1)(y_i - \theta_i) & \text{if } \ y_i - \theta_i < 0, \\ 0 & \text{if } \ y_i - \theta_i = 0. \end{cases}$$

We discuss the three cases separately.

(1) When $y_i - \theta_i > 0$, $\theta_i \leftarrow \arg\min \left\{ \tau(y_i - \theta_i) + \frac{R}{2}(\theta_i - z_i + u_i)^2 \right\}$. Take the derivative and set to 0 to obtain $\theta_i = z_i - u_i + \tau/R$. The condition $y_i - \theta_i > 0$ then becomes $y_i - z_i + u_i > \tau/R$.

(2) When $y_i - \theta_i < 0$, $\theta_i \leftarrow \arg\min\left\{(\tau-1)(y_i-\theta_i) + \frac{R}{2}(\theta_i - z_i + u_i)^2\right\}$. Take the derivative and set to 0 to obtain $\theta_i = z_i - u_i + (\tau-1)/R$. The condition $y_i - \theta_i < 0$ then becomes $y_i - z_i + u_i < (\tau-1)/R$.

(3) When $y_i - \theta_i = 0$, it is simple to get $\theta_i = y_i$.

To summarize, the closed-form solution to the primal (2.13) is

$$\theta_i = \begin{cases} z_i - u_i + \frac{\tau}{R} & \text{if } y_i - z_i + u_i > \frac{\tau}{R}, \\ z_i - u_i + \frac{\tau-1}{R} & \text{if } y_i - z_i + u_i < \frac{\tau-1}{R}, \\ y_i & \text{otherwise;} \end{cases}$$

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

2.7.2 Sensitivity of Initialization in the ADMM Algorithm

In this section, we examine how the convergence of Algorithm 1 is sensitive to different initials. Recall that Algorithm 1 requires an initialization of $\theta^{(0)}$, $z^{(0)}$ and $u^{(0)}$. A practical way to choose these initials is to set $\theta^{(0)} = z^{(0)} = y, u^{(0)} = 0$, where y is the input data, as defined in Algorithm 1. Here, we try three other initializations to assess the sensitivity to the initials:

(1) The setup in Algorithm 1: $\theta^{(0)} = z^{(0)} = y, u^{(0)} = 0$.

(2) We add perturbations to the first setup: $\theta^{(0)} = y + v_1, z^{(0)} = y + v_2, u^{(0)} = v_3$, where $v_1, v_2, v_3 \overset{\text{ind}}{\sim} N(0, I_n)$.

(3) Random initialization: $\theta^{(0)} = w_1, z^{(0)} = w_2, u^{(0)} = w_3$, where $w_1, w_2, w_3 \overset{\text{ind}}{\sim} N(0, I_n)$.

(4) Random initialization with a large multiplication factor: $\theta^{(0)} = 50s_1, z^{(0)} = 50s_2, u^{(0)} = 50s_3$, where $s_1, s_2, s_3 \overset{\text{ind}}{\sim} N(0, I_n)$.

We use the data generation mechanism in Scenario 2 from the main content, and errors are independently drawn from a t-distribution with 3 degrees of freedom. The regularization parameter λ is set to be fixed as 0.5. We replicate the simulation 500 times with a sample size $n = 10000$. Note that the input data and perturbations are generated independently in every replication. For each initialization setup, we record the averaged value of the objective function $\sum_{i=1}^{n} \rho_\tau(y_i - \theta_i) + \lambda \|\nabla_G \theta\|_1$ after each iteration and the number of iterations to converge when we set $\kappa = 0.01$ in Step 3 of Algorithm 1.

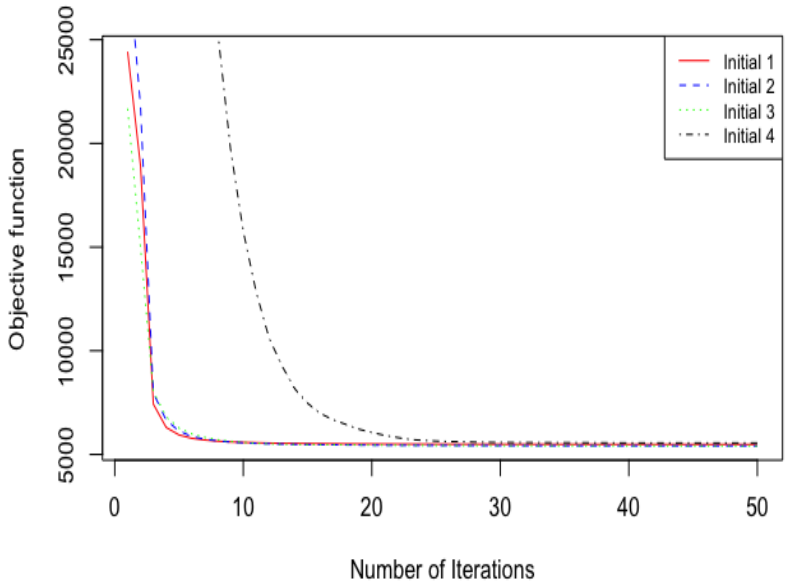

Figure 2.5: Averaged value of objective function (over 500 replications) after each iteration step for different initializations.

From Figure 5, we observe that the value of the objective function converges very fast to the optimal minimum within only tens of iterations, regardless of the random initialization that is used. If we look at the number of iterations shown in Table 4, the first three initializations have relatively the same convergence speed. If there exists a large deviation between the initial value and the true signal like in the last setup, the ADMM algorithm takes slightly longer time to converge. Overall, the ADMM algorithm is insensitive to ran-

dom initialization and fast convergence is observed in practice.

Initialization	Averaged number of iterations to converge
(1)	11
(2)	11.13
(3)	11
(4)	26.65

Table 2.4: Number of iterations to converge under different initializations, averaging over 500 Monte Carlo simulations.

2.7.3 General Lemmas

The following lemmas hold by conditioning on $\{x_i\}_{i=1}^n$.

Definition 2.7.1 *The function Δ^2 is defined as*

$$\Delta^2(\delta) := \sum_{i=1}^{n} \min\{|\delta_i|, \delta_i^2\},$$

where $\delta_i \in \mathbb{R}^n$. We also write $\Delta(\delta) = \{\Delta^2(\delta)\}^{1/2}$.

Definition 2.7.2 *A random variable X is said to be 1-sub-Gaussian if it satisfies the moment condition:*

$$\mathbb{E}[e^{tX}] \leq e^{t^2/2}, \quad \forall t \in \mathbb{R}.$$

Definition 2.7.3 *For a set $S \subset \mathbb{R}^n$, the sub-Gaussian width of S is defined as*

$$SGW(S) = \mathbb{E}\left(\sup_{v \in S} \sum_{i=1}^{n} s_i v_i\right),$$

where $s_1, ..., s_n$ are independent, zero-centered 1-subgaussian random variables.

The notation of sub-Gaussian width is not used very often in literature compared to a similar definition of Gaussian width,

$$GW(S) = \mathbb{E}\left(\sup_{v\in S}\sum_{i=1}^{n} z_i v_i\right),$$

where $z_1, ..., z_n$ are independent standard normal random variables. In fact, the sub-Gaussian width shares many common properties with the Gaussian width. In the following lemmas, we upper bound the sub-Gaussian width by a constant times the Gaussian width using generic chaining.

Lemma 2.7.4 *Let $(Y_v)_{v\in T}$ and $(X_v)_{v\in T}$ be two processes indexed by the set T, defined as*

$$Y_v = v^\top s, \quad X_v = v^\top z,$$

where $s = (s_1, \ldots, s_n)$ consists of independent, zero-centered, 1-sub-Gaussian random variables, and $z = (z_1, \ldots, z_n)$ consists of independent standard normal random variables. Then we have that

$$\mathbb{E}\left(\sup_{u,v\in S}|Y_u - Y_v|\right) \leq L \cdot \mathbb{E}\left(\sup_{v\in S} X_v\right),$$

where $L > 0$ is a universal constant.

Proof. See the proof of Theorem 2.1.5 in [Tal05].

Lemma 2.7.5 *For a set $S \subset \mathbb{R}^n$, following the definitions above, we have that*

$$SGW(S) \leq L \cdot GW(S),$$

where $L > 0$ is the constant from Lemma 2.7.4.

Proof. We define two processes $Y_v = v^\top s$ and $X_v = v^\top z$, where $s = (s_1, \ldots, s_n)$ consists of independent, zero-centered, 1-sub-Gaussian random variables, and $z = (z_1, \ldots, z_n)$ consists of independent standard normal random variables. By symmetry and independence, we have that

$$\sup_{u,v\in S}|Y_u - Y_v| = \sup_{u,v\in S}(Y_u - Y_v) = \sup_{u\in S} Y_u + \sup_{v\in S}(-Y_v).$$

For any fixed $v_0 \in S$, we have that

$$\begin{aligned}
\mathbb{E}\left(\sup_{u,v\in S}|Y_u - Y_v|\right) &= \mathbb{E}\left(\sup_{u\in S} Y_u\right) + \mathbb{E}\left(\sup_{v\in S} -Y_v\right) \\
&\geq \mathbb{E}\left(\sup_{u\in S} Y_u\right) + \mathbb{E}(-Y_{v_0}) \\
&\geq \mathbb{E}\left(\sup_{u\in S} Y_u\right) \\
&= SGW(S).
\end{aligned} \tag{2.20}$$

Thus, by Lemma 2.7.4, we obtain

$$SGW(S) \leq \mathbb{E}\left(\sup_{u,v\in S}|Y_u - Y_v|\right) \leq L \cdot \mathbb{E}\left(\sup_{v\in S} X_v\right) = L \cdot GW(S),$$

for some universal constant $L > 0$. This completes the proof. □

Definition 2.7.6 *We define the empirical loss function*

$$\hat{M}(\theta) = \sum_{i=1}^{n} \hat{M}_i(\theta_i),$$

where

$$\hat{M}_i(\theta_i) = \rho_\tau(y_i - \theta_i) - \rho_\tau(y_i - \theta_i^*).$$

Setting $M_i(\theta_i) = \mathbb{E}(\rho_\tau(y_i - \theta_i) - \rho_\tau(y_i - \theta_i^*))$, the population version of $\hat{M}$ becomes

$$M(\theta) = \sum_{i=1}^{n} M_i(\theta_i).$$

Now, we consider the M-estimator

$$\begin{aligned}
\hat{\theta} \quad = \quad & \underset{\theta\in\mathbb{R}^n}{\arg\min} \qquad \hat{M}(\theta) \\
& \text{subject to} \quad \theta \in S,
\end{aligned} \tag{2.21}$$

and $\theta^* \in \arg\min_{\theta\in\mathbb{R}^n} M(\theta)$. Throughout, we assume that $\theta^* \in S \subset \mathbb{R}^n$.

Lemma 2.7.7 *With the notation from before,*

$$M(\hat{\theta}) \leq \sup_{v\in S}\left\{M(v) - \hat{M}(v)\right\}.$$

Proof. See the proof of Lemma 10 in [PC22].

Lemma 2.7.8 *Suppose that Assumption 2 holds. Then there exists a constant $c_\tau > 0$ such that for all $\delta \in \mathbb{R}^n$, we have*

$$M(\theta + \delta) \geq c_\tau \Delta^2(\delta).$$

Proof. See the proof of Lemma 14 in [PC22].

Corollary 2.7.9 *Under Assumption 2, if $\theta^* \in S$, we have that*

$$\mathbb{E}\left\{\Delta^2(\hat{\theta} - \theta^*)\right\} \leq \mathbb{E}\left[\sup_{v \in S}\left\{M(v) - \hat{M}(v)\right\}\right]. \tag{2.22}$$

Next, we proceed to bound the right hand side of (2.22).

Lemma 2.7.10 *(Symmetrization) Under Assumption 2, we have that*

$$\mathbb{E}\left[\sup_{v \in S}\left\{M(v) - \hat{M}(v)\right\}\right] \leq 2\mathbb{E}\left\{\sup_{v \in S}\sum_{i=1}^{n}\xi_i \hat{M}_i(v_i)\right\},$$

where $\xi_1, ..., \xi_n$ are independent Rademacher variables independent of $y_1, ..., y_n$.

Proof. See the proof of Lemma 11 in [PC22].

Lemma 2.7.11 *(Contraction principle) Under Assumption 2, we have that*

$$\mathbb{E}\left\{\sup_{v \in S}\sum_{i=1}^{n} s_i \hat{M}_i(v_i)\right\} \leq L \cdot GW(S - \theta^*) = L \cdot GW(S),$$

where $s_1, ..., s_n$ are independent, zero-centered 1-subgaussian random variables independent of $y_1, ..., y_n$, and L is a positive constant.

Proof. Recall that $\hat{M}_i(v_i) = \rho_\tau(y_i - v_i) - \rho_\tau(y_i - \theta_i^*)$. Clearly, these are 1-Lipschitz continuous functions. Therefore, if we define independent standard normal random variables $z_1, ..., z_n$, we have that

$$\mathbb{E}\left\{\sup_{v \in S}\sum_{i=1}^{n} s_i \hat{M}_i(v_i)\right\} \leq L \cdot \mathbb{E}\left\{\sup_{v \in S}\sum_{i=1}^{n} z_i \hat{M}_i(v_i)\right\}$$

$$
\begin{aligned}
&= L \cdot \mathbb{E}\left(\mathbb{E}\left\{\sup_{v\in S}\sum_{i=1}^{n} z_i \hat{M}_i(v_i)\Big| y\right\}\right)\\
&\leq L \cdot \mathbb{E}\left(\mathbb{E}\left\{\sup_{v\in S}\sum_{i=1}^{n} z_i v_i\Big| y\right\}\right)\\
&= L \cdot \left[\mathbb{E}\left\{\sup_{v\in S}\sum_{i=1}^{n} z_i (v_i-\theta^*)\right\} + \mathbb{E}\left(\sum_{i=1}^{n} z_i\theta^*\right)\right]\\
&= L \cdot \mathbb{E}\left\{\sup_{v\in S}\sum_{i=1}^{n} z_i (v_i-\theta^*)\right\},
\end{aligned}
$$

where the first inequality follows from Lemma 2.7.5, and the second inequality follows from the Gaussian version of Talagrand's contraction principle, see corallary 3.17 in [LT13] and 7.2.13 in [Ver18]. □

2.7.4 K-NN Embeddings

This section follows from Section D and E in [PSC20] and it originally appeared in the flow-based proof of Theoremr 4 in [LRH14]. The main idea is to embed a mesh on a K-NN graph corresponding to the given observations $X = (x_1, ..., x_n)$ under Assumptions 3-5 so that we are able to bound the total variation along the grid graph from below. In this way, we can further derive an upper bound on the loss of the optimization problem (2.8) and (2.10) with respect to the loss function defined in Definition 2.7.1.

First, we need to construct a valid grid discussed in Definition 17 of [LRH14] with a minor modification. With high probability, a valid grid graph G satisfies the following: (i) the grid width is not too small: each cell of the grid contains at least one of the design points; (ii) the grid width is not too large: points in the same or neighboring cells of the grid are always connected in the K-NN graph.

Given $N \in \mathbb{N}$, a d-dimensional grid graph $G_{\text{lat}} \in [0,1]^d$ has vertex set V_{lat} and edge set E_{lat}. The grid graph has equal side lengths, and the total number of nodes $|V_{\text{lat}}| = N^d$. Without loss of generality, we assume that the nodes of the grid correspond to the points

$$
P_{\text{lat}}(N) = \left\{\left(\frac{i_1}{N} - \frac{1}{2N}, ..., \frac{i_d}{N} - \frac{1}{2N}\right) : i_1, ..., i_d \in \{1, ..., N\}\right\}. \tag{2.23}
$$

Moreover, for $z, z' \in P_{\text{lat}}(N)$, $(z, z') \in E_{\text{lat}}(N)$ if and only if $\|z - z'\|_2 = \frac{1}{N}$.

Now, we define $I(N) = h^{-1}\{P_{\text{lat}}(N)\}$ as the mesh in the covariate space $\mathcal{X}$ corresponding to the grid graph $G_{\text{lat}}(N) \in [0,1]^d$ through the homeomorphism h from Assumption 5. In general, $I(N)$ performs as a quantification in the domain $\mathcal{X}$; see [ALL14] for more details. We denote the elements in $I(N)$ by $u_1, ..., u_{N^d}$, and define a collection of cells $\{C(x)\} \in \mathcal{X}$ for $x \in I(N)$ as

$$C(x) = h^{-1}\left(\left\{z \in [0,1]^d : h(x) = \underset{x' \in P_{\text{lat}}(N)}{\arg\min} \|z - x'\|_\infty\right\}\right). \tag{2.24}$$

In order to analyze the behavior of the proposed estimator $\hat{\theta}$ through the grid embedding, we construct two vectors denoted by $\theta_I \in \mathbb{R}^n$ and $\theta^I \in \mathbb{R}^{N^d}$ for any signal $\theta \in \mathbb{R}^n$.

The first vector, $\theta_I \in \mathbb{R}^n$ incorporates information about the samples $X = (x_1, ..., x_n)$ and the cells $\{C(x)\}$. The idea is to force covariates x_i fallen in the same cell to take the same signal value after mapping with the homeomorphism. Formally, we define

$$(\theta_I)_i = \theta_j \quad \text{where} \quad j = \underset{l=1,...,n}{\arg\min} \|h(P_I(x_i)) - h(x_l)\|_\infty, \tag{2.25}$$

where $P_I(x)$ is the point in $I(N)$ such that $x \in C(P_I(x))$ and if there exists multiple points satisfying the condition, we arbitrarily select one.

The second vector, $\theta^I \in \mathbb{R}^{N^d}$ records coordinates corresponding to the different nodes of the mesh (centers of cells), and is associated with θ_I. We first induce a signal in $\mathbb{R}^{N^d}$ corresponding to the elements in $I(N)$ as

$$I_j = \{i = 1, ..., n : P_I(x_i) = u_j\}, \text{ for } j = 1, ..., N^d.$$

If $I_j \neq \emptyset$, then there exists $i_j \in I_j$ such that $(\theta_I)_i = \theta_{i_j}$ for all $i \in I_j$; if I_j is empty, we require $\theta_{i_j} = 0$. We can thus define

$$\theta^I = \left(\theta_{i_1}, ..., \theta_{i_{N^d}}\right).$$

Note that in the proof of Lemma 11 in [PSC20], the authors showed that under Assumptions 3-5, with probability approaching 1,

$$\max_{x \in I(N)} |C(x)| \leq \text{poly}(\log n). \tag{2.26}$$

We will use this inequality in our proofs later, but we will not make the polynomial function of $\log n$ explicit.

2.7.5 Definition of Piecewise Lipschitz

To make sure Assumption 1 is valid, we require a piecewise Lipschitz condition on the regression function f_0. In this section, we provide the detailed definition of the class of piecewise Lipschitz functions, followed from Definition 1 in [PSC20]. All notations follow the same from the main context, besides an extra notation on the boundary of a set A, denoted by ∂A.

Definition 2.7.12 *Let $\Omega_\epsilon := [0,1]^d \backslash B_\epsilon(\partial[0,1]^d)$. We say that a bounded function $g : [0,1]^d \to \mathbb{R}$ is piecewise Lipschitz if there exists a set $\mathcal{S} \subset (0,1)^d$ that has the following properties:*

- *The set $\mathcal{S}$ has Lebesgue measure zero.*
- *For some constants $C_{\mathcal{S}}, \epsilon_0 > 0$, we have that $\mu(h^{-1}\{B_\epsilon(\mathcal{S}) \cap ([0,1]^d \backslash \Omega_\epsilon)\}) \leq C_{\mathcal{S}}\epsilon$ for all $0 < \epsilon < \epsilon_0$.*
- *There exists a positive constant L_0 such that if z and z' belong to the same connected component of $\Omega_\epsilon \backslash B_\epsilon(\mathcal{S})$, then $|g(z) - g(z')| \leq L_0 \|z - z'\|_2$.*

Roughly speaking, a bounded function g is piecewise Lipschitz if there exists a small set $\mathcal{S}$ that partitions $[0,1]^d$ in such a way that g is Lipschitz within each connected component of the partition. Theorem 2.2.1 in [Zie12] implies that if g is piecewise Lipschitz, then g has bounded variation on any open set within a connected component.

2.7.6 Theorem 2.4.1

2.7.6.1 Notations

For a matrix D, we denote its kernel by $\mathrm{Ker}(D)$ and its Moore–Penrose inverse by $D^\dagger$. We also write Π as the projection matrix onto $\mathrm{Ker}(D)$, and denote $\mathrm{Ker}(D)^\perp$ as the orthogonal complement to the kernel space of D.

Our goal is to upper bound the expectation of $M(\cdot) - \hat{M}(\cdot)$ in the constrained set S defined by

$$S = \left\{\theta \in \mathbb{R}^n \ : \ \|\nabla_G \theta\|_1 \leq V n^{1-1/d} \mathrm{poly}(\log n)\right\}, \tag{2.27}$$

where $V = \|\nabla_G \theta^*\|_1 / [n^{1-1/d}\mathrm{poly}(\log n)]$ and $V \geq V^*$. Through the K-NN embedding, we can instead bound the expected loss in the embedded set defined by

$$\tilde{S} = \left\{\theta \in \mathbb{R}^n \ : \ \|D\theta^I\|_1 \leq V n^{1-1/d} \mathrm{poly}(\log n)\right\}, \tag{2.28}$$

where θ^I follows the definition in Appendix D.

2.7.6.2 Auxiliary lemmas for Proof of Theorem 2.4.1

Lemma 2.7.13 *For $v \in \mathbb{R}^n$, if $\|\nabla_G v\|_1 \leq \tilde{V}$, and $\Delta^2(v - \theta^*) \leq t^2$, then $\|Dv^I\|_1 \leq \tilde{V}$, and $\Delta^2(v^I - \theta^{*,I}) \leq t^2$.*

Proof. Lemma 4 in [PSC20] obtains the inequality

$$\|Dv^I\|_1 \leq \|\nabla_G v\|_1, \ \forall v \in \mathbb{R}^\cdot$$

Hence, the first claim follows.

Next, we observe that for any vector $u \in \mathbb{R}^n$,

$$\Delta^2(u^I) = \left(\sum_{j=1}^{N^d} \min\left\{|u_{i_j}|, u_{i_j}^2\right\}\right) \leq \left(\sum_{i=1}^{n} \min\left\{|u_i|, u_i^2\right\}\right) \leq \Delta^2(u).$$

The second claim follows then. □

Lemma 2.7.13 gives us the fact that

$$\left\{v \in S : \Delta^2(v - \theta^*) \leq t^2\right\} \subseteq \left\{v \in \tilde{S} : \Delta^2(v^I - \theta^{*,I}) \leq t^2\right\}. \tag{2.29}$$

Lemma 2.7.14 *Under Assumptions 1-5, we have that*

$$\begin{aligned}\mathbb{E}\left[\sup_{v\in\tilde{S}:\Delta^2(v^I-\theta^{*,I})\leq t^2}\sum_{i=1}^{n}\xi_i(v_i-\theta_i^*)\right] \;\leq\; & \operatorname{poly}(\log n)\mathbb{E}\left[\sup_{v\in\tilde{S}:\Delta^2(v^I-\theta^{*,I})\leq t^2}\sum_{j=1}^{N^d}\tilde{\xi}_j(v_j^I-\theta_j^{*,I})\right] \\ & +2Vn^{1-1/d}\operatorname{poly}(\log n),\end{aligned}$$

where $\tilde{\xi} \in \mathbb{R}^{N^d}$ is a mean zero 1-subgaussian vector whose coordinates are independent.

Proof. We notice that

$$\begin{aligned}\xi^\top(v-\theta^*) &= \xi^\top(v-v_I) + \xi^\top(v_I-\theta_I^*) + \xi^\top(\theta_I^*-\theta^*) \\ &\leq 2\|\xi\|_\infty\left(\|\nabla_G v\|_1 + \|\nabla_G\theta^*\|_1\right) + \xi^\top(v_I-\theta_I^*),\end{aligned}$$

where the second inequality holds by Lemma 4 in [PSC20]. Moreover,

$$\xi^\top(v_I-\theta_I^*) = \sum_{j=1}^{N^d}\sum_{l\in I_j}\xi_l(v_{i_j}-\theta_{i_j}^*) = \left\{\max_{u\in I}|C(u)|\right\}^{1/2}\tilde{\xi}^\top(v^I-\theta^{*,I})$$

where

$$\tilde{\xi}_j = \left\{\max_{u\in I}|C(u)|\right\}^{-1/2}\sum_{l\in I_j}\xi_l.$$

Clearly, the $\tilde{\xi}_1, ..., \tilde{\xi}_{N^d}$ are independent and also zero-centered 1-subgaussian as the original Rademacher random variables $\xi_1, ..., \xi_n$. Then, following from (2.26), we have

$$\xi^\top(v_I-\theta_I^*) \leq \operatorname{poly}(\log n)\tilde{\xi}^\top(v^I-\theta^{*,I}).$$

Hence, the desired inequality holds. □

Lemma 2.7.15 *Let $\delta \in \mathbb{R}^{N^d}$ with $\Delta^2(\delta) \leq t^2$. Then*

$$\|\Pi\delta\|_\infty \leq \frac{t^2}{N^d} + \frac{t}{\sqrt{N^d}}.$$

Proof. Notice that $\mathrm{Ker}(D) = \mathrm{span}(1_{N^d})$, then $\Pi\delta = \delta^\top v \cdot v$, where $v = \frac{1}{\sqrt{N^d}} 1_{N^d}$. Hence

$$\|\Pi\delta\|_\infty = \|\delta^\top v \cdot v\|_\infty \leq |\delta^\top v| \cdot \|v\|_\infty = \frac{|\delta^\top v|}{\sqrt{N^d}}. \tag{2.30}$$

Now, we have that

$$\begin{aligned}
|\delta^\top v| &\leq \sum_{i=1}^{N^d} |\delta_i||v_i| \\
&= \sum_{i=1}^{N^d} |\delta_i||v_i| 1_{\{|\delta_i|>1\}} \ + \sum_{i=1}^{N^d} |\delta_i||v_i| 1_{\{|\delta_i|\leq 1\}} \\
&\leq \|v\|_\infty \sum_{i=1}^{N^d} |\delta_i| 1_{\{|\delta_i|>1\}} + \|v\| \left(\sum_{i=1}^{N^d} \delta_i^2 1_{\{|\delta_i|\leq 1\}} \right)^{1/2} \\
&\leq \frac{t^2}{\sqrt{N^d}} + t,
\end{aligned} \tag{2.31}$$

where the first inequality follows from the triangle inequality, the second from Hölder's and Cauchy Schwarz inequalities. The claim follows combining (2.30) with (2.31). □

Lemma 2.7.16 *Under Assumptions 1-5, for $N \asymp n^{1/d}$ and some positive constant C_d, we have that*

$$GW\left(\left\{ \delta : \delta \in \tilde{S}, \Delta^2(\delta) \leq t^2 \right\} \right) \leq C_d \left(\frac{t^2}{\sqrt{n}} + t + V n^{1-1/d} \mathrm{poly}(\log n) \right).$$

Proof. Recall that the projection on $\mathrm{Ker}(D)^\perp$, $D^\dagger D = I - \Pi$, which yields

$$\tilde{\xi}^\top \delta = \tilde{\xi}^\top \Pi\delta + \tilde{\xi}^\top D^\dagger D\delta.$$

Then we have

$$\begin{aligned}
\mathbb{E}\left[\sup_{\delta\in\tilde{S}:\Delta^2(\delta)\leq t^2} \tilde{\xi}^\top \delta \right] &\leq \mathbb{E}\left[\sup_{\delta\in\tilde{S}:\Delta^2(\delta)\leq t^2} \tilde{\xi}^\top \Pi\delta \right] + \mathbb{E}\left[\sup_{\delta\in\tilde{S}:\Delta^2(\delta)\leq t^2} \tilde{\xi}^\top D^\dagger D\delta \right] \\
&=: \ T_1 + T_2.
\end{aligned} \tag{2.32}$$

We first bound T_1. Notice that Π is idempotent, i.e., $\Pi^2 = \Pi$, thus

$$\tilde{\xi}^\top \Pi\delta = \tilde{\xi}^\top \Pi\Pi\delta$$

$$
\begin{aligned}
&\leq \|\tilde{\xi}^\top \Pi\|_1 \|\Pi\delta\|_\infty \\
&= \left| \sum_{j=1}^{N^d} \tilde{\xi}_j \right| \cdot \|\Pi\delta\|_\infty \\
&= \left| \sum_{j=1}^{N^d} \frac{\tilde{\xi}_j}{N^{d/2}} \right| \cdot N^{d/2} \|\Pi\delta\|_\infty
\end{aligned} \tag{2.33}
$$

where the inequality follow from Hölder's inequality, and $v = \frac{1}{\sqrt{N^d}} 1_{N^d}$. Then, from Lemma 2.7.15,

$$
T_1 \leq \mathbb{E}\left[\left| \sum_{j=1}^{N^d} \frac{\tilde{\xi}_j}{N^{d/2}} \right| \right] \left(\frac{t^2}{\sqrt{N^d}} + t \right) \leq C_d \left(\frac{t^2}{\sqrt{n}} + t \right) \tag{2.34}
$$

for some positive constant C_d.

Next, we bound T_2. Notice that, by Hölder's inequality and the definition of $\tilde{S}$ in (2.28), it follows that

$$
\begin{aligned}
T_2 &\leq \mathbb{E}\left[\sup_{\delta \in \tilde{S}: \Delta^2(\delta) \leq t^2} \|\tilde{\xi}^\top D^\dagger\|_\infty \|D\delta\|_1 \right] \\
&\leq V n^{1-1/d} \text{poly}(\log n) \mathbb{E}\left[\|\tilde{\xi}^\top D^\dagger\|_\infty \right],
\end{aligned} \tag{2.35}
$$

and thus we only need to bound $\mathbb{E}\left[\|\tilde{\xi}^\top D^\dagger\|_\infty \right]$. From Section 3 in [HR16], we write $D^\dagger = [s_1, ..., s_m]$, and $\max_{j=1,...,m} \|s_j\|_2$ is bounded above by bounded by $(\log n)^{1/2}$ for $d = 2$ or by a constant for $d > 2$. Then,

$$
\begin{aligned}
\mathbb{E}\left[\|\tilde{\xi}^\top D^\dagger\|_\infty \right] &= \mathbb{E}\left[\max_{j=1,...,m} |s_j^\top \tilde{\xi}| \right] \\
&= \mathbb{E}\left[\max_{j=1,...,m} |\tilde{s}_j^\top \tilde{\xi}| \right] \max_{j=1,...,m} \|s_j\|_2,
\end{aligned} \tag{2.36}
$$

where

$$
\tilde{s}_j = \frac{s_j}{\max_{j=1,...,m} \|s_j\|_2}.
$$

Moreover, $\tilde{s}_j^\top \tilde{\xi}$ is also zero-centered sub-Gaussian but with parameter at most 1. Combining (2.35) with (2.36), we obtain that

$$
T_2 \leq C_d V n^{1-1/d} \text{poly}(\log n),
$$

for some positive constant C_d. The conclusion follows then since standard normal random variables are 1-subgaussian. $\square$

Theorem 2.7.17 *Suppose that*

$$L \cdot GW\left(\{\delta \in \tilde{S} : \Delta^2(\delta^I) \leq \eta^2\}\right) \leq \kappa(\eta),$$

for a function $\kappa : \mathbb{R} \to \mathbb{R}$. Then for all $\eta > 0$, we have that

$$\mathbb{P}\left(\Delta^2(\hat{\delta}) > \eta^2\right) \leq \frac{L\text{poly}(\log n)\kappa(\eta)}{c_\tau \eta^2} + \frac{Vn^{1-1/d}\text{poly}(\log n)}{c_\tau \eta^2},$$

where L is the constant from Lemma 2.7.4 and c_τ is the constant from Lemma 2.7.8. Furthermore, if $\{r_n\}$ is a sequence such that

$$\lim_{t\to\infty} \sup_n \left[\frac{L\text{poly}(\log n)\kappa(tr_n n^{1/2})}{t^2 r_n^2 n} + \frac{Vn^{1-1/d}\text{poly}(\log n)}{t^2 r_n^2 n}\right] \to 0,$$

then

$$\frac{1}{n}\Delta^2(\hat{\theta} - \theta^*) = O_{\mathbb{P}}(r_n^2).$$

Proof. Let $\hat{\delta} = \hat{\theta} - \theta^*$ and suppose that

$$\frac{1}{n}\hat{\delta}^2 > \frac{\eta^2}{n}. \tag{2.37}$$

Next, let $q^2 = \Delta^2(\hat{\delta})$. Then define $g : [0,1] \to \mathbb{R}$ as $g(t) = \Delta^2(t\hat{\delta})$. Clearly, g is a continuous function with $g(0) = 0$ and $g(1) = q^2$. Therefore, there exists $t_{\hat{\delta}}$ such that $g(t_{\hat{\delta}}) = \eta^2$. Hence, letting $\tilde{\delta} = t_{\hat{\delta}}$ we observe that by the basic inequality $\hat{M}(\theta^* + \tilde{\delta}) \leq 0, \delta \in S$ by convexity of S, and $\Delta^2(\tilde{\delta}) = \eta^2$ by construction. This implies, along with Lemma 2.7.8, that

$$\sup_{v \in S : \Delta^2(v-\theta^*) \leq \eta^2} M(v) - \hat{M}(v) \geq M(\theta^* + \tilde{\delta}) - \hat{M}(\theta^* + \tilde{\delta}) \geq M(\theta^* + \tilde{\delta}) \geq c_\tau \eta^2.$$

Therefore, we have that

$$\begin{aligned}
\mathbb{P}\left\{\Delta^2(\hat{\delta}) > \eta^2\right\} &\leq \mathbb{P}\left\{\sup_{v \in S : \Delta^2(v-\theta^*) \leq \eta^2} M(v) - \hat{M}(v) \geq c_\tau \eta^2\right\} \\
&\leq \frac{1}{c_\tau \eta^2}\mathbb{E}\left\{\sup_{v \in S : \Delta^2(v-\theta^*) \leq \eta^2} M(v) - \hat{M}(v)\right\}
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{1}{c_\tau \eta^2} \mathbb{E}\left\{ \sup_{v \in \tilde{S}: \Delta^2(v-\theta^*) \leq \eta^2} M(v) - \hat{M}(v) \right\} \\
&\leq \frac{L \text{poly}(\log n)}{c_\tau \eta^2} GW\left(\{\delta \in \tilde{S} : \Delta^2(\delta^I) \leq \eta^2\} \right) + \\
&\quad \frac{V n^{1-1/d} \text{poly}(\log n)}{c_\tau \eta^2} \\
&\leq \frac{L \text{poly}(\log n) \kappa(\eta)}{c_\tau \eta^2} + \frac{V n^{1-1/d} \text{poly}(\log n)}{c_\tau \eta^2},
\end{aligned}
$$

where the second inequality follows from Markov's inequality, the third inequality from (2.29), the fourth inequality by combining Corollary 2.7.9, Lemmas 2.7.10, 2.7.11 and 2.7.14, and the last from Lemma 2.7.16. This completes the proof. □

2.7.6.3 Proof of Theorem 2.4.1

Proof. The claim follows immediately from Lemmas 2.7.14 and 2.7.16 and Theorem 2.7.17 by setting

$$r_n \asymp n^{-1/2d} \text{poly}(\log n).$$

□

2.7.7 Theorem 2.4.2

2.7.7.1 Auxiliary lemmas for Proof of Theorem 2.4.2

Throughout, we assume that Assumptions 1-5 hold, and all notations follow the same as in the proof of Theorem 1.

Lemma 2.7.18 *Let $\epsilon \in (0,1)$, then there exists a choice*

$$\lambda = \begin{cases} \Theta\{\log n\} & \text{for } d = 2, \\ \Theta\left\{(\log n)^{1/2}\right\} & \text{for } d > 2, \end{cases}$$

such that for a constant $C_0 > 0$, we have that, with probability at least $1 - \epsilon/4$,

$$\kappa(\hat{\theta} - \theta^*) \in \mathcal{A},$$

with

$$\mathcal{A} := \left\{ \delta : \|\nabla_G \delta\|_1 \leq C_0 \left(\|\nabla_G \theta^*\|_1 + \frac{R_1}{R_2} \left[\frac{\Delta^2(\delta)}{n^{1/2}} + \Delta(\delta) \right] \right) \right\}$$

for all $\kappa \in [0,1]$, *where*

$$R_1 = C_d \log \left\{ \frac{n}{\epsilon} \right\}^{1/2},$$

$$R_2 = \begin{cases} C_d (\log n)^{1/2} \left[\log \left\{ \frac{c_k n}{\epsilon} \right\} \right]^{1/2} & \text{for } d = 2, \\ C_d \left[\log \left\{ \frac{c_k n}{\epsilon} \right\} \right]^{1/2} & \text{for } d > 2, \end{cases}$$

and C_0, C_d *are positive constants.*

Proof. Pick $\kappa \in [0,1]$ fixed, and let $\tilde{\delta} = \kappa(\hat{\theta} - \theta^*)$. Then by the optimality of $\hat{\theta}$ and the convexity of (2.8), we have that

$$\sum_{i=1}^n \rho_\tau(y_i - \tilde{\theta}_i) + \lambda \|\nabla_G \tilde{\theta}\|_1 \leq \sum_{i=1}^n \rho_\tau(y_i - \theta_i^*) + \lambda \|\nabla_G \theta^*\|_1,$$

where $\tilde{\theta} = \theta^* + \tilde{\delta}$. Then as in the proof of Lemma 3 from [BC11],

$$0 \leq \lambda \left[\|\nabla_G \theta^*\|_1 - \|\nabla_G \tilde{\theta}\|_1 \right] + (\tilde{\theta} - \theta^*)^\top a^*, \tag{2.38}$$

where $a_i^* = \tau - 1\{y_i \leq \theta_i^*\}$ for $i = 1, ..., n$.

Next, we bound the second term of the right hand of (2.38). From Lemma 2.7.14, we know it is sufficient to bound

$$\tilde{a}^\top (\tilde{\theta}^I - \theta^{*,I}),$$

where

$$\tilde{a}_j = \left\{ \max_{u \in I} |C(u)| \right\}^{-1/2} \sum_{l \in I_j} a_l^*.$$

Now,

$$\begin{aligned} \tilde{a}^\top (\tilde{\theta}^I - \theta^{*,I}) &= \tilde{a}^\top \Pi (\tilde{\theta}^I - \theta^{*,I}) + \tilde{a}^\top D^\dagger D (\tilde{\theta}^I - \theta^{*,I}) \\ &=: A_1 + A_2 \end{aligned} \tag{2.39}$$

From Lemmas 2.7.13, 2.7.15 and (2.33), we obtain

$$A_1 \leq \|\tilde{a}\|_\infty \left(\frac{\Delta^2(\tilde{\theta}^I - \theta^{*,I})}{n^{1/2}} + \Delta(\tilde{\theta}^I - \theta^{*,I}) \right)$$

$$\leq \|\tilde{a}\|_\infty \left(\frac{\Delta^2(\tilde{\theta} - \theta^*)}{n^{1/2}} + \Delta(\tilde{\theta} - \theta^*) \right) \tag{2.40}$$

To bound A_2, we use the result of Lemma 2.7.13 to obtain

$$\begin{aligned} A_2 &\leq \|(D^\dagger)^\top \tilde{a}\|_\infty \left[\|D\tilde{\theta}^I\|_1 + \|D\theta^{*,I}\|_1 \right] \\ &\leq \|(D^\dagger)^\top \tilde{a}\|_\infty \left[\|\nabla_G \tilde{\theta}\|_1 + \|\nabla_G \theta^*\|_1 \right] \end{aligned} \tag{2.41}$$

Since a^* is bounded and is thus sub-Gaussian, $\tilde{a}$ is also sub-Gaussian. As in the proof of Theorem 2 from [HR16], we have that the following two inequalities hold simultaneously on an event of probability at least $1 - 2\epsilon$,

$$\|\tilde{a}\|_\infty \leq R_1 := C_d \log\left\{\frac{n}{\epsilon}\right\}^{1/2}, \quad \|(D^\dagger)^\top \tilde{a}\|_\infty \leq R_2 := \begin{cases} C_d (\log n)^{1/2} \left[\log\left\{\frac{c_k n}{\epsilon}\right\}\right]^{1/2} & \text{for } d = 2, \\ C_d \left[\log\left\{\frac{c_k n}{\epsilon}\right\}\right]^{1/2} & \text{for } d > 2, \end{cases}$$

for some constant c_k.

Then, with probability at least $1 - 2\epsilon$, By choosing $\lambda = 2R_2$, we obtain

$$\kappa(\hat{\theta} - \theta^*) \in \mathcal{A} := \left\{ \delta : \|\nabla_G \delta\|_1 \leq C_0 \left(\|\nabla_G \theta^*\|_1 + \frac{R_1}{R_2} \left[\frac{\Delta^2(\delta)}{n^{1/2}} + \Delta(\delta) \right] \right) \right\},$$

for some positive constant C_0. □

2.7.7.2 Proof of Theorem 2.4.2

Proof. Let $\epsilon \in (0, 1)$. By Lemma 2.7.18 we can suppose that the following event

$$\Omega = \left\{ \kappa(\hat{\theta} - \theta^*) \in \mathcal{A},\ \forall \kappa \in [0, 1] \right\} \tag{2.42}$$

happen with probability at least $1 - \epsilon/2$ with $\mathcal{A}$ as in Lemma 2.7.18. Then,

$$\mathbb{P}\left\{ \Delta^2(\hat{\delta}) > \eta^2 \right\} \leq \mathbb{P}\left[\left\{ \Delta^2(\hat{\delta}) > \eta^2 \right\} \cap \Omega \right] + \frac{\epsilon}{2}.$$

Now suppose that the event

$$\left\{ \Delta^2(\hat{\delta}) > \eta^2 \right\} \cap \Omega$$

holds. As in the proof of Theorem 2.7.17, there exists $\tilde{\delta} = t_{\hat{\delta}}\hat{\delta}$ with $t_{\hat{\delta}} \in [0,1]$ such that $\tilde{\delta} \in \mathcal{A}$, $\Delta^2(\tilde{\delta}) = \eta^2$. Hence, by the basic inequality,

$$\hat{M}(\theta^* + \tilde{\delta}) + \lambda \left[\|\nabla_G(\theta^* + \tilde{\delta})\|_1 - \|\nabla_G \theta^*\|_1 \right] \leq 0.$$

Then,

$$\begin{aligned} \sup_{\delta \in \mathcal{A}, \Delta^2(\delta) \leq \eta^2} \left[M(\theta^* + \delta) - \hat{M}(\theta^* + \delta) + \lambda \left\{ \|\nabla_G \theta^*\|_1 - \|\nabla_G(\theta^* + \tilde{\delta})\|_1 \right\} \right] &\geq M(\theta^* + \delta) \\ &\geq c_\tau \eta^2, \end{aligned}$$

where the second inequality follows from Lemma 2.7.8. Therefore,

$$\begin{aligned} \mathbb{P}\left[\left\{ \Delta^2(\hat{\delta}) > \eta^2 \right\} \cap \Omega \right] &\leq \mathbb{P}\Bigg(\Bigg\{ \sup_{\delta \in \mathcal{A}, \Delta^2(\delta) \leq \eta^2} \Big[M(\theta^* + \delta) - \hat{M}(\theta^* + \delta) \\ &\qquad + \lambda \Big\{ \|\nabla_G \theta^*\|_1 - \|\nabla_G(\theta^* + \delta)\|_1 \Big\} \Big] \geq c_\tau \eta^2 \Bigg\} \cap \Omega \Bigg) \\ &\leq \frac{1}{c_\tau \eta^2} \mathbb{E}\Bigg(\mathbf{1}_\Omega \sup_{\delta \in \mathcal{A}, \Delta^2(\delta) \leq \eta^2} \Big[M(\theta^* + \delta) - \hat{M}(\theta^* + \delta) \\ &\qquad + \lambda \Big\{ \|\nabla_G \theta^*\|_1 - \|\nabla_G(\theta^* + \delta)\|_1 \Big\} \Big] \Bigg) \\ &\leq \frac{1}{c_\tau \eta^2} \mathbb{E}\left(\mathbf{1}_\Omega \sup_{\delta \in \mathcal{A}, \Delta^2(\delta) \leq \eta^2} \Big[M(\theta^* + \delta) - \hat{M}(\theta^* + \delta) \Big] \right) \\ &\quad + \frac{\lambda}{c_\tau \eta^2} \mathbb{E}\left(\mathbf{1}_\Omega \sup_{\delta \in \mathcal{A}, \Delta^2(\delta) \leq \eta^2} \left[\|\nabla_G \theta^*\|_1 - \|\nabla_G(\theta^* + \delta)\|_1 \right] \right) \\ &\leq \frac{1}{c_\tau \eta^2} \mathbb{E}\left(\mathbf{1}_\Omega \sup_{\delta \in \mathcal{A}, \Delta^2(\delta) \leq \eta^2} \Big[M(\theta^* + \delta) - \hat{M}(\theta^* + \delta) \Big] \right) \\ &\quad + \frac{\lambda}{c_\tau \eta^2} \mathbb{E}\left(\mathbf{1}_\Omega \sup_{\delta \in \mathcal{A}, \Delta^2(\delta) \leq \eta^2} \|\nabla_G \delta\|_1 \right), \end{aligned} \tag{2.43}$$

where the second inequality follows from Markov's inequality, and the last from the triangle inequality.

Next, define

$$\mathcal{H}(\eta) = \{\delta \in \mathcal{A} : \Delta(\delta) \leq \eta\}.$$

Hence, if $\delta \in \mathcal{H}(\eta)$ and Ω holds, then

$$\|\nabla_G \delta\|_1 \leq C_0 \left(\|\nabla_G \theta^*\|_1 + \frac{R_1}{R_2} \left[\frac{\Delta^2(\delta)}{n^{1/2}} + \Delta(\delta) \right] \right) \tag{2.44}$$

where the inequality follow from the definition of $\mathcal{H}(\eta)$ and Lemma 2.7.18.

We now define

$$\mathcal{L}(\eta) = \left\{ \delta : \|\nabla_G \delta\|_1 \leq C_0 \left\{ \|\nabla_G \theta^*\|_1 + \frac{R_1}{R_2} \left[\frac{\Delta^2(\delta)}{n^{1/2}} + \Delta(\delta) \right] \right\}, \Delta(\delta) \leq \eta \right\},$$

$$\tilde{\mathcal{L}}(\eta) = \left\{ \delta : \|D\delta^I\|_1 \leq C_0 \left\{ \|\nabla_G \theta^*\|_1 + \frac{R_1}{R_2} \left[\frac{\Delta^2(\delta)}{n^{1/2}} + \Delta(\delta) \right] \right\}, \Delta(\delta^I) \leq \eta \right\}.$$

Then

$$\begin{aligned}
\mathbb{P}\left[\left\{ \Delta^2(\hat{\delta}) > \eta^2 \right\} \cap \Omega \right] &\leq \frac{1}{c_\tau \eta^2} \mathbb{E} \left(\sup_{\delta \in \mathcal{L}(\eta)} \left[M(\theta^* + \delta) - \hat{M}(\theta^* + \delta) \right] \right) + \frac{\lambda}{c_\tau \eta^2} \sup_{\delta \in \mathcal{L}(\eta)} \|\nabla_G \delta\|_1 \\
&\leq \frac{2\text{poly}(\log n)}{c_\tau \eta^2} \mathbb{E} \left(\sup_{\delta \in \tilde{\mathcal{L}}(\eta)} \sum_{i=1}^{N^d} \tilde{\xi}_i \delta_i^I \right) + \frac{2C_0 \left[V^* n^{1-1/d} \text{poly}(\log n) + \frac{R_1}{R_2} \left(\frac{\eta^2}{n^{1/2}} + \eta \right) \right]}{c_\tau \eta^2} \\
&\quad + \frac{\lambda}{c_\tau \eta^2} \sup_{\delta \in \mathcal{L}(\eta)} \|\nabla_G \delta\|_1 \\
&\leq \frac{2L\text{poly}(\log n)}{c_\tau \eta^2} C_d \left[\frac{\eta^2}{n^{1/2}} + \eta + C_0 V^* n^{1-1/d} \text{poly}(\log n) + C_0 \frac{R_1}{R_2} \left(\frac{\eta^2}{n^{1/2}} + \eta \right) \right] \\
&\quad + \frac{2C_0}{c_\tau \eta^2} \left[V^* n^{1-1/d} \text{poly}(\log n) + \frac{R_1}{R_2} \left(\frac{\eta^2}{n^{1/2}} + \eta \right) \right] \\
&\quad + \frac{\lambda}{c_\tau \eta^2} C_0 \left[V^* n^{1-1/d} \text{poly}(\log n) + \frac{R_1}{R_2} \left(\frac{\eta^2}{n^{1/2}} + \eta \right) \right],
\end{aligned} \tag{2.45}$$

where the second inequality follows with the same argument from Lemmas 2.7.10, 2.7.11, and 2.7.14, and the last inequality follows from Lemma 2.7.16.

Hence given our choice of λ, by choosing

$$\eta = \ c_\gamma n^{\frac{1}{2}(1-1/d)} \text{poly}(\log n)$$

for some $c_\gamma > 1$, we conclude that

$$\mathbb{P}\left[\left\{ \Delta^2(\hat{\delta}) > \eta^2 \right\} \cap \Omega \right] \leq \epsilon,$$

provided that c_γ is large enough. $\square$

CHAPTER 3

2D Score Based Estimation of Heterogeneous Treatment Effects

3.1 Introduction

The questions that motivate many scientific studies in disciplines such as economics, epidemiology, medicine, and political science, are not associational but causal in nature. In the study of causal inference, many researchers are interested in inferring average treatment effects, which provide a good sense of whether treatment is likely to deliver more benefit than the control among a whole community. However, the same treatment may affect different individuals very differently. Therefore, a substantial amount of works focus on analyzing heterogeneity in treatment effects, of which the term refers to variation in the effects of treatment across individuals. This variation may provide theoretical insights, revealing how the effect of interventions depends on participants' characteristics or how varying features of a treatment alters the effect of an intervention.

We follow the potential outcome framework for causal inference [Spl23, Rub74], where each unit is assigned into either the treatment or the control group. Each unit has an observed outcome variable with a set of covariates. In randomized experiments and observational studies, it is desirable to replicate a sample as closely as possible by obtaining subjects from the treatment and control groups with similar covariate distributions when estimating causal effects. However, it is almost impossible to match observations exactly the same in both treatment and control groups in observational studies. To address this problem, it is usually preferred to define prespecified subgroups under certain conditions and estimate the

treatment effects varying among subgroups. Accordingly, the conditional average treatment effect (CATE; [Hah98]) function is designed to capture heterogeneity of a treatment effect across subpopulations. The function is often conditioned on some component(s) of the covariates or a single statistic, like propensity score [RR83] and prognostic score [Han08]. Propensity scores are the probabilities of receiving the treatment of interest; prognostic scores model the potential outcome under a control group assignment. To understand treatment effect heterogeneity in terms of propensity and prognostic scores, we assume that equal or similar treatment effects are observed along some intervals of the two scores.

We target at constructing an estimator for treatment effects that conditions on only two quantities, propensity and prognostic scores, and we assume a piecewise constant structure in treatment effects. We take a step further from score-based matching algorithms and propose a data-driven approach that integrates the joint use of propensity and prognostic scores in a matching algorithm and a partition over the entire population via a non-parametric regression tree. In the first step, we estimate propensity scores and prognostic scores for each observed unit in the data. Secondly, we perform a K-nearest-neighbor matching of units of the treatment and control groups based on the two estimated scores and forth construct a proxy of individual treatment effects for all units. The last step involves growing a binary tree regressed on the two estimated scores.

The complementary nature of propensity and prognostic score methods supports that conditioning on both the propensity and prognostic scores has the potential to reduce bias and improve the precision of treatment effect estimates, and it is affirmed in the simulation studies by [LS14] and [ACP18]. We also demonstrate such advantage for our proposed estimator across almost all scenarios examined in the simulation experiments.

Besides high precision in estimation, our proposed estimator demonstrates its superiority over state-of-arts methods with a few attractive properties as follows:

- The estimator is computationally efficient. Propensity and prognostic scores can be easily estimated through simple regression techniques. Our matching algorithm based

on the two scores largely reduces dimensionality compared to full matching on the complete covariates. Moreover, growing a single regression tree saves much time over other tree-based estimation methods, such as BART [HMC20] and random forests [WA18, ATW19].

- Many previous works in subgroup analysis, such as [APE00] and [ACW18], set stratification on the sample with a fixed number of subgroups before estimating treatment effects. These approaches require a pre-determination on the number of subgroups contained in the data, and they inevitably introduce arbitrariness into the causal inference. In comparison, our proposed method simultaneously identifies the underlying subgroups in observations through binary split according to propensity and prognostic scores and provides a consequential estimation of treatment effects on each subgroups.

- Although random forests based methods [WA18, ATW19] achieve great performance in minimizing bias in estimating treatment effects, these ensemble methods are often referred to as "black boxes". It is hard to capture the underlying reason why the collective decision with the high number of operations involved is made in their estimation process, especially in the case when there are many features considered in the model. On the contrary, our proposed method provides a straightforward and low-dimensional summary of treatment effects simply based on two scores that are relatively easy to estimate. As a result, given the covariates of an observation, one can easily deduce the positiveness and magnitude of its treatment effect according to its probability of treatment receipt and potential outcome following the structure of the regression tree without carrying out feature selection process.

- Instead of relying on a single feature or a subset of features, identifying subgroups with the joint distribution of all features will be of particular interest to policymakers seeking to target policies on those most likely to benefit. Therefore, many existing researches focus on subgroup stratification based on a single index that combines baseline characteristics. For instance, [ACW18] utilizes potential outcomes without treatment

(prognostic scores) for endogenous stratification, and [KH07] stratifies experimental subjects based on their predicted probability of certain risks (propensity scores). Our method can be considered as a natural extension of the two previous works, which allows researchers and policy makers to merge a large number of baseline characteristics into simply two indices and identify the subgroup who are most in need of help in a population.

We review relevant literature on matching algorithms and estimation of heterogeneous treatment effects in Section 3.2. In Section 3.3, we provide the theoretical framework and preliminaries for the causal inference model. We propose our method for estimation and prediction in Section 3.4. Section 3.5 lists the results of numerical experiments on multiple simulated data sets and two real-world data sets, following with the comparison with state-of-the-art methods in existing literature and the discussion on policy implications under different realistic scenarios.

3.2 Relevant Literature

Statistical analysis of causality can be dated back to [Spl23]. Causal inference can be viewed as an identification problem [Kee15], for which statisticians are dedicated to learn the true causality behind the data. In reality, however, we do not have enough information to determine the true value due to a limited number of observations for analysis. This problem is also summarized as a "missing data" problem [DL18], which stems from the *fundamental problem of causal inference* [Hol86], that is, for each unit at most one of the potential outcomes is observed. Importantly, the causal effect identification problem, especially for estimating treatment effects, can only be resolved through assumptions. Several key theoretical frameworks have been proposed over the past decades. The potential outcome framework by [Rub74], often referred to as the Rubin Causal Model (RCM; [Hol86]), is a common model of causality in statistics at the moment. [Daw00] develops a decision theoretic approach to causality that rejects counterfactuals. [Pea95, Pea09] advocate for a model of causality

based on non-parametric structural equations and path diagrams.

Matching

To tackle the "missing data" problem when estimating treatment effects in randomized experiments in practice, matching serves as a very powerful tool. The main goal of matching is to find matched groups with similar or balanced observed covariate distributions [Stu10]. The exact K-nearest-neighbor matching [Rub74] is one of the most common, and easiest to implement and understand methods; and ratio matching [Smi97, RT00, MR01a], which finds multiple good matches for each treated individual, performs well when there is a large number of control individuals. [Ros89, GR93, Zub12, ZK17] developed various optimal matching algorithms to minimize the total sum of distances between treated units and matched controls in a global sense. [AI06] studied the consistency of covariate matching estimators under large sample assumptions. Instead of greedy matching on the entire covariates, propensity score matching (PSM) by [RT96] is an alternative algorithm that does not guarantee optimal balance among covariates and reduces dimension sufficiently. [Imb04] improved propensity score matching with regression adjustment. The additional matching on prognostic factors in propensity score matching was first considered by [RT00]. Later, [LS14] demonstrated the superiority of the joint use of propensity and prognostic scores in matching over single-score based matching in low-dimensional settings through extensive simulation studies. [ACP18] extended the method to fit to high-dimensional settings and derived asymptotic results for the so-called doubly robust matching estimators. The sequential works by [AGB20] and [AB22] pioneered in visualizing the relationship between propensity and prognostic scores as auxiliary in the estimation of treatment effects.

Subclassification

To understand heterogeneity of treatment effects in the data, subclassification, first used in [Coc68], is another important research problem. The key idea is to form subgroups over

the entire population based on characteristics that are either immutable or observed before randomization. [RR83, RR85], and [LD04] examined how creating a fixed number of subclasses according to propensity scores removes the bias in the estimated treatment effects, and [YIC16] developed a similar methodology in settings with more than two treatment levels. Full matching [Ros91, Han04, SG08] is a more sophisticated form of subclassification that selects the number of subclasses automatically by creating a series of matched sets. [SM15] presented three measures derived using the theory of order statistics to claim heterogeneity of treatment effect across subgroups. [STW09] pioneered in exploiting standard regression tree methods [BFS84] in subgroup treatment effect analysis. Further, [AI16] derived a recursive partition of the population according to treatment effect heterogeneity. [Hil11] was the first work to advocate for the use of Bayesian additive regression tree models (BART; [CGM10]) for estimating heterogeneous treatment effects, followed by a significant number of research papers focusing on the seminal methodology, including [GK12], [HS13] and [HMC20]. [ACW18] introduced endogenous stratification to estimate subgroup effects for a fixed number of subgroups based on certain quantiles of the prognostic score. More recently, [PCR21] combined the fused lasso estimator with score matching methods to lead to a data-adaptive subgroup effects estimator.

Machine Learning for Causal Inference

For the goal of analyzing treatment effect heterogeneity, supervised machine learning methods play an important role. One of the more common ways for accurate estimation with experimental and observational data is to apply regression [IR15] or tree-based methods [IS11]. From a Bayesian perspective, [HLP14] provided a principled way of adding priors to regression models, and [TGC16] developed Bayesian non-parametric approaches for both linear regression and tree models. The recent breakthrough work by [WA18] proposed the causal forest estimator arising from random forests from [Bre01]. More recently, [ATW19] took a step forward and enhanced the previous estimator based on generalized random forests. [IR13] adapted an estimator from the Support Vector Machine (SVM) classifier with

hinge loss [Wah02]. [BLZ16] studied treatment effect estimators with lasso regularization [Tib96] when the number of covariates is large, and [KVW18] applied group lasso for simultaneous covariate selection and robust estimation of causal effects. In the meantime, a series of papers including [QM11], [KSB19] and [SLO19], focused on developing meta-learners for heterogeneous treatment effects that can take advantage of various machine learning algorithms and data structures.

Applied Work

On the application side, the estimation of heterogeneous treatment effects is particularly an intriguing topic in causal inference with broad applications in scientific research. [GK11] estimated heterogeneous treatment effects in randomized experiments in the context of political science. [DW02] explored the use of propensity score matching for nonexperimental causal studies with application in economics. [DHK16] investigated heterogeneous treatment effects to provide the evidence base for precision medicine and patient-centred care. [ZLL17] proposed the Survival Causal Tree (SCT) method to discover patient subgroups with heterogeneous treatment effects from censored observational data. [RPR20] examined three classes of approaches to identify heterogeneity of treatment effect within a randomized clinical trial, and [TTT17] rendered a subgroup treatment estimate for drug trials.

3.3 Preliminaries

Before we introduce our method, we need to provide some mathematical background for treatment effect estimation. We follow Rubin's framework on causal inference [Rub74], and assume a superpopulation or distribution $\mathcal{P}$ from which a realization of n independent random variables is given as the training data. That is, we are given $\{(Y_i(0), Y_i(1), X_i, Z_i)\}_{i=1}^n$ independent copies of $(Y(1), Y(0), X, Z)$, where $X_i \in \mathbb{R}^d$ is a d-dimensional covariate or feature vector, $Z_i \in \{0, 1\}$ is the treatment-assignment indicator, $Y_i(0) \in \mathbb{R}$ is the potential outcome of unit i when i is assigned to the control group, and $Y_i(1)$ is the potential outcome

when i is assigned to the treatment group.

One important and commonly used measure of causality in a binary treatment model is the average treatment effect (ATE; [Imb04]), that is, the mean outcome difference between the treatment and control groups. Formally, we write the ATE as

$$\text{ATE} := \mathbb{E}[Y(1) - Y(0)].$$

With the n units in the study, we further define the individual treatment effect (ITE) of unit i denoted by D_i as

$$D_i := Y_i(1) - Y_i(0).$$

Then, an unbiased estimate of the ATE is the sample average treatment effect

$$\bar{Y}(1) - \bar{Y}(0) = \frac{1}{n}\sum_{i=1}^{n} D_i.$$

However, we cannot observe D_i for any unit because a unit is either in the treatment group or in the control group, but not in both.

To analyze heterogeneous treatment effects, it is natural to divide the data into subgroups (e.g., by gender, or by race), and investigate if the average treatment effects are different across subgroups. Therefore, instead of estimating the ATE or the ITE directly, statisticians seek to estimate the conditional average treatment effect (CATE), defined by

$$\tau(x) := \mathbb{E}[Y(1) - Y(0) \mid X = x]. \tag{3.1}$$

The CATE can be viewed as an ATE in a subpopulation defined by $\{X = x\}$, i.e. the ATE conditioned on membership in the subgroup.

We also recall the propensity score [RR83], denoted by $e(X)$, and defined as

$$e(X) = \mathbb{P}(Z = 1 \mid X).$$

Thus, $e(X)$ is the the probability of receiving treatment for a unit with covariate X. In addition, we consider prognostic scores, denoted by $p(X)$, for potential outcomes. The prognostic score is first defined by [Han08] as any quantity $p(X)$ that satisfies

$$Y(0) \perp\!\!\!\perp X \mid p(X),$$

and in this chapter, we use the conventional definition as the predicted outcome under the control condition [AGB20]:

$$p(X) = E[Y(0) \mid X].$$

We are interested in constructing a 2d summary of treatment effects based on propensity and prognostic scores. Instead of conditioning on the entire covariates or a subset of it in the CATE function, we express our estimand, named as scored-based subgroup CATE, by conditioning on the two scores:

$$\tau(x) := \mathbb{E}[Y(1) - Y(0) \mid e = e(x), p = p(x)]. \tag{3.2}$$

For interpretability, we assume that treatment effects are piecewise constant over a 2d grid of propensity and prognostic scores. Specifically, there exists a partition of intervals $\{I_1^e, ..., I_s^e\}$ of $[0, 1]$ and another partition of intervals $\{I_1^p, ..., I_t^p\}$ of $\mathbb{R}$ such that for any $i \in \{1, ..., s\}$ and $j \in \{1, ..., t\}$, we have

$$\tau(x) \equiv C_{i,j} \quad \text{for } x \text{ s.t. } e(x) \in I_i^e, p(x) \in I_j^p,$$

where $C_{i,j} \in \mathbb{R}$ is a constant.

Moreover, our estimation of treatment effects relies on the following assumptions:

Assumption 6 *Throughout the chapter, we maintain the Stable Unit Treatment Value Assumption (SUTVA; [IR15]), which consists of two components: no interference and no hidden variations of treatment. Mathematically, for unit $i = 1, ..., n$ with outcome Y_i and treatment indicator Z_i, it holds that*

$$Y_i(Z_1, Z_2, ..., Z_n) = Y_i(Z_i).$$

Thus, the SUTVA requires that the potential outcomes of one unit should be unaffected by the particular assignment of treatments to the other units. Furthermore, for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes.

Assumption 7 *The assumption of probabilistic assignment holds. This requires the assignment mechanism to imply a non-zero probability for each treatment value, for every unit. For the given covariates X and treatment-assignment indicator Z, we must have*

$$0 < \mathbb{P}(Z = 1 \mid X) < 1,$$

almost surely.

This condition, regarding the joint distribution of treatments and covariates, is also known as overlap in some literature (see Assumption 2.2 in [Imb04] and [DDF21]), and it is necessary for estimating treatment effects everywhere in the defined covariate space. Note that $\mathbb{P}(Z_i = 1 \mid X_i)$ is the propensity score. In other words, Assumption 7 requires that the propensity score, for all values of the treatment and all combinations of values of the confounders, be strictly between 0 and 1.

Assumption 8 *We make the assumption that*

$$(Y(0), Y(1)) \perp\!\!\!\perp Z \mid e(X), p(X)$$

holds.

This assumption is an implication of the usual unconfoundedness assumption:

$$(Y(0), Y(1)) \perp\!\!\!\perp Z \mid X \tag{3.3}$$

Combining Assumption 7 and that in Equation (3.3), the conditions are typically referred as *strong ignorability* defined in [RR83]. Strong ignorability states which outcomes are observed or missing is independent of the missing data conditional on the observed data. It allows statisticians to address the challenge that the "ground truth" for the causal effect is not observed for any individual unit. We rewrite the conventional assumption by replacing the vector of covariates x with the joint of propensity score $e(x)$ and $p(x)$ to accord with our estimation target.

Provided that Assumptions 6-8 hold, it follows that

$$\mathbb{E}[Y(z) \mid e = e(x), p = p(x)] = \mathbb{E}[Y \mid e = e(x), p = p(x), Z = z],$$

and thus our estimand (3.2) is equivalent to

$$\tau(x) = \mathbb{E}[Y \mid e = e(x), p = p(x), Z = 1] - \mathbb{E}[Y \mid e = e(x), p = p(x), Z = 0]. \tag{3.4}$$

Thus, in this chapter we focus on estimating (3.4), which is equivalent to (3.2) if the assumptions above hold, but might be different if Assumption 8 is violated.

3.4 Methodology

We now formally introduce our proposal of a three-step method for estimating heterogeneous treatment effects and the estimation rule for a given new observation. We assume a sample of size n with covariate X, treatment indicator Z, and outcome variable Y, where the notations inherit from the previous section. Generally, we consider a low-dimensional set-up, where the sample size n is larger than the covariate dimension d. An extension of our proposed method to the high-dimensional case is discussed in this section as well.

Step 1

We first estimate propensity and prognostic scores for all observations in the sample. For propensity score, we apply a logistic regression on the entire covariate X and the treatment indicator Z by solving the optimization problem

$$\hat{\alpha} = \underset{\alpha \in \mathbb{R}^d}{\arg\min} \; -\sum_{i=1}^{n} \Big[Z_i \log \sigma(X_i^\top \alpha) + (1 - Z_i) \log(1 - \sigma(X_i^\top \alpha)) \Big], \tag{3.5}$$

where $\sigma(x) = \frac{1}{1+\exp(-x)}$ is the logistic function. With the coefficient vector $\hat{\alpha}$, we compute the estimated propensity scores $\hat{e}$ by

$$\hat{e}_i = \sigma\left(X_i^\top \hat{\alpha}\right).$$

For prognostic score, we restrict to the controlled group, and regress the outcome variable Y on the covariate X through ordinary least squares: we solve

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^d} \sum_{i:Z_i=0} (Y_i - X_i^\top \theta)^2, \tag{3.6}$$

and we estimate prognostic scores as

$$\hat{p}_i = X_i^\top \hat{\theta}.$$

Step 2

Next, we perform a nearest-neighbor matching based on the two estimated scores from the previous step. We adapt the notation from [AI06], and use the Mahalanobis distance norm as the distance metric in the matching algorithm. Mahalanobis distance matching, first proposed by [Han06], has been widely used in matching algorithms for causal inference [LS14, AGB20]. It has been found to perform well when the dimension of covariates are relatively low [Stu10], and thus it is more suitable than using standard Euclidean distance in our case when we have only propensity scores and prognostic scores as covariates.

Formally, for the units i and j with estimated propensity scores $\hat{e}_i, \hat{e}_j$ and propensity scores $\hat{p}_i, \hat{p}_j$, we define the score-based Mahalanobis distance between i and j by

$$d(i,j) = \left[\begin{pmatrix} \hat{e}_i \\ \hat{p}_i \end{pmatrix} - \begin{pmatrix} \hat{e}_j \\ \hat{p}_j \end{pmatrix} \right]^\top \Sigma^{-1} \left[\begin{pmatrix} \hat{e}_i \\ \hat{p}_i \end{pmatrix} - \begin{pmatrix} \hat{e}_j \\ \hat{p}_j \end{pmatrix} \right],$$

where Σ denotes the variance-covariance matrix of $(\hat{e}, \hat{p})^\top$.

Let $j_k(i)$ be the index $j \in \{1, 2, ..., n\}$ that solves $Z_j = 1 - Z_i$ and

$$\sum_{l:Z_l=1-Z_i} \mathbf{1}\{d(l,i) \leq d(j,i)\} = k,$$

where $\mathbf{1}\{\cdot\}$ is the indicator function. This is the index of the unit that is the kth closest to unit i in terms of the distance between two scores, among the units with the treatment opposite to that of unit i. We can now construct the K-nearest-neighbor set for unit i by

the set of indices for the first K matches for unit i,

$$\mathcal{J}_K(i) = \{j_1(i), ..., j_K(i)\}.$$

We then compute

$$\tilde{Y}_i = (2Z_i - 1)\left(Y_i - \frac{1}{K}\sum_{j \in \mathcal{J}_K(i)} Y_j\right). \tag{3.7}$$

Intuitively, the construction of $\tilde{Y}$ gives a proxy of the individual treatment effect (ITE) on each unit. We find K matches for each unit in the opposite treatment group based on the similarity of their propensity and prognostics scores, and the mean of the K matches is used to estimate the unobserved potential outcome for each unit.

Step 3

The last step involves denoising of the point estimates of the individual treatment effects $\tilde{Y}$ obtained from Step 2. The goal is to partition all units into subgroups such that the estimated treatment effects would be constant over some 2d intervals of propensity and prognostic scores (see the left of Figure 3.1).

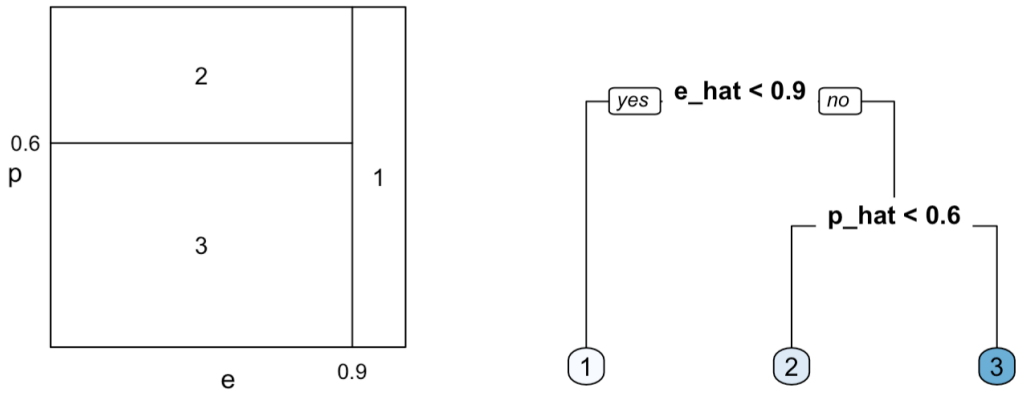

Figure 3.1: *Left:* a hypothetical partition over the 2d space of propensity and prognostic scores with the true values of piecewise constant treatment effects; *Right:* a sample regression tree T constructed in Step 3.

To perform such stratification, we grow a regression tree on $\tilde{Y}$, denoted by T, and the regressors are the estimated propensity scores $\hat{e}$ and the estimated prognostic scores $\hat{p}$ from Step 1. We follow the very general rule of binary recursive partitioning to build the tree T: allocate the data into the first two branches, using every possible binary split on every covariate; select the split that minimizes Gini impurity, continue the optimal splits over each branch along the covariate's values until the minimum node size is reached. To avoid overfitting, we set the minimum node size as 20 in our model. Choosing other criteria such as information gain instead of Gini impurity is another option for splitting criteria. A 10-fold cross validation is also performed at meantime to prune the large tree T for deciding the value of cost complexity. Cost complexity is the minimum improvement in the model needed at each node. The pruning rule follows that if one split does not improve the overall error of the model by the chosen cost complexity, then that split is decreed to be not worth pursuing (see more details in Section 9.2 of [HTF01]).

The final tree T (see the right plot of Figure 3.1) contains a few terminal nodes, and these are the predicted treatment effects for all units in the data. The values exactly represent a piecewise constant stratification over the 2d space of propensity and prognostic scores.

Estimation on a new unit

After we obtain the regression tree model T in Step 3, we can now estimate the value of the individual treatment effect corresponding to a new unit with covariate x_{new}.

We first compute the estimated propensity and prognostic scores for the new observation by

$$\hat{e}_{\text{new}} = \sigma(x_{\text{new}}^{\top}\hat{\alpha}),\ \hat{p}_{\text{new}} = x_{\text{new}}^{\top}\hat{\theta},$$

where $\hat{\alpha}$ and $\hat{\theta}$ are the solutions to Equations (3.5) and (3.6) respectively. Then with the estimated propensity score $\hat{e}_{\text{new}}$ and prognostic score $\hat{p}_{\text{new}}$, we can get an estimate of the treatment effect for this unit following the binary predictive rules in the tree T.

High-Dimensional Estimator

In a high-dimensional setting where the covariate dimension d is much larger than the sample size n, we can estimate the propensity and prognostic scores by adding a lasso ($l1$-based) penalty [Tib96] instead. This strategy was first proposed and named as "doubly robust matching estimators" (DRME) by [ACP18]. The corresponding optimization problems for the two scores can be written as

$$\hat{\alpha} = \underset{\alpha \in \mathbb{R}^d}{\arg\min} \; -\sum_{i=1}^{n} \Big[Y_i \log \sigma(X_i^\top \alpha) + (1 - Y_i) \log(1 - \sigma(X_i^\top \alpha)) \Big] + \lambda_1 \sum_{j=1}^{d} |\alpha_j|,$$

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^d}{\arg\min} \; \sum_{i: Z_i = 0} (Y_i - X_i^\top \theta)^2 + \lambda_2 \sum_{j=1}^{d} |\theta_j|.$$

The selection of the tuning parameter λ_1, λ_2 can be determined by any information criteria (AIC, BIC, and etc.). In practice, we use 10-fold cross-validation (CV) to select the value of λ. Then we perform the K-nearest-neighbor matching based on propensity and prognostic scores to get the estimates of individual treatment effects using Equation (3.7).

We extend the above estimator with our proposal of applying a regression tree on the estimated propensity and prognostic scores. The procedure for estimation of subgroup heterogeneous treatment effects and estimation on a new unit remains the same as Step 3 for low-dimensional set-ups.

Remark 3 *The choice of the number of nearest neighbors is a challenging problem. Updating distance metrics for every observation is computationally expensive, and choosing a value that is too small leads to a higher influence of noise on estimation. With regards to the application of nearest neighbor matching in causal inference, [AI06] derived large sample properties of matching estimators of average treatment effects with a fixed number of nearest neighbors, but the authors did not provide any details on how to select the exact number of neighbors. Conventional settings on the number of nearest neighbors in current literature is to set $K = 1$ (one-to-one matching; [Stu10, AS16]). However, [MR01b] suggested that in observational studies, substantially greater bias reduction is possible through matching with a variable number of controls rather than exact pair matching.*

In Appendix 3.7.1, we conduct a simulation study following one of the generative models from Section 3.5 to show how sensitive estimation accuracy is to the number of nearest neighbors selected and setting K to a large number other than 1 is more 'sensible' to reduce estimation bias. Although it is usually difficult to select a perfect value of K in practice, simply setting $K \approx \log(n)$ as suggested by [BCQ97] leads to reasonable results for a data sample of size n. Throughout all our experiments in the next section, setting K to the integer closest to the value $\log(n)$ provides estimates with high accuracy and does not require too much computational cost.

Computational Complexity

Our method is composed of three steps as introduced before. We first need to implement a logistic regression for estimating propensity score for a sample with size n and ambient dimension d. To solve a logistic regression optimization problem involves using the proximal Newton method, and its computational complexity is of # of iterations $\cdot\, O(nd)$ [YHL12]. In general, the algorithm converges after a small number of iterations, and we can safely assume it is of a constant order. As a result, the overall time complexity of solving a logistic regression is $O(nd)$. The estimation of prognostic scores also requires a computational cost of $O(nd)$, and for high-dimensional settings, the cost of solving the solution via coordinate descent remains at $O(nd)$ [FHT10] . The complexity of a K-nearest-neighbor matching based on the two estimated scores in the second step is of $O(Kn)$ [Lux07], and the selection of $K \approx \log(n)$ leads to a complexity of $O(n \log n)$. In the third step, we grow a regression tree based on two estimated scores, and it requires a computational complexity of $O(n \log n)$.

The overall computational complexity of our method depends on the comparison between the order of d and $\log(n)$. Our method attains a computational complexity of $O(nd)$ if the order of d is greater than that of $\log(n)$. Otherwise, the complexity becomes $O(n \log n)$.

3.5 Experiments

In this section, we will examine the performance of our proposed estimator (PP) in a variety of simulated and real data sets. The baseline estimators we compete against are leave-one-out endogenous stratification (ST; [ACW18]), causal forest (CF; [WA18]), single-score matching including propensity-score matching (PSM) and prognostic-score matching. Note that in the original research by [ACW18], the authors restricted their attention to randomized experiments, because this is the setting where endogenous stratification is typically used. However, they mentioned the possibility of applying the method on observational studies. We take this into consideration, and make their method as one of our competitors.

We implement our methods in R, using the packages "MatchIt" for K-nearest-neighbor Mahalanobis distance matching and "caret" for growing a non-parametric regression tree with cross validation over complexity parameter. Throughout, we set the number of nearest neighbors, K, to be the closest integer to $\log(n)$, where n is the sample size. For causal forest, we directly use the R package "grf" developed by [ATW19], following with an automatic cross-validation to select values of hyper-parameters [AW19] and a K-fold cross-fitting under a conventional setting of $K = 10$ recommended by [NW21]. Software that replicate all the simulations is available on the authors' Github page.

We evaluate the performance of each method according to two aspects, accuracy and uncertainty quantification. The results for single-score matching algorithms are not reported in this chapter because of very poor performance throughout all scenarios.

3.5.1 Simulated Data

We first examine on the following simulated data sets under six different data generation mechanisms. We get insights from the simulation study in [LS14] for the models considered in Scenarios 1-3. The propensity score and outcome (prognosis) models in Scenarios 1 and 3 are characterized by additivity and linearity (main effects only), but with different piecewise constant structures in the true treatment effects over a 2d grid of the two scores. We add

non-additivity and non-linear terms to both propensity and prognosis models in Scenarios 2. In other words, both propensity and prognostic scores are expected to be misspecified in these two models if we apply generalized linear models directly in estimation. Scenario 4 comes from [ACW18], with a constant treatment effect over all observations. Scenario 5 is modified from [WA18] (see Equation 27 there), in which the propensity model follows a continuous distribution instead of a linear structure. In addition, the true treatment effect is not a function of propensity and prognostic scores. In Scenario 6, we break the mold assumption of nicely delineated subgroups in treatment effects over 2d space of the two scores, and use a smooth function instead. A high-dimensional setting ($d >> n$) is examined in Scenario 7, where the generative model inherits from Scenario 1. We also include a table summarizing all simulation scenarios in Appendix 3.7.2 to help the audience interpret the result of the simulation study more readily.

We first introduce some notations used in the experiments: the sample size n, the ambient dimension d, as well as the following functions:

$$\begin{aligned}
&\text{true treatment effect: } \tau^*(X) = \mathbb{E}\left[Y(1) - Y(0)\middle|X\right], \\
&\text{treatment propensity: } e(x) = \mathbb{P}(Z = 1|X = x), \\
&\text{treatment logit: } \operatorname{logit}(e(x)) = \log\left(\frac{e(x)}{1 - e(x)}\right).
\end{aligned}$$

Throughout all the models we consider, we maintain the unconfoundedness assumption discussed in Section 3.3, generate the covariate X following a certain distribution, and entail homoscedastic Gaussian noise ϵ.

We evaluate the accuracy of an estimator $\tau(X)$ by the mean-squared error for estimating $\tau^*(X)$ at a random example X, with $m = n/10$, defined by

$$\operatorname{MSE}(\hat{\tau}(X)) := \frac{1}{m}\sum_{i=1}^{m}\left[\hat{\tau}_i(X) - \tau_i^*(X)\right]^2.$$

We record the averaged MSE over 1000 Monte Carlo trials for each scenario. For each Monte Carlo trial, we use train-test split to evaluate model performance. In detail, we first obtain n training sample and train models on this set. We then generate a separate testing sample under the same data generation mechanism with sample size $n/10$. We make predictions on the

test data using the trained models and compute the MSEs from different models accordingly. In terms of uncertainty quantification, we measure the coverage probability of $\tau(X)$ with a target coverage rate of 0.95. For endogenous stratification and our proposed method, we use non-parametric bootstrap with replacement to construct the empirical quantiles for each unit. The details on the implementation of non-parametric bootstrap methods are presented in Appendix 3.7.3. For causal forest, we construct 95% confidence intervals by estimating the standard errors of estimation using the "grf" package. We do not include the results for Scenario 7, a high-dimensional setting, because the asymptotic normality guarantees of non-parametric bootstrap and random forests are not valid in high dimensions [WA18, KP18].

Scenario 1. With $d \in \{2, 10, 50\}$, $n \in \{1000, 5000\}$, for $i = 1, ..., n$, we generate the data as follows:

$$
\begin{aligned}
&Y_i = p(X_i) + Z_i \cdot \tau_i^* + \epsilon_i, \\
&\tau_i^* = \mathbf{1}_{\{e(X_i)<0.6, p(X_i)<0\}}, \\
&\text{logit}(e(X_i)) = X_i^\top \beta^e, \\
&p(X_i) = X_i^\top \beta^p, \\
&X_i \overset{i.i.d.}{\sim} \mathcal{U}[0,1]^d, \\
&\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1),
\end{aligned}
$$

where β^e and β^p are randomly generated from $\mathcal{U}[-1, 1]^d$ in each iteration.

Scenario 2. We now add some interaction and non-linear terms to the propensity and prognostic models in Scenario 1, while keeping the set-ups of the covariate X, the response Y, the true treatment effect τ^* and the error term ϵ unchanged. We set $d = 10$ and $n = 5000$ in this case.

$$
\begin{aligned}
\text{logit}(e(X_i)) = X_i^\top \beta^e &+ 0.5 X_{i1} X_{i3} + 0.7 X_{i2} X_{i4} + 0.5 X_{i3} X_{i5} \\
&+ 0.7 X_{i4} X_{i6} + 0.5 X_{i5} X_{i7} + 0.5 X_{i1} X_{i6}
\end{aligned}
$$

$$
\begin{aligned}
&+ 0.7X_{i2}X_{i3} + 0.5X_{i3}X_{i4} + 0.5X_{i4}X_{i5} \\
&+ 0.5X_{i5}X_{i6} + X_{i2}^2 + X_{i4}^2 - X_{i7}^2 \\
p(X_i) = X_i^\top \beta^p &+ 0.5X_{i1}X_{i3} + 0.7X_{i2}X_{i4} + 0.5X_{i3}X_{i8}, \\
&+ 0.7X_{i4}X_{i9} + 0.5X_{i8}X_{i10} + 0.5X_{i1}X_{i9} \\
&+ 0.7X_{i2}X_{i3} + 0.5X_{i3}X_{i4} + 0.5X_{i4}X_{i8} \\
&+ 0.5X_{i8}X_{i9} + X_{i3}^2 + X_{i4}^2 - X_{i10}^2.
\end{aligned}
$$

Again for each simulation, β^e and β^p are randomly generated from $\mathcal{U}[-1,1]^d$.

Scenario 3. In this case, we define the true treatment effect with a more complicated piecewise constant structure over the 2d grid, under the same model used in Scenario 1, with $d = 10$ and $n = 5000$:

$$
\tau_i^* = \begin{cases} 0 & \text{if } e(X_i) \leq 0.6, p(X_i) \leq 0, \\ 1 & \text{if } e(X_i) \leq 0.6, p(X_i) > 0 \text{ or } e(X_i) > 0.6, p(X_i) \leq 0, \\ 2 & \text{if } e(X_i) > 0.6, p(X_i) > 0. \end{cases}
$$

Scenario 4. Setting $d = 10$ and $n = 5000$, the data is generated as:

$$
\begin{aligned}
Y_i &= 1 + \beta^\top X_i + \epsilon_i, \\
X_i &\overset{i.i.d.}{\sim} \mathcal{N}(0, \mathbf{I}_{d\times d}), \\
\epsilon_i &\overset{i.i.d.}{\sim} \mathcal{N}(0, 100 - d),
\end{aligned}
$$

where $\beta = (1, ..., 1)^\top \in \mathbb{R}^d$. Moreover, the treatment indicators for the simulations are such that $\sum_i Z_i = \lceil n/2 \rceil$. By construction, the vector of treatment effects satisfies $\tau^* = 0$.

Scenario 5. The data satisfies

$$
\begin{aligned}
\tau_i^* &= (X_{i1} + X_{i2} + ... + X_{i10})/10, \\
Y_i &= 2X_i^\top \mathbf{e}_1 - 1 + Z_i \cdot \tau_i^* + \epsilon_i,
\end{aligned}
$$

$$
\begin{aligned}
Z_i &\sim \text{Binom}(1, e(X_i)),\\
X_i &\overset{i.i.d.}{\sim} \mathcal{U}[0,1]^d,\\
e(X_i) &= \frac{1}{4}[1 + \beta_{2,4}(X_i^\top \mathbf{e}_1)],\\
\epsilon_i &\overset{i.i.d.}{\sim} \mathcal{N}(0,1),
\end{aligned}
$$

where $\mathbf{e}_1 = (1, 0, ..., 0)^\top$. We compare the results of different methods under the setting: $d = 10, n = 5000$.

Scenario 6. For this scenario, we examine the performance when the treatment effect takes on a smooth interval of propensity and prognostic scores. We use the same model in Scenario 1 for generating the two scores with $d = 10$ and $n = 5000$, and we inherit and modify the generative function (28) from [WA18] for treatment effects as follows:

$$
\tau_i^* = \zeta(e(X_i)) \cdot \zeta(p(X_i)),\ \zeta(x) = 1 + \frac{1}{1 + e^{-20(x-1/3)}}.
$$

Scenario 7. In the last case, we study the performance of different estimators on a high-dimensional data. The data model follows

$$
\begin{aligned}
&X_i \overset{i.i.d.}{\sim} \mathcal{U}[0,1]^d,\\
&Y_i = p(X_i) + Z_i \cdot \tau_i^* + \epsilon_i,\\
&\tau_i^* = \mathbf{1}_{\{e(X_i)<0.6, p(X_i)<0\}},\\
&\text{logit}(e(X_i)) = 0.4X_{i1} + 0.9X_{i2} - 0.4X_{i3} - 0.7X_{i4} - 0.3X_{i5} + 0.6X_{i6},\\
&p(X_i) = 0.9X_{i1} - 0.9X_{i2} + 0.2X_{i3} - 0.2X_{i4} + 0.9X_{i5} - 0.9X_{i6},\\
&\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1).
\end{aligned}
$$

We select $n = 3000$ and $d = 5000$ for analysis.

The boxplots that depict the distribution of MSEs obtained under all scenarios are presented in Figure 3.2. We can see that for Scenario 1, our proposed estimator achieves better

accuracy when the sample size n is large, and it is the best among the three estimators in these cases. The good performance of our method is consistent when we assume a more complex partition on the defined 2d grid as in Scenario 3. In addition, variation in accuracy, measured by the difference between the upper quartiles and the lower quartiles (also referred as the interquartile ranges) in each boxplot becomes smaller, accompanied with the increase in n. The reason for the enhanced performance when applying our method on a large-size data is due to the nature of nearest neighbor algorithm, which benefits more in learning the boundaries of clusters accurately with a larger sample size. Moreover, causal random forest suffers more in accuracy especially as d gets larger [WA18], while our method does not impact significantly because it focuses on a lower-dimensional space of the two scores instead of the overall covariate space. In Scenario 2, we introduce non-additivity and non-linear terms into the data model. Although linear assumptions are violated for both propensity and prognostic models, our method performs better compared to the other two methods regarding accuracy and variability. For a potential outcome model with randomized assignment of treatment and constant treatment effects, as in Scenario 4, our method has the best accuracy among all candidates, even though large noise is added to the true signal. In Scenario 5, our estimator performs slightly worse than causal forests in accuracy since the true treatment function is not directly dependent on propensity and prognostic scores, but it still outperforms endogenous stratification. However, for the previous two cases, our method shows larger variation in accuracy than the others. One of the reasons for this phenomenon is due to the misspecification in estimating the true scores and the large variance in constructing tree models. When the assumption of a sharp stratification of treatment effects in the 2d space of the two scores is invalid, as in Scenario 6, our method maintains its superiority over the benchmarks. With subgroups identified over the 2d interval of the scores, the mean squared errors are reduced through our estimator comparably better than the other methods. In a high-dimensional setting such as Scenario 7, we consider modified methods with lasso-regularized regressions for both our methodology and endogenous stratification, and our method remains its high performance as in the low-dimensional set-ups

since the piecewise constant structure in treatment effects in the setting does not change.

We now take a careful look at the visualization comparison between the true treatment effect and the predictions obtained from our method for Scenarios 1, 2, 3, 6 and 7. We confine both the true signal and the predictive model in a 2d grid scaled on the true propensity and prognostic scores, as shown in Figure 3.3. It is not surprising that our proposed estimators provide a descent recovery of the piecewise constant partition in the true treatment effects over the 2d grid as in Scenarios 1, 2, 3, and 7, with only a small difference in the magnitude of treatment effects. In the setting when treatment effects are smoothly distributed across the grid of the two scores as in Scenario 6, our estimation does not exactly capture the soft boundary of the shape, but still provides us with a simplified delineation that help us understand general heterogeneity in treatment effects over the 2d space.

With regard to uncertainty quantification, we examine coverage rates with a target confidence level of 0.95 for each method under different scenarios, and the corresponding results, along with the within iteration standard error estimates, are recorded in Table 3.1. It is quite clear that our proposed method achieves nominal coverage over the other two methods in almost all scenarios. It is worth to notice that although our estimator achieves a better coverage rate than the benchmarks, it suffers from large variation due to the nature of nearest-neighbor matching, especially in the case with small sample size. In general, our method that incorporates both propensity and prognostic scores in identifying subgroups and estimating treatment effects show its advantage in reducing bias, but the resulting large variance is worrisome and it remains space for future study to solve this issue.

We also compare the computational cost of our proposed scored-based method and causal forest, with simulated data from Scenario 1. Endogenous stratification is not considered in this comparison because the method exerts an arbitrary subclassification into a predetermined number of subgroups and it does not involve a cross-validation process. We record the averaged time consumed to obtain the estimate over 1000 simulations for both methods. Figure 4 demonstrates that our proposed method (PP) is comparably time-efficient over its competitor, and causal forest (CF) can be very expensive in operational time for large-size

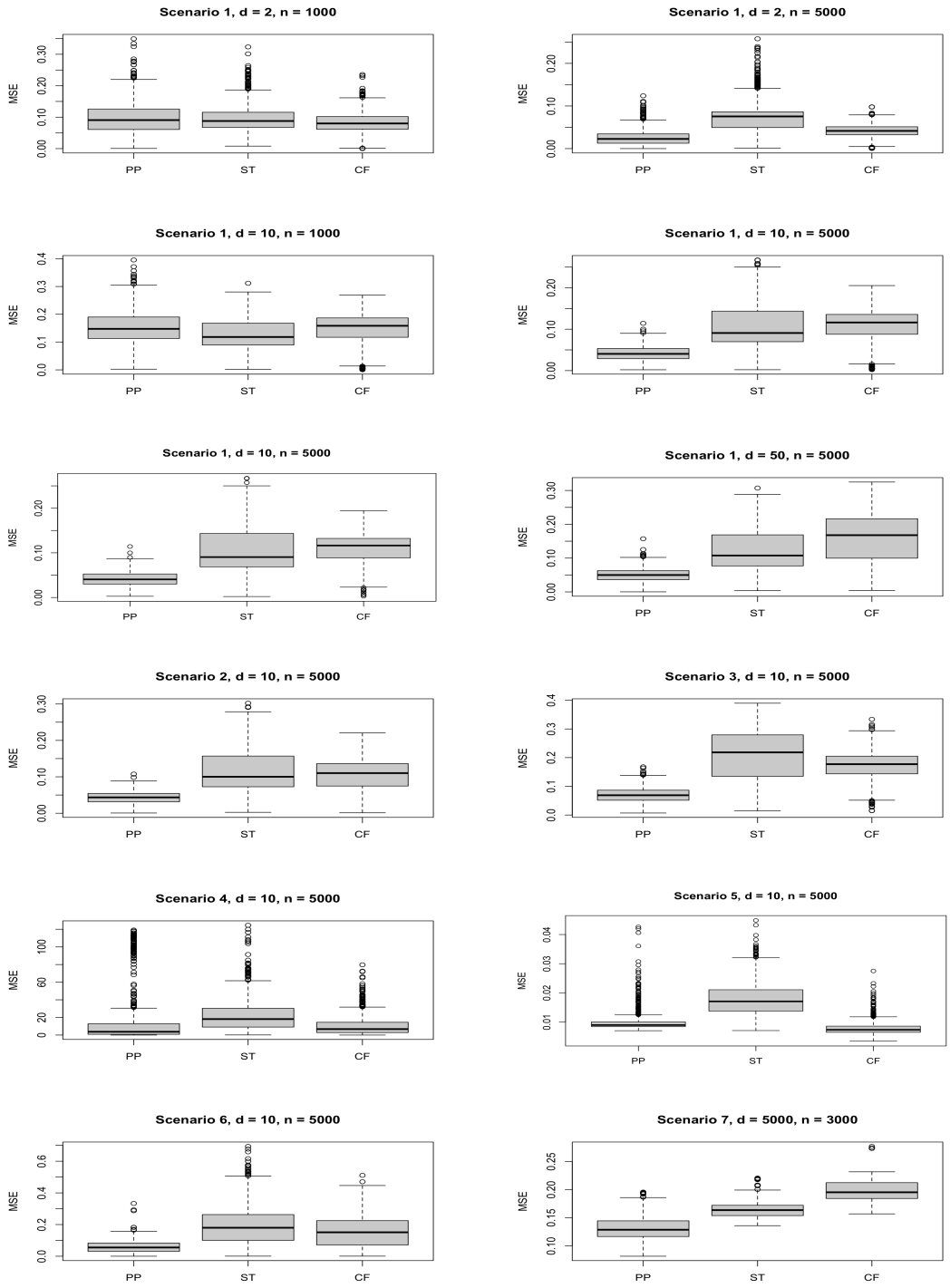

Figure 3.2: Comparison of mean squared errors over 1000 Monte Carlo simulations under different generative models for our proposed method (PP), endogenous stratification (ST) and causal forest (CF).

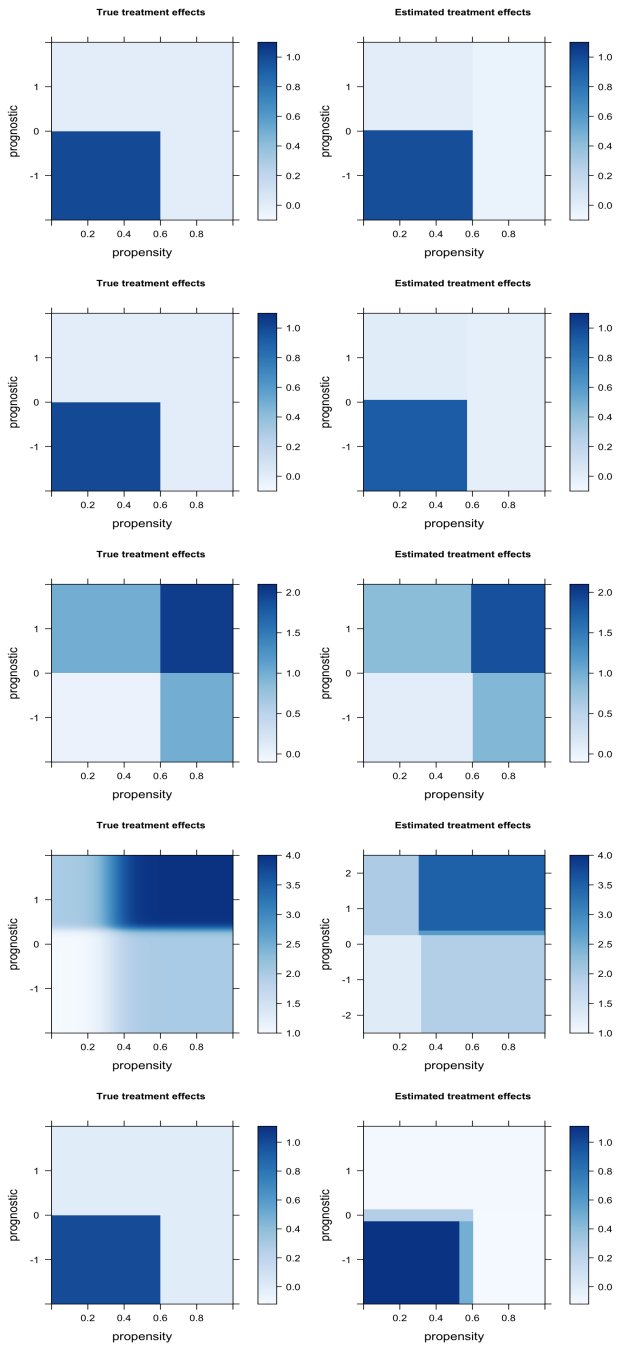

Figure 3.3: One instance comparison between the true treatment effects and the estimates from the score-based method for Scenarios 1, 2, 3, 6 and 7 (from the top to the bottom). Plots on the left column depict the true treatment effect on the 2d grid of propensity and prognostic scores, and the right plots demonstrate the estimated treatment effects from our proposed method over the same grid.

Table 3.1: Reported coverage rates with a target confidence level of 0.95 and within iteration standard error estimates (in parentheses).

Scenario	n	d	PP	ST	CF
1	1000	2	0.996 (0.177)	0.748 (0.158)	0.969 (0.128)
	5000	10	0.987 (0.116)	0.731 (0.062)	0.736 (0.061)
2	1000	10	0.994 (0.230)	0.753 (0.152)	0.797 (0.157)
	5000	10	0.956 (0.141)	0.619 (0.064)	0.829 (0.224)
3	1000	10	0.989 (0.382)	0.420 (0.163)	0.769 (0.297)
	5000	10	0.949 (0.164)	0.307 (0.075)	0.750 (0.252)
4	1000	10	1.000 (28.621)	0.933 (16.578)	0.993 (18.067)
	5000	10	1.000 (6.693)	0.998(6.049)	0.973 (5.957)
5	1000	10	1.000 (0.524)	0.950 (0.353)	0.978 (0.429)
	5000	10	0.845 (0.336)	0.622 (0.287)	0.716 (0.321)
6	1000	10	0.995 (0.472)	0.495 (0.161)	0.865 (0.474)
	5000	10	0.999 (0.179)	0.578 (0.089)	0.821 (0.259)

problems.

These results highlight the promise of our method for accurate estimation of subgroups in heterogeneous treatment effects, all while emphasizing avenues for further work. An immediate challenge is to control the bias in estimation of propensity and prognostic scores. Using more powerful models instead of simple linear regression in estimation is a good way to reduce bias by enabling the trees to focus more closely on the coordinates with the greatest signal. The study of splitting rules for trees designed to estimate causal effects is still in its infancy and improvements may be possible.

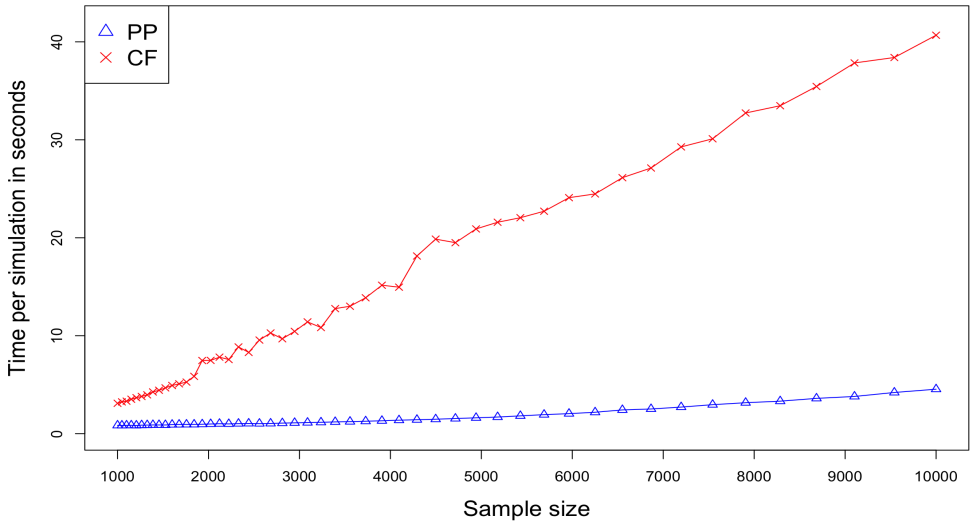

Figure 3.4: A plot of time per simulation in seconds for our proposed method (PP) and causal forest (CF) against problem size n (50 values from 1000 up to 10000). For each method, the time to compute the estimate for one simulated data is averaged over 1000 Monte Carlo simulations.

3.5.2 Real Data Analysis

To illustrate the behavior of our estimator, we apply our method on the two real-world data sets, one from a clinical study and the other from a complex social survey. Propensity score based methods are frequently used for confounding adjustment in observational studies, where baseline characteristics can affect the outcome of policy interventions. Therefore, the results from our method are expected to provide meaningful implications for these real data sets. However, one potential limitation of our proposed tree-based approach is that it imposes arbitrary dichotomization with sharp thresholds and thus may obscure potential continuous gradient over the space. For real data analysis in general, our method provides a reasonable simplification that is easy to interpret for the audience and gives some framework for thinking about when such a simplification would be appropriate.

3.5.2.1 Right Heart Catheterization Analysis

While randomized control trials (RCT) are widely encouraged as the ideal methodology for causal inference in clinical and medical research, the lack of randomized data due to high costs and potential high risks leads to the studies based on observational data. In this section, we are interested in examining the association between the use of right heart catheterization (RHC) during the first 24 hours of care in the intensive care unit (ICU) and the short-term survival conditions of the patients. Right Heart Catheterization (RHC) is a procedure for directly measuring how well the heart is pumping blood to the lungs. RHC is often applied to critically ill patients for directing immediate and subsequent treatment. However, RHC imposes a small risk of causing serious complications when administering the procedure. Therefore, the use of RHC is controversial among practitioners, and scientists want to statistically validate the causal effects of RHC treatments. The causal study using observational data can be dated back to [CSD96], where the authors implemented propensity score matching and concluded that RHC treatment lead to lower survival than not performing the treatment. Later, [HI01] proposed a more efficient propensity-score based method and the recent study by [LV21] using a modified propensity score model suggested RHC significantly affected mortality rate in a short-term period.

A dataset for analysis was first used in [CSD96], and it is suitable for the purpose of applying our method because of its its extremely well-balanced distribution of confounders across levels of the treatment [SMR21]. The treatment variable Z in the data indicates whether or not a patient received a RHC within 24 hours of admission. The binary outcome Y is defined based on whether a patient died at any time up to 180 days since admission. The original data consisted of 5735 participants with 73 covariates. We preprocess the full data in the way suggested in [HI01] and [LV21], by removing all observations that contain null values in covariates, dropping the singular covariate in the reduced data, and encoding categorical variables into dummy variables. The resulted data contains 2707 observations and 72 covariates, with 1103 in the treated group ($Z = 1$) and 1604 in the control group ($Z = 0$). Among the 72 observed covariates, there are 21 continuous, 25 binary, and 26

dummy variables transformed from the original 6 categorical variables.

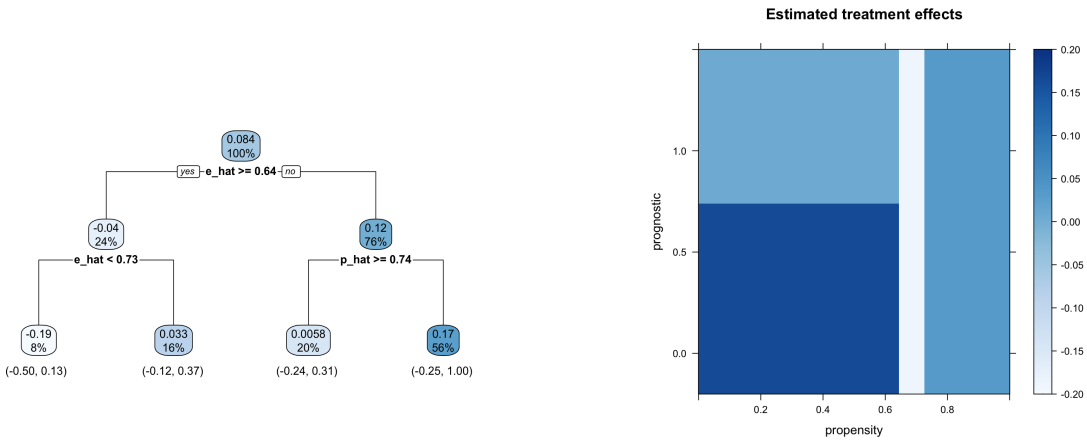

Figure 3.5: Prediction model for Right Heart Catheterization (RHC) data. *Left:* a tree diagram of predictions on each split based on the two scores, with 95 % confidence intervals provided on the bottom. *Right:* a 2d plot of prediction stratification on the intervals of propensity and prognostic scores, with a color scale on magnitude of treatment effects.

The result of the prediction model from our proposed method is reported in Figure 3.5 below. We observe that the sign of estimated treatment effects varies depending on the value of the propensity score and prognostic score. This particular pattern implies that RHC procedures indeed offer both benefits and risks in affecting patients' short-term survival conditions. Specifically, we are interested in the occurrence of large positive treatment effects (increase in chance of death) from the estimation. An estimated treatment effect of 0.17 (with a 95% confidence interval between -0.25 and 1.00) is observed on the group of patients with propensity scores less than 0.64 and prognostic scores less than 0.74, and this group accounts for 56% of the entire sample. For the rest of the sample, we observe relatively small treatment effects in magnitude. Under the scenario of RHC data, a smaller propensity score means that the patient is less likely to receive RHC procedures after admitting to the ICU, and it is related to the availability of RHC procedures at the hospital to which the patient is admitted. A smaller prognostic score tells that the patient has lower underlying chance of death. One possible explanation for this significant positive treatment on this certain

group is that drastic change in treatment procedures that were applied to patients who do not actually need the aggressive style of care largely undermine patients' health conditions after admission and increase the mortality rate. In summary, our findings generally agree with the results and explanations in [CSD96] and they offer some insights for practitioners to decide whether they should apply RHC procedures to patients.

3.5.2.2 National Medical Expenditure Survey

For the next experiment, we analyze a complex social survey data. In many complex surveys, data are not usually well-balanced due to potential biased sampling procedure. To incorporate score-based methods with complex survey data requires an appropriate estimation on propensity and prognostic scores. [DSS14] suggested that combining a propensity score method and survey weighting is necessary to achieve unbiased treatment effect estimates that are generalizable to the original survey target population. [AJC18] conducted numerical experiments and showed that greater balance in measured baseline covariates and decreased bias is observed when natural retained weights are used in propensity score matching. Therefore, we include sampling weight as an baseline covariate when estimating propensity and prognostic scores in our analysis.

In this study, we aim to answer the research question: how smoking habit affects medical expenditure over lifetime, and we use the same data set as in [JDG03], which is originally extracted from the 1987 National Medical Expenditure Survey (NMES). The NMES included detailed information about frequency and duration of smoking with a large nationally representative data base of nearly 30,000 adults, and that 1987 medical costs are verified by multiple interviews and additional data from clinicians and hospitals. A large amount of literature focus on applying various statistical methods to analyze the causal effects of smoking on medical expenditures using the NMES data. In the original study by [JDG03], the authors first estimated the effects of smoking on certain diseases and then examined how much those diseases increased medical costs. In contrast, [Rub01], [ID04] and [ZDI20] proposed to directly estimate the effects of smoking on medical expenditures using propensity-score

based matching and subclassification. [HMC20] applied Bayesian regression tree models to assess heterogeneous treatment effects.

For our analysis, we explore the effects of extensive exposure to cigarettes on medical expenditures, and we use pack-year as a measurement of cigarette measurement, the same as in [ID04] and [HMC20]. Pack-year is a clinical quantification of cigarette smoking used to measure a person's exposure to tobacco, defined by

$$\text{pack-year} = \frac{\text{number of cigarettes per day}}{20} \times \text{number of years smoked.}$$

Following that, we determine the treatment indicator Z by the question whether the observation has a heavy lifetime smoking habit, which we define to be greater than 17 pack-years, the equivalent of 17 years of pack-a-day smoking.

The subject-level covariates X in our analysis include age at the times of the survey (between 19 and 94), age when the individual started smoking, gender (male, female), race (white, black, other), marriage status (married, widowed, divorced, separated, never married), education level (college graduate, some college, high school graduate, other), census region (Northeast, Midwest, South, West), poverty status (poor, near poor, low income, middle income, high income), seat belt usage (rarely, sometimes, always/almost always), and sample weight. We select the natural logarithm of annual medical expenditures as the outcome variable Y to maintain the assumption of heteroscedasticity in random errors. We preprocess the raw data set by omitting any observations with missing values in the covariates and excluding those who had zero medical expenditure. The resulting restricted data set contains 7903 individuals, with 4014 in the treated group ($Z = 1$) and 3889 in the controlled group ($Z = 0$).

The prediction model obtained from our method, as shown in Figure 3.6, is simple and easy to interpret. We derive a positive treatment effect across the entire sample, and the effect becomes significant when the predicted potential outcome is relatively low (less than 5.7). These results indicate that more reliance on smoking will deteriorate one's health condition, especially for those who currently do not have a large amount of medical expenditure.

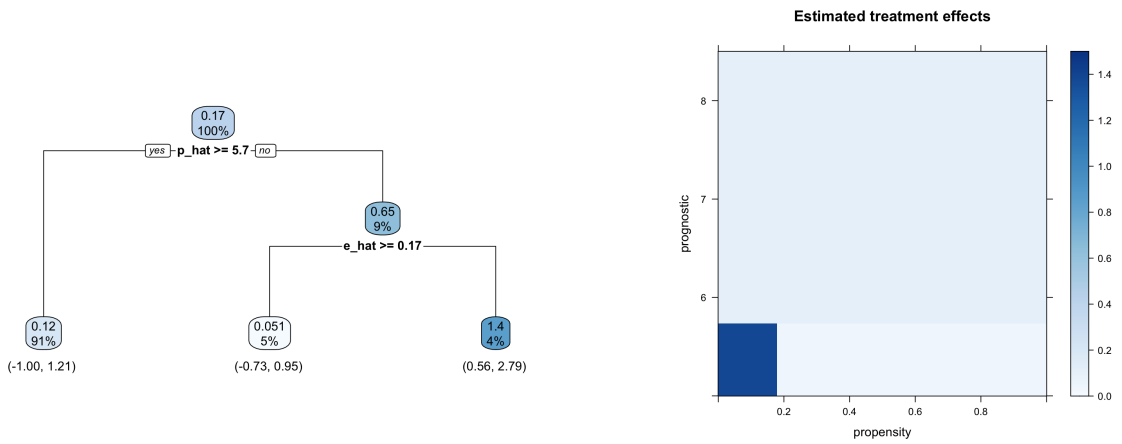

Figure 3.6: Prediction model for National Medical Expenditure Survey (NMES) data. *Left:* a tree diagram of predictions on each split based on the two scores, with 95 % confidence intervals provided on the bottom. *Right:* a 2d plot of prediction stratification on the intervals of propensity and prognostic scores, with a color scale on magnitude of treatment effects.

Moreover, we observe a significant positive treatment effect of 1.4 (with a 95% confidence interval between 0.56 and 2.79), in other word, a certain and substantial increase in medical expenditure for the subgroup with propensity score less than 0.17. It is intuitive to assume that a smaller possibility of engaging in excessive tobacco exposure is associated with healthier living styles. This phenomenon is another evidence that individuals who are more likely to stay healthy may suffer more from excessive exposure to tobacco products. In all, these results support policymakers and social activists who advocate for nationwide smoking ban.

3.6 Discussions

Our method is different from existing methods on estimating heterogeneous treatment effects in a way that we incorporate both matching algorithms and non-parametric regression trees in estimation, and the final estimate can be regarded as a 2d summary on treatment effects. Moreover, our method exercises a simultaneous stratification across the entire population

into subgroups with the same treatment effects. Subgroup treatment effect analysis is an important but challenging research topic in observational studies, and our method can be served as an efficient tool to reach a reasonable partition.

Our numerical experiments on various simulated and real-life data lay out empirical evidence of the superiority of our estimator over state-of-the-art methods in accuracy and our proposal provides insightful interpretation on heterogeneity in treatment effects across different subgroups. We also discovered that our method is powerful in investigating subpopulations with significant treatment effects. Identifying representative subpopulations that receive extreme results after treatment is a paramount task in many practical contexts. Through empirical experiments on two real-world data sets from observational studies, our method demonstrates its ability in identifying these significant effects.

Although our method shows its outstanding performance in estimating treatments effects under the piecewise constant structure assumption, it remains meaningful and requires further study to develop more accurate recovery of such structure. For example, a potential shortcoming of using conventional regression trees for subclassification is that the binary partition over the true signals is not necessarily unique. Using some variants of CART, like optimal trees [BD17] and dyadic regression trees [Don97], would be more appropriate for estimation under additional assumptions. Another particular concern with tree-based methods is the curse of dimensionality. Applying other non-parametric regression techniques, such as K-nearest-neighbor fused lasso [PSC20] and locally estimated scatterplot smoothing (LOESS), may offer a compromised solution to deal with this issue and learn a more smoothed structure in treatment effects other than a rectangular partition on 2d data. It is also worth improving the estimation of propensity and prognostic scores using similar non-parametric based methods if a piecewise constant assumption hold for the two scores as well.

Moreover, we acknowledge that fitting propensity score and prognostic scores normally would demand more than main terms linear and logistic regressions as used in the current method to have hope of eliminating bias in estimating the two scores. The key takeaway

of our proposal is not to estimate the two scores more accurately, but to provide a lower-dimensional summary on heterogeneous treatment effects. If more foreknowledge on the structure of the underlying functions of the two scores can be incorporated, we can certainly select a more ideal algorithm or machine learning model for estimation. To compromise the lack of such information under a usual scenario, a SuperLearner ensemble estimator, proposed by [LPH07], may offer a more promising solution for estimating the two scores by taking a weighted average of multiple prediction models. Yet, an important challenge for future work is to design rules that can automatically choose the best model to fit the data.

Another question to consider is the source of heterogeneity in causal effects. Our proposal that utilizes two scores can be seen as a reasonable starting point to analyze heterogeneity from a low-dimensional perspective, but it cannot comprehensively capture the sources of such effects. For instance, it is seemingly possible to exist a covariate that leads to substantial heterogeneous treatment effects on the response but does not itself have a large effect on propensity or prognostic scores. There is no particular guiding theories about this issue in current literature, and thus it remains huge space for further investigation.

3.7 Appendix

3.7.1 Study on the Choice of the Number of Nearest Neighbors

In this section, we examine how the number of nearest neighbors in the matching algorithm affects the estimation accuracy. Recall in Step 2 of our proposed method, we implement a K-nearest-neighbor algorithm based on the two estimated scores for a sample of size n. The computational complexity of this K-NN algorithm is of $O(Kn)$. Although a larger K produce a promisingly higher estimation accuracy, more computational costs become the corresponding side-effect. Moreover, the issue of "underfitting" - the increase of bias occurs when we make such selection. It is because a selection of too large K defeats the underlying principle behind nearest neighbor algorithms - that neighbours that are nearer share similar densities. Therefore, a smart choice of K is essential to balance the trade-off among accuracy,

computational expense, and generability.

We follow the same generative model in Scenario 1 from Section 3.5 and compute the averaged mean squared error over 1000 Monte Carlo simulations for $K = 1, ..., 50$ with a fixed sample size $n = 5000$. The results in Figure 3.6 show that the averaged MSE continuously decreases as the number of nearest neighbors K selected in the matching algorithm grows. However, the speed of improvement in accuracy gradually slows down when K exceeds 10, which is close to log(5000). This suggests that an empirical choice of $K \approx \log(n)$ is sufficient to produce a reasonable estimate on the target parameter and this choice is more 'sensible' than the conventional setting of $K = 1$.

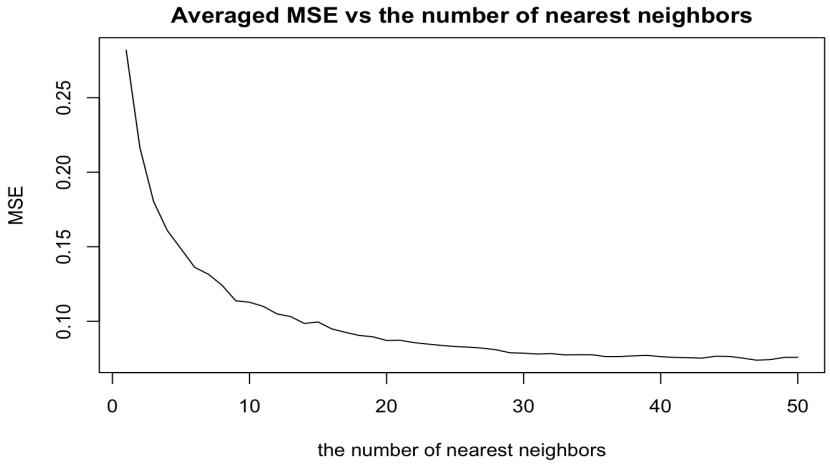

Figure 3.7: The plot of averaged MSE against the number of nearest neighbors.

3.7.2 A Summary Table of Simulation Studies

In this appendix, we include a table to provide a text summary of all simulation models considered in Section 5.1. We add some notes to help the audience understand similarities, differences, and sources of these data generative models.

Scenario	Propensity Model	Prognostic Model	Treatment Effect Model	Note
1	linearity in covariates	linearity in covariates	a sharp delineation over the 2d grid of propensity and prognostic scores with 2 subgroups	
2	linear model with additional interactions and non-linear terms	linear model with additional interactions and non-linear terms	a sharp delineation over the 2d grid of propensity and prognostic scores with 2 subgroups	the same treatment effect model as in Scenario 1; both propensity and prognostic scores are misspecified with linear model
3	linearity in covariates	linearity in covariates	a sharp delineation over the 2d grid of propensity and prognostic scores with 4 subgroups	the same propensity and prognostic models as in Scenario 1; more subgroups in treatment effects
4	random with a fixed number of treated units	linearity in covariates with an intercept	constant zero	the same setting from [ACW18]
5	a smooth function of single covariate	proportional to single covariate	a function of the covariates	modified from Equation 27 of [WA18]
6	linearity in covariates	linearity in covariates	a smooth function of the two scores	modified from Equation 28 of [WA18]
7	linearity in a few covariates with sparsity in others	linearity in a few covariates with sparsity in others	a sharp delineation over the 2d grid of propensity and prognostic scores with 2 subgroups	a high-dimensional setting corresponding to Scenario 1

3.7.3 Non-parametric Bootstrap in Simulation Studies

In Section 3.5, we use non-parametric bootstrap to construct confidence intervals for endogenous stratification and our proposed method. We use these bootstrap samples to compute coverage rates with respect to a target level of 95% as a measurement of uncertainty. The bootstrap method, introduced by [Efr79], is a simple but powerful tool to obtain a robust non-parametric estimate of the confidence intervals through sampling from the empirical distribution function of the observed data. It is worth noting that causal forests or generalized random forests can compute confidence intervals quickly without requiring a bootstrap approach on which our method relies.

In this appendix, we introduce the details on how we implement non-parametric bootstrap for the purpose of computing coverage rates in the simulation experiments. For each scenario in Section 3.5, we start with generating a sample following the defined data generation model with a sample size n. Next, we create 1000 random samples with replacement from this single set of data, also with the sample size n. We then apply both methods on these simulation repetitions, and obtain a series of estimations on each unit in the original set. Following these estimations, we calculate the corresponding 2.5% and 97.5% quantiles for all units in the original sample. Coverage rates of a 95% prediction level are thus the frequencies of the original units falling inside the intervals between the two quantiles computed in the previous step.

CHAPTER 4

Causal Effect Estimation via Fused Lasso over Graph

4.1 Introduction

Causal inference finds broad applications across diverse disciplines including economics [Var16, AI17], epidemiology [HHC88, HR20], medical science [Van15, OE16], and political science [Bla13, MW14], with growing focus on its pivotal role in criminal analysis for enhancing community safety and security [Ber87, CFM15]. By scrutinizing criminal activity patterns [MR09] and identifying hot spots [BB08], causal inference helps inform precise law enforcement strategies. Integrating causal inference methods enriches the quantification of causality, transcending mere correlations and revealing the root causes of crime. This deeper comprehension of crime dynamics empowers policymakers to implement effective, evidence-based interventions tailored to specific contexts.

One critical aspect of criminal analysis is that crime is inherently localized in nature [And11, BML22]. Crime exhibits varying rates across different regions, but it tends to concentrate in specific neighborhoods or areas. Effective intervention strategies necessitate a localized understanding of crime dynamics, enabling policymakers to optimize efforts and reduce crime more efficiently. This focus has aroused interest in estimating heterogeneous treatment effects [PBF22], where leveraging graph structures allows for a local perspective by assuming that interactive causality exists over adjacent nodes on a graph [WCA23]. However, there remains a research gap in applying more generalized graph structures to causal inference, particularly for estimating heterogeneous treatment effects.

To address this lack in current literature, this chapter proposes novel methodologies

for learning heterogeneous treatment effects from graph structures that capture localized patterns in treatment effects. Our proposed methodologies incorporate fused lasso, a locally adaptive function model [ROF92, MG97, TSR05, KKB09, Tib14, WSS16, PSS18]. This chapter highlights three major advantages of our proposed methodologies:

- Our methods show local adaptability by leveraging adjacency patterns derived from a graph structure. The methods are built on the simple assumption that treatment effects may vary across units, but they share a piecewise constant structure over the graph. This flexibility allows our methods to be applied to any general graphs, not limited to commonly used graphs such as chain graphs and grid graphs. Thereby, our approaches are more locally adaptive to capture nuanced relationships between observations.

- Our methods are characterized by their simplicity and efficiency in computation. The algorithms for computing our proposed estimators are based on the l_1-penalized optimization framework, which can be easily solved using existing algorithms. This makes our algorithms not only straightforward to implement but also highly practical for researchers and practitioners in related fields.

- We rigorously study the theoretical robustness of our methods, showing that, under mild assumptions on data, the estimators we propose achieve fast convergence rates for estimating treatment effects. The rate for the model when a graph is provided is consistent for estimation over general graph structures, and the rate for the model when a graph is not provided aligns with those of well-established statistical models when we construct and integrate a K-nearest-neighbor graph in the estimation process.

4.1.1 Related Literature

The application of causal inference in criminal analysis dates back to the study by [Ber87], which evaluated the impact of criminal justice programs. More recently, [MWE12] conducted

a comprehensive analysis of drug courts, employing causal inference methods to assess their effectiveness in reducing recidivism. Similarly, [AFS15] used randomized controlled trials to estimate the causal effect of police body-worn cameras on use of force incidents and citizen complaints. In the realm of hot spots policing, [BTP19] conducted a thorough review of intervention effects on crime reduction, applying meta-analysis methods to quantify their efficacy.

Estimating heterogeneous treatment effects is crucial for understanding causal effects in various applications, where advanced machine learning methods play an important role in achieving accurate estimations. For instance, [WA18] and [ATW19] introduced random forests for treatment effect estimation, while [PQJ18] applied boosting methods in high-dimensional data, and [SJS17] used neural networks for individual treatment effect estimation. Recent research, including [KSB19] and [SLO19], has focused on developing meta-learner frameworks, exploring diverse machine learning algorithms and data structures. For flexible and high-performance estimation, [FWW11] investigated doubly robust estimation methods for observational data, [OR21] reviewed empirical Bayes approaches, and [CCD18] pioneered double machine learning for estimating treatment effects and structural parameters. In recent innovations, [YCP23] proposed a two-dimensional score-based approach for heterogeneous treatment effect estimation. [MZL22] and [MZL23] developed methods integrating lasso and support vector machines for optimal group-structured treatment effect estimation.

Recent literature has explored the use of diverse graph structures in causal effect estimation. For example, [OSL20] employed chain graph models to infer treatment effects on social networks, [PBF22] utilized bipartite graphs based on geographic locations for randomization tests of causal effects, and [LW22] studied treatment effect estimation under network interference using random graph asymptotics. Another realm that intersects the estimation of treatment effects with graph-based models is the recent development of network community detection for causal inference. For instance, [FMW22] introduced a Bayesian generalized propensity score method for estimating causal effects under network interference. Addi-

tionally, [GD23] addressed regression-based analysis and design-based properties in causal inference for network experiments.

4.1.2 Outline

The remainder of the chapter is organized as follows. Section 4.2 introduces the statistical frameworks for estimating heterogeneous treatment effects via fused lasso over graph and compares the proposed models with related research in the current literature. In Section 4.3, we propose new estimators and the corresponding algorithms for computing numerical estimates. The theoretical properties of the proposed estimators are then studied in Section 4.4. Section 4.5 presents the results of numerical experiments conducted on multiple simulated datasets, and in Section 4.6, we present the analysis of a case study on U.S. transportation network crime data. In Section 4.7, we summarize our contributions and discuss some limitations and potential future directions for the proposed methodologies.

4.2 Preliminaries

4.2.1 Notation

Before presenting our main proposals, we introduce some notation that will be used consistently throughout this chapter. The Euclidean norm, denoted as $\|x\|_2$, for a vector $x \in \mathbb{R}^d$, is defined as $\|x\|_2 = \sqrt{x_1^2 + \ldots + x_d^2}$. The l_1 norm of x is represented as $\|x\|_1 = |x_1| + \ldots + |x_d|$, while the infinity norm is denoted by $\|x\|_\infty = \max_i |x_i|$. Within the covariate space $\mathcal{X}$, we consider the Borel sigma algebra, $\mathcal{B}(\mathcal{X})$, induced by the metric $d_{\mathcal{X}}$. We also introduce the measure μ defined on $\mathcal{B}(\mathcal{X})$. In this chapter, we assume that the covariates X have a dimension greater than one and follow an independent distribution, i.e., $X_i \overset{\text{ind}}{\sim} p(x)$. Here, p represents the probability density function associated with the distribution of X_i within the measure space $(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mu)$. Moreover, we employ the notation $a_n = O_{\mathbb{P}}(b_n)$ to denote that, for any $\epsilon > 0$, there exists a constant $C > 0$ such that $\mathbb{P}(a_n \geq C b_n) < \epsilon$ for all n. Sim-

ilarly, we use the notation $a_n = \Theta_{\mathbb{P}}(b_n)$ to indicate that both $a_n = O_{\mathbb{P}}(b_n)$ and $b_n = O_{\mathbb{P}}(a_n)$, meaning that a_n and b_n are of the same order in probability. For a matrix A, we denote its Moore–Penrose inverse by $A^{\dagger}$. For two matrices A and B of the same dimension $m \times n$, we denote the Hadamard product as $A \circ B$, with its elements calculated as $(A \circ B)_{ij} = (A)_{ij}(B)_{ij}$. We also write poly$(x)$ for a polynomial function of x, but the functions vary from case to case in the content below.

4.2.2 The Basic Model Conditioned on a Graph

We formalize our problem in terms of the potential outcomes framework proposed by [Spl23, Rub74]. Suppose that we have n independent and identically distributed measurements (Y_i, Z_i) for $i = 1, \ldots, n$, where $Y_i \in \mathbb{R}$ is the observed outcome, and $Z_i \in \{0, 1\}$ is the treatment assignment. We posit the existence of potential outcomes $\{Y_i(0), Y_i(1)\}$ corresponding to the outcome we would have observed given the treatment assignment $Z_i = 0$ or 1, respectively, such that $Y_i = Y_i(Z_i)$. We can write:

$$Y_i = Y_i(1) \cdot \mathbf{1}_{\{Z_i=1\}} + Y_i(0) \cdot \mathbf{1}_{\{Z_i=0\}}, \tag{4.1}$$

where $Y_i(0), Y_i(1) \in \mathbb{R}$ are random variables.

At the same time, we assume that there exists a graph, denoted as G, which encapsulates some local structures in the observed data. The graph G is characterized by a vertex set $V = \{1, \ldots, n\}$ corresponding to the observed data and an edge set E connecting data pairs (i, j) if an intrinsic relationship between i and j is assumed. For instance, in scenarios where data (Y_i, Z_i) for $i = 1, \ldots, n$ represent geographic locations on a map, the graph G may take a simple grid graph structure, linking neighboring data points on the map. In scenarios such as time-series or longitudinal data, where each (Y_i, Z_i) corresponds to a time point t_i, an edge of the graph G could connect the data pair (i, j) if there is temporal proximity between t_i and t_j. Another example is using a more general graph G to represent social networks, where edges join data points if there is interaction between individuals.

We seek to estimate the treatment effect function conditioned on the graph G

$$\tau_i^* := \mathbb{E}\left[Y_i(1) - Y_i(0) \mid G\right]. \tag{4.2}$$

The basic assumption of our model is that the true treatment effects τ_i^* can vary across units, but the graph G contains information about the local smoothness of the treatment effects. Specifically, we assume that treatment effects are piecewise constant along the graph. If data units i and j are connected as an edge of G, i.e., $(i, j) \in E$, then

$$\tau_i^* \approx \tau_j^* \quad \text{or} \quad \tau_i^* = \tau_j^*$$

for most edge pairs $(i, j) \in E$.

In order to identify the treatment effect function, we assume that the treatment assignment is randomized if we control for the graph G.

Assumption 9 *For $i = 1, \ldots, n$,*

$$Y_i(0), Y_i(1) \perp\!\!\!\perp Z_i \mid G.$$

The above assumption is modified from the unconfoundedness assumption in [RR83], which requires conditioning on a covariate set instead.

If we write the conditional response surfaces as $\mu_i(Z_i) := \mathbb{E}[Y_i \mid G, Z_i]$, and define $\epsilon_i(Z_i) := Y_i(Z_i) - [\mu_i(0) + Z_i\tau_i^*]$, with the shorthand $\epsilon_i := \epsilon_i(Z_i)$, the following lemma leads to the statistical model of interest.

Lemma 4.2.1 *Under Assumption 9, $\mathbb{E}[\epsilon_i \mid G, Z_i] = 0$ for $i = 1, \ldots, n$, and*

$$Y_i - m_i^* = (Z_i - e_i^*)\tau_i^* + \epsilon_i, \tag{4.3}$$

where $m_i^* = \mathbb{E}[Y_i \mid G]$ is the conditional mean outcome or the main effect, and $e_i^* = \mathbb{P}(Z_i = 1 \mid G)$ is the propensity score.

Proof. See Appendix 4.8.4.

The decomposition (4.3) was originated from [Rob88] in order to estimate parametric components in partially linear models, and recent literature in causal inference, including [ATW19, AW19, NW21], rely on it for statistical inference on treatment effects.

Assumption 10 *For $i = 1, \ldots, n$, the propensity score is defined by $e_i^* := \mathbb{P}(Z_i = 1 \mid G)$, and there exist constants $e_{\min}, e_{\max} > 0$ such that*

$$0 < e_{\min} < \min_{i=1,\ldots,n} e_i^* \leq \max_{i=1,\ldots,n} e_i^* < e_{\max} < 1$$

almost surely.

Assumption 10 guarantees that there is a substantial degree of overlap between the populations receiving the treatment and the control [DDF21]. In other words, the assumption states that there exists a range of values, bounded by constants $e_{\min}$ and $e_{\max}$, within which the propensity scores e_i^* fall for all individuals in the sample.

Assumption 11 *Let $\epsilon_i(Z_i)$ defined as above for $i = 1, ..., n$. Then we assume that the vector $(\epsilon_1(Z_1), ..., \epsilon_n(Z_n))^\top$ has independent coordinates that are mean zero sub-Gaussian with parameter $\sigma > 0$.*

Assumption 11 requires that the errors are independent and sub-Gaussian distributed with mean zero. See more details on the definition of sub-Gaussian random variables in Appendix 4.8.3. This characterizes the behavior of the error terms in the model, ensuring that extreme values or outliers in the error terms are limited, making the behavior of the model more stable. This condition is standard in non-parametric statistical model analysis. Also note that in Assumption 11, we only assume that $\epsilon_i(Z_i)$ and $\epsilon_j(Z_j)$ are independent for any units $i \neq j$, but for the same unit i, $\epsilon_i(1)$ and $\epsilon_i(0)$ are not necessarily independent.

We propose to estimate heterogeneous treatment effects using fused lasso over a graph, with a computation method inspired by the quasi-oracle estimator from [NW21]. Assume that we obtain an estimator $\hat{m}$ for the true main effect $m^* = (m_1^*, \ldots, m_n^*)^\top$ and an estimator

$\hat{e}$ for the true propensity score $e^* = (e_1^*, \ldots, e_n^*)^\top$. Then, our estimator for the treatment effect under the model setup (4.3) is defined as follows:

$$\hat{\tau} := \underset{\tau \in \mathbb{R}^n}{\arg\min} \left\{ \frac{1}{2} \sum_{i=1}^{n} (Y_i - \hat{m}_i - (Z_i - \hat{e}_i)\tau_i)^2 + \lambda_\tau \|D\tau\|_1 \right\}, \tag{4.4}$$

where $\lambda_\tau > 0$ is a tuning parameter, and D denotes the oriented incidence matrix of the graph G. The matrix D has a row for each edge and a column for each vertex of G. If the p-th edge in G connects the i-th and j-th observations, then

$$(D)_{p,q} = \begin{cases} 1 & \text{if } q = i, \\ -1 & \text{if } q = j, \\ 0 & \text{otherwise.} \end{cases}$$

The existence of the estimators $\hat{m}$ and $\hat{e}$ addresses the difficulty that in practice we never know the true main effect m^* and usually do not know the true propensity score e^* either. We will introduce how to obtain such estimators for our proposed model in Section 4.1. The term $\|D\tau\|_1$, the total variation of τ along the graph G, can be interpreted as a regularizer on the model complexity of the treatment effect function. This regularization is commonly applied in penalized regression like generalized lasso and its variations, and it can be explicitly expressed as

$$\|D\tau^*\|_1 := \sum_{(i,j) \in E} |\tau_i^* - \tau_j^*|.$$

4.2.3 A Model with Linear Representation on Covariates

In practical applications, researchers are often interested not only in understanding treatment effect heterogeneity but also in quantifying the individual contribution of each covariate to the outcome. To address this, we consider the setting where data (X_i, Y_i, Z_i) is available, where $X = (X_1, ..., X_n)$ is a covariate of dimension d, and (Y_i, Z_i) follow the same definitions as in the previous model.

A classical approach to modeling treatment effect heterogeneity involves linear regression with treatment-covariate interactions [Cox84, GC07, CHI08], which assumes the following

model:

$$Y_i = X_i^\top \beta^* + Z_i X_i^\top \gamma^* + u_i, \tag{4.5}$$

where the main effect m_i^* is assumed to follow a linear relationship with the covariates $m_i^* = X_i^\top \beta^*$, with $\beta^* \in \mathbb{R}^d$ representing a vector of coefficients associated with the covariates. The treatment effect τ_i^* is then decomposed into a systematic component, $X_i^\top \gamma^*$, which captures treatment effect variation explained by the observed covariates, and an idiosyncratic component, u_i, which represents variation unexplained by the covariates [DFM19].

Now, instead of assuming a linear representation of the covariates on treatment effects, we adhere to the assumption inherited from our first model that the true treatment effect vector τ^* is piecewise constant along the graph G. At the same time, we rewrite Assumption 9 as follows:

Assumption 12 *For $i = 1, \ldots, n$,*

$$Y_i(0), Y_i(1) \perp\!\!\!\perp Z_i \mid X_i, G.$$

We also make an assumption regarding confounding factors in data.

Assumption 13 *Let Γ be a set of confounding variables such that for $i = 1, \ldots, n$,*

$$X_i \not\perp\!\!\!\perp Y_i \mid \Gamma, \text{ and } Z_i \perp\!\!\!\perp \Gamma \mid X_i, G.$$

Then the following condition holds:

$$(Y_i \mid X_i, Z_i, G, \Gamma) \overset{d}{=} (Y_i \mid X_i, Z_i, G).$$

The assumption guarantees that, despite the presence of confounding factors between the observed covariates and outcomes, these confounding factors remain independent of the potential outcomes if conditioned on the underlying graph structure G. This means that the graph structure can fully explain the relationships between the observed variables, and therefore, confounding effects can be disregarded in the analysis.

Remark 4 *In Section 4.6, our analysis may not account for all hidden factors that could influence the outcome. For example, factors such as educational level may influence where individuals live, which in turn may have a hidden impact on the outcome. However, under the assumption that the chosen graph structure captures all relevant local patterns, we assume that these hidden factors can be ignored, as they do not introduce additional confounding in our analysis.*

With Assumptions 12 and 13, we reformulate the model (4.5) as:

$$Y_i = X_i^\top \beta^* + Z_i \tau_i^* + \epsilon_i, \tag{4.6}$$

where

$$\tau_i^* := \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i, G],$$

and the random error term ϵ_i is short for $\epsilon_i(Z_i)$. We assume that Assumption 11 holds under this model setup, and thereby ϵ_i satisfies $\mathbb{E}[\epsilon_i \mid X_i, Z_i, G] = 0$.

We propose our second estimator, named by causal covariate graph fused lasso, as the solution to the following optimization problem:

$$\hat{\beta},\ \hat{\tau} := \underset{\beta \in \mathbb{R}^d,\ \tau \in \mathbb{R}^n}{\arg\min} \ \Big\{ \frac{1}{2} \sum_{i=1}^{n} (Y_i - X_i^\top \beta - Z_i \tau_i)^2 + \lambda_\tau \|D\tau\|_1 \Big\}, \tag{4.7}$$

where λ_τ is a positive tuning parameter. This proposed estimator adopts a formulation similar to the semi-parametric regression estimator in [Rob88]. The treatment effect τ can be considered as an explanatory variable, entering the equation as a linear term, alongside a corresponding penalization with respect to the graph G. The method to find the solution to the optimization problem (4.7) will be introduced in Section 4.2.

The main benefit of this estimator lies in its capacity to provide not only estimates of treatment effects but also coefficients. These coefficients offer an additional channel that enables researchers to simultaneously understand the impact of individual features within the covariates. Another advantage of this estimator is that we avoid explicit estimation of the propensity score, typically a challenging task to estimate accurately in practice.

4.2.4 K-Nearest-Neighbor Graph

Given the assumption that the true treatment effect vector τ^* exhibits local structure along the graph G, an ideal graph for estimation should be connected and satisfy $\|D\tau^*\|_1 \ll n$, where the number of edges follows $|E| = O_{\mathbb{P}}(n)$. This condition implies that the total variation of τ^* along the graph is small, indicating sparsity in the edge set. In other words, the graph should avoid unnecessary edges that do not provide meaningful information about local patterns in the data, ensuring that the underlying structure is captured effectively while minimizing overfitting and redundancy. Chain graphs [OSL20], bipartite graphs [PBF22], and directed acyclic graphs [LPZ24] provide examples of the integration of connected graphs into causal estimation. However, as discussed in Chapter 2, predefined graph structures are often unavailable in practice. In such cases, when an observed covariate is present, a K-nearest-neighbor graph constructed from the covariate serves as a natural choice for preserving local data patterns. Therefore, in the following analysis and experiments, we focus on utilizing K-NN graphs in scenarios where no predefined graph structure is available.

The K-NN graph G is constructed as follows: given the d-dimensional covariate X corresponding to the data (Y, Z), the edge set E_K of G contains the pair (i, j), where $i \in V$, $j \in V$, and $i \neq j$, if and only if X_i is among the K-nearest neighbors of X_j, with respect to the metric $d_{\mathcal{X}}$, or vice versa. Throughout, we assume that the ambient dimension d is a fixed constant with $d \geq 2$. Accordingly, we denote ∇_G as the oriented incidence matrix of the K-NN graph G, following the same definition as the incidence matrix D of a general graph. Then by integrating the K-NN graph introduced above, the K-NN graph version of causal covariate graph fused lasso estimator (4.7) can be rewritten by replacing the oriented incidence matrix D of a general graph G by that of the K-NN graph, ∇_G in the optimization problem.

4.2.5 Comparison with Related Work

Conventional methods in causal inference, such as tree-based methods [AI16, ATW19, HMC20, CKS23], meta-learners [IR13, LL16, KSB19, NW21], and doubly robust methods [ZZL15, CCD18], typically assume that treatment effects can be estimated by conditioning on a covariate X from observed data (Y, Z). These approaches primarily aim at estimating the conditional average treatment effect (CATE),

$$\tau_i^*(X) := \mathbb{E}\left[Y_i(1) - Y_i(0) \mid X\right],$$

which is conditioned on the covariate X.

In contrast, our proposed method, causal graph fused lasso, allows for direct inference on heterogeneous treatment effects using graph structures, without requiring an observed covariate. The model (4.3) is formulated under Assumptions 9 and 10, conditioned on the graph G, rather than on a covariate X. The target estimand (4.3) is also conditioned on the graph itself, and the corresponding estimation from the optimization problem (4.4) leverages the inherent structure of the graph solely. This approach not only offers a departure from traditional covariate-based methods but also underscores the capability of graph-based methodologies to capture local relationships and dependencies that may not be fully captured by traditional covariate-centric models. This flexibility makes our method particularly suitable for scenarios where direct observation of covariates is limited or where graph-based data representations naturally reflect underlying causal mechanisms.

In the context of l_1 penalized optimization for estimating heterogeneous treatment effects, [IR13] stood out as a pioneer by introducing l_1 constraints to model treatment effect heterogeneity. Their seminal work formulated the estimator as:

$$(\hat{\beta}, \hat{\gamma}) := \underset{(\beta,\gamma)}{\arg\min} \left\{ \sum_{i=1}^{n} z_i \cdot |1 - Y_i^* \cdot (\mu + \beta^\top W_i + \gamma^\top V_i)|_+^2 + \lambda_W \sum_{i=1}^{L_W} |\beta_j| + \lambda_V \sum_{j=1}^{L_V} |\gamma_j| \right\}. \quad (4.8)$$

Here, the penalized lasso constraints are applied to treatment effect heterogeneity variables, enhancing the optimization problem's robustness.

Another contribution utilizing a lasso penalty for estimating heterogeneous treatment effects is found in [PCR21]. They proposed a two-step procedure where samples are initially ordered based on propensity or prognostic scores, followed by applying the fused lasso to obtain piecewise constant treatment effects respecting the predefined ordering of the score.

The proposed model setups (4.3) and (4.6) in this chapter align with the framework introduced by [Rob04]. Extending the same framework, [NW21] introduced a comprehensive quasi-oracle framework for specifying heterogeneous treatment effect estimators through empirical loss minimization:

$$\hat{\tau} := \underset{\tau \in \mathbb{R}^n}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - m^*(X_i) - (Z_i - \hat{e}_i)\tau_i)^2 + \Lambda_n(\tau) \right\}. \tag{4.9}$$

where m^* and e^* are considered as a priori functions and $\Lambda_n(\tau)$ acts as a regularizer on the complexity of the treatment effect function.

Compared to existing literature, the two methods introduced in this chapter represent a pioneering effort in integrating graph structures with l_1-penalized estimation of heterogeneous treatment effects. We build upon the framework (4.9) proposed by [NW21] by incorporating a fused lasso over graph regularization into our methods. By penalizing the total variation of the estimator over the graph, our proposed estimators enhance local adaptivity through the flexibility of accommodating diverse graph structures. Such integration represents a significant advancement in uncovering local patterns inherent in graph data commonly observed in real-world applications.

4.3 Methodology and Computation

4.3.1 Causal Graph Fused Lasso

Our first estimator, named causal graph fused lasso, solves the optimization problem (4.4). Our estimation involves a multi-step process by sequentially estimating the main effect, the propensity score, and eventually the estimand, the treatment effect conditioned on the graph, through the fused lasso over a graph G.

Step 1. The first step is to estimate the main effect vector $m^* = (m_1^*, ..., m_n^*)^\top$, where $m_i^* = \mathbb{E}[Y_i|G]$. We solve a fused lasso over graph with a tuning parameter $\lambda_m > 0$ along the observed outcome Y to obtain $\hat{m}$:

$$\hat{m} := \underset{m \in \mathbb{R}^n}{\arg\min} \left\{ \frac{1}{2} \sum_{i=1}^{n} (Y_i - m_i)^2 + \lambda_m \|Dm\|_1 \right\}. \tag{4.10}$$

Note that this optimization problem shares the same form as a generalized lasso problem that can be solved with the parametric max-flow algorithm from [CD09].

Step 2.

Case 1. If the true propensity score e^* is known, we set $\hat{e} = e^*$.

Case 2. If the true propensity score e^* is unknown, we need to provide an estimate $\hat{e}$, which we obtain via the fused lasso over graph against the treatment indicator Z:

$$\hat{e}_i = \begin{cases} \tilde{e}_i & \text{if } \tilde{e}_i \in [e_{\min}, e_{\max}], \\ e_{\min} & \text{if } \tilde{e}_i < e_{\min}, \\ e_{\max} & \text{if } \tilde{e}_i > e_{\max}, \end{cases} \tag{4.11}$$

where

$$\tilde{e} := \underset{e \in \mathbb{R}^n}{\arg\min} \left\{ \frac{1}{2} \sum_{i=1}^{n} (Z_i - e_i)^2 + \lambda_e \|De\|_1 \right\}, \tag{4.12}$$

and λ_e is a positive tuning parameter.

The objective of setting constraints on the estimated propensity score is to ensure that our estimators adhere to the overlap condition outlined in Assumption 10. Following the practical guideline of the 0.1 rule-of-thumb mentioned in [CHI09], we can empirically set two bounds for the estimated propensity score, denoted as $\{e_{\min}, e_{\max}\}$, with suggested values of $\{0.1, 0.9\}$. This empirical choice of bounds aims to provide sufficient variation in the estimated propensity score while preventing extreme values that could compromise the overlap condition.

Step 3. Estimate treatment effect $\hat{\tau}$ via fused lasso over graph:

$$\hat{\tau} := \underset{\tau \in \mathbb{R}^n}{\arg\min} \left\{ \frac{1}{2} \sum_{i=1}^{n} (Y_i - \hat{m}_i - (Z_i - \hat{e}_i)\tau_i)^2 + \lambda_\tau \|D\tau\|_1 \right\}, \tag{4.13}$$

with the tuning parameter $\lambda_\tau > 0$.

The optimization problem (4.13) is a penalized optimization problem, which can be easily solved via the alternating directions method of multipliers (ADMM) algorithm [BPC11]. We first reformulate the equation (4.13) as:

$$\begin{aligned} &\underset{\tau \in \mathbb{R}^n, w \in \mathbb{R}^n}{\text{minimize}} \ \frac{1}{2} \sum_{i=1}^{n} (Y_i - \hat{m}_i - (Z_i - \hat{e}_i)\tau_i)^2 + \lambda_\tau \|Dw\|_1 \\ &\text{s.t.} \qquad\qquad w = \tau, \end{aligned} \tag{4.14}$$

and the corresponding augmented Lagrangian can then be written as:

$$L_R(\tau, w, u) = \frac{1}{2} \sum_{i=1}^{n} (Y_i - \hat{m}_i - (Z_i - \hat{e}_i)\tau_i)^2 + \lambda_\tau \|Dw\|_1 + \frac{R}{2} \|\tau - w + u\|_2^2,$$

where $R > 0$ is the penalty parameter that controls the step size in the update and is conventionally set to 0.5. Thus, we can solve this optimization problem by iteratively updating the primal:

$$\tau \leftarrow \underset{\tau \in \mathbb{R}^n}{\arg\min} \left\{ \frac{1}{2} \sum_{i=1}^{n} (Y_i - \hat{m}_i - (Z_i - \hat{e}_i)\tau_i)^2 + \frac{R}{2} \|\tau - w + u\|_2^2 \right\}, \tag{4.15}$$

and the dual:

$$w \leftarrow \underset{w \in \mathbb{R}^n}{\arg\min} \left\{ \frac{1}{2} \|\tau + u - w\|_2^2 + \frac{\lambda_\tau}{R} \|Dw\|_1 \right\}. \tag{4.16}$$

The primal problem (4.15) can be solved in closed form using a coordinate-wise approach by differentiating with respect to τ_i:

$$\begin{aligned} \frac{\partial}{\partial \tau_i} &= (Z_i - \hat{e}_i)^2 \tau_i - (Y_i - \hat{m}_i)(Z_i - \hat{e}_i) + R\tau_i - R(w_i - u_i) = 0, \\ &\Rightarrow \tau_i = \frac{R(z_i - u_i) + (Z_i - \hat{e}_i)(Y_i - \hat{m}_i)}{(Z_i - \hat{e}_i)^2 + R}, \end{aligned} \tag{4.17}$$

for $i = 1, ..., n$, and the dual problem (4.16) can be solved with the parametric max-flow algorithm similar to solving the optimization problem (4.10).

We provide the entire computing procedure for the first proposed estimator in Algorithm 3. Note that the causal graph fused lasso estimator is non-parametric to some extent since the estimator does not assume a specific functional form of the covariates. Local structure on data is learned exclusively through the representation of a graph. Therefore, this estimator can learn complex and potentially nonlinear relationships in treatment effect heterogeneity.

4.3.2 Causal Covariate Graph Fused Lasso

To compute the causal covariate graph fused lasso estimator, we need to find the solution to the optimization problem (4.7). We observe that this objective function can be considered bi-convex in β and τ, similar to the loss function (9) in [LPZ24], where we can separate the optimization problem into two separate convex blocks for β and τ. Hence, we can address the above problem via a block coordinate descent algorithm. The convergence of coordinate descent algorithms to a stationary point in bi-convex non-differentiable problems has been well-established in Theorem 4.1 of [Tse01], where the stationary point is defined by a point where all directional sub-gradients are non-negative.

Suppose that we already have a putative $\hat{\beta}$, we only need to solve the optimization problem

$$\hat{\tau} := \underset{\tau \in \mathbb{R}^n}{\arg\min} \left\{ \frac{1}{2} \sum_{i=1}^{n} (Y_i - X_i^\top \hat{\beta} - Z_i \tau_i)^2 + \lambda_\tau \|D\tau\|_1 \right\} \tag{4.18}$$

to obtain $\hat{\tau}$. This problem is again a weighted penalized optimization problem same as (4.13), therefore we can apply a similar ADMM algorithm as in Algorithm 1. The details of the computation are provided in Appendix 4.8.1.

We can then go ahead and plug in the estimated $\hat{\tau}$ into the original optimization expression to update $\hat{\beta}$ as the solution:

$$\hat{\beta} := \underset{\beta \in \mathbb{R}^d}{\arg\min} \left\{ \frac{1}{2} \sum_{i=1}^{n} (Y_i - X_i^\top \beta - Z_i \hat{\tau}_i)^2 \right\}, \tag{4.19}$$

where we remove the penalty term from (4.7) since we are no longer optimizing for $\hat{\tau}$. The

Algorithm 3: Compute causal graph fused lasso estimators

Input: penalty parameters: $\lambda_m > 0$, $\lambda_e > 0$ and $\lambda_\tau > 0$, maximum iteration: N_{iter}, tolerance: κ

Data: $Y \in \mathbb{R}^n$, $Z \in \mathbb{R}^n$, oriented incidence matrix D, optional $e \in \mathbb{R}^n$

Output: $\hat{\tau} \in \mathbb{R}^n$

1. Compute

$$\hat{m} \leftarrow \arg\min \left\{ \frac{1}{2} \sum_{i=1}^{n} (Y_i - m_i)^2 + \lambda_m \|Dm\|_1 \right\}.$$

2. If e^* is given, set $\hat{e} = e^*$. If e^* is not given, for $i = 1, ..., n$, compute

$$\hat{e}_i = \begin{cases} \tilde{e}_i & \text{if } \tilde{e}_i \in [0.1, 0.9], \\ 0.1 & \text{if } \tilde{e}_i < 0.1, \\ 0.9 & \text{if } \tilde{e}_i > 0.9, \end{cases}$$

where

$$\tilde{e} \leftarrow \arg\min \left\{ \frac{1}{2} \sum_{i=1}^{n} (Z_i - e_i)^2 + \lambda_e \|De\|_1 \right\}.$$

3. Initialize $\tau^{(0)} = Y - \hat{m}$, $w^{(0)} = Y - \hat{m}$, $u^{(0)} = 0$.

4. For $k = 1, 2, ...$, until $\|\tau^{(k)} - \tau^{(k-1)}\|_2 \leq \kappa$ or the procedure reaches N_{iter}:

(a) For $i = 1, ..., n$, update

$$\tau_i^{(k)} \leftarrow \frac{R(w_i^{(k-1)} - u_i^{(k-1)}) + (Z_i - \hat{e}_i)(Y_i - \hat{m}_i)}{(Z_i - \hat{e}_i)^2 + R}.$$

(b) Update

$$w^{(k)} \leftarrow \arg\min \left\{ \frac{1}{2} \|\tau^{(k)} + u^{(k-1)} - w\|_2^2 + \frac{\lambda_\tau}{R} \|Dw\|_1 \right\}.$$

(c) Update

$$u^{(k)} \leftarrow u^{(k-1)} + \tau^{(k)} - w^{(k)}.$$

closed-form solution to this linear regression problem is simply:

$$\hat{\beta} = (X^\top X)^{-1} X^\top [Y - Z^\top \hat{\tau}],$$

if we assume $X^\top X$ is invertible.

One remaining problem is that we need to set up an initial $\beta^{(0)}$ for optimizing $\hat{\tau}$. A practical approach [ACW18] is as follows: we randomly select a subset of indices $S \in \{1, ..., n\}$ from the control group (for example, 50% of the available data in the control group). Then, we denote $X_S \subset X$ as a subset of the covariate X, and Y_S as the sub-vector of the response variable Y, both corresponding to the indices in the set S. Next, we perform ordinary least squares (OLS) regression on Y_S against X_S. In the case when $X_S^\top X_S$ is invertible, with an initialization $\tau^{(0)} = 0$, we have that

$$\hat{\beta}^{(0)} = (X_S^\top X_S)^{-1} X_S^\top Y_S,$$

as the solution to (4.19) in the first iteration of the block coordinate descent algorithm.

We now present the entire procedure for computing causal covariate graph fused lasso estimator in Algorithm 4.

Algorithm 4: Compute causal covariate graph fused lasso estimators

---

Input: penalty parameter: $\lambda_\tau > 0$, maximum iteration: N_{iter}, tolerance: κ

Data: $Y \in \mathbb{R}^n$, $Z \in \mathbb{R}^n$, oriented incidence matrix D

Output: $\hat{\beta} \in \mathbb{R}^d$, $\hat{\tau} \in \mathbb{R}^n$

1. Set $\hat{\tau}^{(0)} = 0$, and randomly select a subset S from the control group and perform ordinary least squares to initialize $\hat{\beta}^{(0)}$ as

$$\hat{\beta}^{(0)} \leftarrow (X_S^\top X_S)^{-1} X_S^\top Y_S.$$

2. For $k = 1, 2, ...,$ until $\|\tau^{(k)} - \tau^{(k-1)}\|_2 \leq \kappa$ or the procedure reaches N_{iter}:

(a) Update

$$\hat{\tau}^{(k)} \leftarrow \arg\min \Big\{ \frac{1}{2} \sum_{i=1}^{n} (Y_i - X_i^\top \hat{\beta}^{(k-1)} - Z_i \tau_i)^2 + \lambda_\tau \|D\tau\|_1 \Big\}.$$

(b) Update

$$\hat{\beta}^{(k)} \leftarrow (X^\top X)^{-1} X^\top [Y - Z^\top \hat{\tau}^{(k)}].$$

4.3.3 Model Selection

The selection of tuning parameters, including λ_m in (4.10), λ_e in (4.12), and λ_τ in (4.4) and (4.7), constitutes an essential practical consideration in estimation as their value controls the smoothness of the corresponding estimators. Optimal values for the tuning parameters can be determined through cross-validation. Alternatively, we can employ Bayesian Information Criteria [BIC; Sch78] for regularization parameter selection.

In the context of a regularized regression problem with a given tuning parameter λ [WLT07], the BIC for the obtained estimator $\hat{\theta}_\lambda$ can be computed as:

$$\text{BIC}(\lambda) = \frac{1}{n}\sum(y_i - \hat{\theta}_\lambda)^2 + \log(n) \cdot df_L(\lambda),$$

where df_L represents the degree of freedom of the estimator for a given tuning parameter λ. During tuning parameter selection, we calculate the BIC for each λ from a set of candidates and choose the one that minimizes the statistic.

Following the approach of [TT12], we can estimate the degrees of freedom df_L by the number of connected components in the graph G after removing all edges j where $|(D\hat{\theta})_j|$ exceeds a threshold γ, typically very small (e.g., 10^{-2}).

4.4 Theoretical Analysis

4.4.1 Analysis on Causal Graph Fused Lasso Estimator

We first study the theoretical properties of the causal graph fused lasso estimator from the formulation (4.4).

Theorem 4.4.1 *Suppose that Assumptions 9-11 hold and* $\|\tau^*\|_\infty = O_{\mathbb{P}}(1)$. *If the estimated propensity score* $\hat{e}$ *satisfies that* $\hat{e}_i \in [e_{\min}, e_{\max}]$ *for all* $i = 1, ..., n$, *and*

$$\frac{1}{n}\|\hat{e} - e^*\|_2^2 = O_{\mathbb{P}}(r_{e^*}),\ \frac{1}{n}\|\hat{m} - m^*\|_2^2 = O_{\mathbb{P}}(r_{m^*}),$$

then for $\lambda_\tau \asymp \|D\tau^*\|_1^{-1/3} n^{1/3}$, *the estimated treatment effect* $\hat{\tau}$ *from (2.3) satisfies*

$$\frac{1}{n}\|\hat{\tau} - \tau^*\|_2^2 = O_{\mathbb{P}}\left(\frac{1}{n} + \frac{\|D\tau^*\|_1^{2/3}}{n^{2/3}} + r_{e^*} + r_{m^*}\right).$$

Theorem 4.4.1 establishes the consistency of the estimator $\hat{\tau}$ defined in (4.4) for estimating the heterogeneous treatment effects τ^*. It is worth noting that the upper bound on the mean squared error of $\hat{\tau}$ is determined by $\|D\tau^*\|_1$, the total variation of τ^* across the graph G. If both r_{e^*} and r_{m^*} converge faster than $\|D\tau^*\|_1^{2/3} n^{-2/3}$, we obtain a convergence rate of $\|D\tau^*\|_1^{2/3} n^{-2/3}$ for the proposed estimator from (4.4), which is applicable to any connected graph structure. This rate aligns with the rate of $t^{2/3} n^{-2/3}$ for the graph fused lasso estimator in [PSS18], where the authors assume that the underlying signal belongs to a bounded variation class, defined with respect to the graph, with a total variation of t.

Corollary 4.4.2 represents a direct extension of Theorem 4.4.1, which corresponds to our first estimator, causal graph fused lasso.

Corollary 4.4.2 *Given that the notation and assumptions from Theorem 4.4.1 are inherited. If* e^* *is known and* $\hat{m}$ *is estimated as in (4.10) with* $\lambda_m \asymp \|Dm^*\|^{-1/3} n^{1/3}$, *then the estimated* $\hat{\tau}$ *from (4.4) satisfies*

$$\frac{1}{n}\|\hat{\tau} - \tau^*\|_2^2 = O_{\mathbb{P}}\left(\frac{1}{n} + \frac{\|D\tau^*\|_1^{2/3} + \|Dm^*\|_1^{2/3}}{n^{2/3}}\right).$$

If instead, e^* *is unknown and we construct* $\hat{e}$ *as in (4.11) and (4.12) with* $\lambda_e \asymp \|De^*\|^{-1/3} n^{1/3}$, *then*

$$\frac{1}{n}\|\hat{\tau} - \tau^*\|_2^2 = O_{\mathbb{P}}\left(\frac{1}{n} + \frac{\|D\tau^*\|_1^{2/3} + \|Dm^*\|_1^{2/3} + \|De^*\|_1^{2/3}}{n^{2/3}}\right).$$

This corollary establishes the upper bound on the mean squared error of our proposed estimator $\hat{\tau}$ from (4.4) as a function of the total variations over the graph G, $\|D\tau^*\|_1$, $\|Dm^*\|_1$, and $\|De^*\|_1$, when we derive both the estimates $\hat{m}$ and $\hat{e}$ using fused lasso over graph. Again, this rate aligns with the rate in [PSS18], and it remains valid for arbitrary connected graphs, regardless of the structure of the graph G.

It is worth emphasizing that under appropriate boundedness conditions, particularly assuming that the total variations $\|D\tau^*\|_1$, $\|Dm^*\|_1$, and $\|De^*\|_1$ are bounded by some $t > 0$ over the graph G, our proposed estimator achieves an upper bound on the mean squared error of order $t^{2/3}n^{-2/3}$.

4.4.2 Causal Covariate Fused Lasso Estimator with K-NN Graph

In this section, we analyze the theoretical properties of the K-nearest-neighbor version of the causal covariate fused lasso estimator, in particular focusing on the convergence rates of the estimated treatment effects $\hat{\tau}$. Previous works by [LRH14, PSC20, YP21] have laid down theoretical foundations for non-parametric estimators involving the K-NN graph, upon which we build our theoretical analysis.

Given that the estimation process relies on a K-NN graph, additional assumptions on the data are necessary to establish the results. The following assumption regards the distribution of the covariate X.

Assumption 14 *The random vectors $X_1, ..., X_n$ 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 probability density function of X, p_X, is bounded. This amounts to*

$$0 < p_{\min} \leq \inf_{x \in [a,b]} p(x) \leq p_{\max},$$

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

We highlight that Assumption 14 aligns with the standard practice in non-parametric regression, as detailed in [PSC20]. It offers a broader scope than the assumption of uniform covariate distribution in $[0, 1]^d$, as seen in non-parametric regression models introduced in [GKK06], as well as in the context of heterogeneous treatment effects described in [WA18].

In addition to Assumption 14, the following two conditions, inherited from [LRH14], are essential when we confine our analysis to a K-NN graph instead of a general graph for our proposed estimator.

Assumption 15 *The base measure μ in the metric space $\mathcal{X}$, in which X is defined, satisfies*

$$c_{1,d} r^d \leq \mu\{B_r(x)\} \leq c_{2,d} r^d, \text{ for all } x \in \mathcal{X},$$

for all $0 < r < r_0$, where $r_0, c_{1,d}, c_{2,d}$ are positive constants, and $d \in \mathbb{N}\backslash\{0,1\}$ is the intrinsic dimension of $\mathcal{X}$.

While $\mathcal{X}$ may not necessarily be a Euclidean space, we require in this condition that balls in $\mathcal{X}$ possess volume with respect to denoted as μ on the Borel sigma algebra, $\mathcal{B}(\mathcal{X})$. This measure exhibits behavior similar to the Lebesgue measure of balls within $\mathbb{R}^d$. See details in Definition 2 in [LRH14].

Assumption 16 *There exists a homeomorphism $h : \mathcal{X} \to [0,1]^d$, such that*

$$L_{\min} d_{\mathcal{X}}(x, x') \leq \|h(x) - h(x')\|_2 \leq L_{\max} d_{\mathcal{X}}(x, x'),$$

for all $x, x' \in \mathcal{X}$ and for some positive constants $L_{\min}, L_{\max}$, where $d \in \mathbb{N}\backslash\{0,1\}$ is the intrinsic dimension of $\mathcal{X}$.

The existence of a continuous bijection between $\mathcal{X}$ and $[0,1]^d$ in Assumption 16 ensures a space without holes and topological equivalence to $[0,1]^d$, requiring $d > 1$, as generally assumed throughout this chapter.

Assumption 17 *The total variation of τ^* over the K-nearest-neighbor graph G satisfies the condition that $\|\nabla_G \tau^*\|_1 = O_{\mathbb{P}}(n^{1-1/d}\text{poly}(\log n))$.*

Under Assumptions 14–16, we can assert that the bounded total variation condition in Assumption 17 holds, with an additional condition on the true function of τ^* being piecewise Lipschitz. The class of piecewise Lipschitz functions is defined in previous papers by [PSC20] and [YP21], and we provide the corresponding details in Appendix 4.8.7. Notably, [PSC20] also introduces an alternative condition different from piecewise Lipschitz, to ensure the validity of their results; refer to Assumption 5 in the same paper.

Theorem 4.4.3 *Suppose that Assumptions 10-17 hold. We also assume that* $\max(\|X\beta^*\|_\infty, \|\tau^*\|_\infty) = O_{\mathbb{P}}(1)$, $\max_{i\in\{1,\dots,n\},j\in\{1,\dots,d\}} |X_{ij}| = C_X$ *for some positive constant* C_X, *and* $Z_i \perp\!\!\!\perp \epsilon_i$ *for* $i = 1, \dots, n$. *Consider the first iteration of the block coordinate descent algorithm in Algorithm 2 for solving the K-NN version of the optimization problem (4.7), subject to the constraint* $\|\beta\|_\infty \leq M$ *for some positive constant* M. *The number of nearest neighbors* K *is chosen as* $K = \log^{1+r} n$ *for some* $r > 0$. *If the initialized* $\hat{\beta}^{(0)}$ *is obtained such that* $\|\hat{\beta}^{(0)} - \beta^*\|_2^2 = O_{\mathbb{P}}(n^{-\alpha})$ *for* $\alpha \geq \frac{1}{d}$, *then for an appropriate choice of the tuning parameter* λ_τ *as in (4.49), the estimated treatment effect* $\hat{\tau}^{(1)}$ *from the first iteration of Algorithm 2 satisfies*

$$\frac{1}{n}\|\hat{\tau}^{(1)} - \tau^*\|_2^2 = O_{\mathbb{P}}\left(n^{-1/d}\text{poly}(\log n)\right).$$

Theorem 4.4.3 asserts that under mild assumptions on the data generation mechanism of the d-dimensional data and computation through the proposed algorithm, causal K-NN graph fused lasso estimator achieves a convergence rate of $n^{-1/d}$ within the first iteration of a block coordinate descent algorithm, up to a logarithmic factor. Importantly, this rate aligns with the convergence rate of non-parametric K-NN estimators, as demonstrated in [PSC20] and [YP21].

Remark 5 *As stated in [LPZ24], there are technical difficulties in establishing the consistency of the estimated* $\hat{\tau}$ *after the first iteration due to the dependence between* $\hat{\tau}$ *and* $\hat{\beta}$ *across iterations. Therefore, in Theorem 4.4.3, we demonstrate that, provided an appropriate initial estimate* $\hat{\beta}^{(0)}$ *that meets the assumptions specified in the claim, Algorithm 4 can produce a consistent estimator for the treatment effect after one iteration.*

In Algorithm 4, we initialize $\hat{\beta}^{(0)}$ *using ordinary least squares on* Y_S *against* X_S *over a subset* S *of the control group:*

$$\beta^{(0)} = (X_S^\top X_S)^{-1} X_S^\top Y_S.$$

This approach typically results in $\|\hat{\beta}^{(0)} - \beta^*\|_2^2 = O_{\mathbb{P}}(n^{-\alpha})$ *with* $\alpha \geq \frac{1}{d}$ *when* $d \geq 2$ *[Sha91], thereby the assumption in the theorem is satisfied with such initialization in general. Consequently, within the first iteration of the block coordinate descent algorithm introduced in*

Section 4.3, we establish the main claim in Theorem 4.4.3. While this result implies a convergence rate of order $n^{-1/d}$ for the treatment estimator $\hat{\tau}^{(1)}$ within the first iteration, the initialization of $\hat{\beta}^{(0)}$ remains arbitrary. Hence, we perform multiple iterations for updating both $\hat{\beta}$ and $\hat{\tau}$ in our algorithm to help convergence in practice.

4.5 Simulation Experiments

In preparation for the simulation experiments, we establish several key notations, including the sample size denoted as n, the ambient dimension represented as d, as well as the following essential functions at a random sample x:

$$\begin{aligned}
&\text{true main effect: } m^*(x) = \mathbb{E}[Y|X = x],\\
&\text{true treatment propensity: } e^*(x) = \mathbb{P}(Z = 1|X = x),\\
&\text{true treatment effect: } \tau^*(x) = \mathbb{E}\left[Y(1) - Y(0)\middle|X = x\right].
\end{aligned}$$

Throughout all data scenarios, we generate a covariate matrix $X = (X_1, \ldots, X_n)$ with random samples $X_i \sim [0,1]^d$ for $i = 1, \ldots, n$, and formulate the main effect $m^*(X) = (m^*(X_1), \ldots, m^*(X_n))$, the propensity score $e^*(X) = (e^*(X_1), \ldots, e^*(X_n))$, and the treatment effect $\tau^*(X) = (\tau^*(X_1), \ldots, \tau^*(X_n))$ under some defined functions for each Monte Carlo trial. Then the treatment indicator Z is generated as

$$Z_i \sim \text{Binom}(1, e^*(X_i)).$$

The observed outcome Y is defined as a function of m^*, e^*, Z, and τ^*, and we apply different methods to estimate τ^* with $\hat{\tau}$. For each trail, we evaluate the accuracy of an estimator $\hat{\tau}(X)$ by its mean-squared error (MSE) for estimating $\tau^*(X)$ at a sample covariate data X, defined by

$$\text{MSE}(\hat{\tau}(X)) := \frac{1}{n}\sum_{i=1}^{n}\left[\hat{\tau}(X_i) - \tau^*(X_i)\right]^2.$$

We repeat the process for 500 Monte Carlo trials, and record the averaged MSE over these trials for each estimator.

For our proposed estimators, causal graph fused lasso (GFL-1) and causal covariate fused lasso (GFL-2), if a graph is provided in the data generation mechanism, we compute the estimates directly by integrating the oriented incidence matrix of the graph. In scenarios where a graph is not available, we construct a K-nearest-neighbor graph with $K = 5$ using the observed covariate X, and then compute the estimates with this constructed K-NN graph. We benchmark the performance of our proposed estimators against established competitors, including R-learner [NW21], S-learner [IR13], T-learner [AI16], X-learner [KSB19], and causal forests [CF; ATW19]. To explore the application of community detection methods in estimating treatment effects, we augment the covariate set of causal forests with graph statistics (CF-G) commonly used in community detection, including closeness centrality [NG04], clustering coefficient [WS98], and average neighbor degree [BBP04] of the same K-NN graph built from our methods. We analyze whether incorporating these graph patterns enhances the performance of causal forests.

For meta-learners, we employ functions from the R package "rlearner" developed by [NW21]. For causal forests, we directly use the R package "grf" [ATW19], following with an automatic cross-validation to select values of hyper-parameters [AW19] and a K-fold cross-fitting under a conventional setting of $K = 10$ recommended by [NW21].

Scenario 1: Random trials with grid graph

We construct a covariate $X \in \mathbb{R}^2$, with $X_{i,j} = (\frac{i}{N}, \frac{j}{N})$ for $i, j = 1, ..., N$ where $N > 0$ is a predefined constant. Additionally, we assume a 2D grid graph over $[0, 1]^2$, with lattice defined on the data points of X and vertices connected adjacent points in X. The main effect and the true treatment effect are then defined as follows:

$$m^*(X_{i,j}) = \begin{cases} 1 & \text{if } \frac{i}{4N} + \frac{3j}{4N} < 0.5, \\ 0 & \text{otherwise.} \end{cases}$$

$$\tau^*(X_{i,j}) = \begin{cases} 1 & \text{if } \frac{i}{N} < 0.6 \text{ and } \frac{j}{N} < 0.6, \\ 0 & \text{otherwise.} \end{cases}$$

We set $e^*(X_{i,j}) = 0.5$ for all $i, j = 1, ..., N$ and with $\epsilon_{i,j} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$, let the outcome Y be

$$Y_{i,j} = m^*(X_{i,j}) + (Z_{i,j} - e^*(X_{i,j}))\tau^*(X_{i,j}) + \epsilon_{i,j}.$$

We record the performance of an estimator at $N = 30, 50, 100$.

Scenario 2: Random trials with unknown graph

With $d = 5$, we set the propensity score $e^*(X)$ as 0.5 across all observations, and for $i = 1, ..., n$, we generate the data following the model:

$$\begin{aligned}
m^*(X_i) &= \sin(\pi X_{i,1} X_{i,2}) + 2(X_{i,3} - 0.5)^2 + X_{i,4} + 0.5 X_{i,5}, \\
\tau^*(X_i) &= \begin{cases} 0.5 & \text{if } \min(X_{i,1}, X_{i,2}, ..., X_{i,5}) \geq \frac{1}{4}, \\ -0.5 & \text{if } \max(X_{i,1}, X_{i,2}, ..., X_{i,5}) < \frac{1}{4}, \\ 0 & \text{otherwise,} \end{cases} \\
Y_i &= m^*(X_i) + (Z_i - e^*(X_i))\tau^*(X_i) + \epsilon_i,
\end{aligned}$$

with $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, 0.75^2)$.

Scenario 3: Multi-dimensional covariate with unknown graph and unknown propensity score

Assuming the treatment effect $\tau^*(X)$ being constant 1 across all observations, we generate the data under the following model for $i = 1, ..., n$:

$$\begin{aligned}
m^*(X_i) &= \begin{cases} 2 & \text{if } X_i^\top \beta > [d/2] - 1, \\ -2 & \text{otherwise,} \end{cases} \\
e^*(X_i) &= \begin{cases} 0.75 & \text{if } (X_i^2)^\top \beta > \lfloor d^{1/2} \rfloor - 1, \\ 0.25 & \text{otherwise,} \end{cases} \\
Y_i &= m^*(X_i) + (Z_i - e^*(X_i))\tau^*(X_i) + \epsilon_i,
\end{aligned}$$

where $\beta \sim \text{Unif}(0,1)^d$ and $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, 0.5^2)$. We examine the performance of estimators with the ambient dimension $d = 10$.

Scenario 4: Data based on graph statistics

We generate a covariate X of dimension $d = 10$ and construct a corresponding K-nearest-neighbor graph G based on the covariate X with $K = 5$. Then, we construct the data as follows:

$$m^*(X_i) = \begin{cases} 1 & \text{if } C_C(X_i) > 0.1 \text{ and } CL(X_i) > 0.1, \\ -2 & \text{otherwise,} \end{cases}$$

$$e^*(X_i) = \begin{cases} 0.75 & \text{if } n_{\text{avg}}(X_i) > 12, \\ 0.5 & \text{if } 10 < n_{\text{avg}}(X_i) \leq 12, \\ 0.25 & \text{otherwise,} \end{cases}$$

$$\tau^*(X_i) = \begin{cases} 1.5 & \text{if } C_C(X_i) > 0.15 \text{ and } n_{\text{avg}}(X_i) > 10, \\ 0.5 & \text{if } C_C(X_i) < 0.15 \text{ and } n_{\text{avg}}X_i) < 10, \\ 1 & \text{otherwise,} \end{cases}$$

$$Y_i = m^*(X_i) + (Z_i - e^*(X_i))\tau^*(X_i) + \epsilon_i,$$

where $C_C(X_i)$ denotes the closeness centrality of the node in the graph G that corresponds to the data X_i, $CL(X_i)$ denotes the clustering coefficient of that node, and $n_{\text{avg}}(X_i)$ denotes the average degree of all connected neighbors of the same node. We set the error term $\epsilon_i \sim \overset{i.i.d.}{\sim} \mathcal{N}(0, 0.5^2)$.

Scenario 5: Data with the model on covariates

With $d = 10$, we generate the data under the following mechanism:

$$e^*(X_i) = \begin{cases} 0.8 & \text{if } \|X_i - \frac{1}{2}1_d\|_2^2 \leq \frac{4}{100}, \\ 0.2 & \text{otherwise,} \end{cases}$$

$$\tau^*(X_i) = \begin{cases} 2 & \text{if } \|X_i - q_1\|_2 < \min\{\|X_i - q_2\|_2, \|X_i - q_3\|_2, \|X_i - q_4\|_2\}, \\ 1 & \text{if } \|X_i - q_2\|_2 < \min\{\|X_i - q_1\|_2, \|X_i - q_3\|_2, \|X_i - q_4\|_2\}, \\ 0 & \text{if } \|X_i - q_3\|_2 < \min\{\|X_i - q_1\|_2, \|X_i - q_2\|_2, \|X_i - q_4\|_2\}, \\ -1 & \text{otherwise,} \end{cases}$$

where $q_1 = \left(\frac{1}{4}1_{\lfloor d/2 \rfloor}^\top, \frac{1}{2}1_{d-\lfloor d/2 \rfloor}^\top\right)^\top$, $q_2 = \left(\frac{1}{2}1_{\lfloor d/2 \rfloor}^\top, \frac{1}{4}1_{d-\lfloor d/2 \rfloor}^\top\right)^\top$, $q_3 = \left(\frac{3}{4}1_{\lfloor d/2 \rfloor}^\top, \frac{1}{2}1_{d-\lfloor d/2 \rfloor}^\top\right)^\top$, and $q_4 = \left(\frac{1}{2}1_{\lfloor d/2 \rfloor}^\top, \frac{3}{4}1_{d-\lfloor d/2 \rfloor}^\top\right)^\top$, and

$$Y_i = X_i^\top \beta^* + Z_i \tau^*(X_i) + \epsilon_i,$$

with $\beta^* = (0.1, 0.7, 0.2, 0.4, 0.6, -0.3, -0.5, -0.9, 0.8, -1)^\top$ and $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, 0.5^2)$.

In Scenario 1, we conduct random trials on a 2D grid, defining the true main effect based on a threshold condition. Scenario 2 involves random trials with multi-dimensional covariates, capturing non-linear relationships in true main effects, and also features a piecewise constant true treatment effect. Scenario 3 generates data with constant treatment effects, where the main effect and the propensity score are structured by conditional assignment based on covariates rather than graph. In Scenario 4, we generate data based on local structures in graph statistics such as closeness centrality, clustering coefficient and average neighbor degree. Throughout these scenarios, we assume the data model for the observed outcome follows our first model detailed in Section 4.2. Scenario 5 incorporates the second model from Section 4.2, featuring a linear representation of covariates in the main effect and a treatment effect determined by a piecewise constant structure. In Scenarios 1 and 4, we consider the graph as predefined. For Scenarios 2, 3, and 5, where no graph is provided, we construct a K-nearest-neighbor graph, setting $K = 5$, and use this graph in the estimation process of our proposed methods.

Table 4.1 summarizes the average mean squared errors from 500 Monte Carlo simulations across different data scenarios outlined earlier. Our first approach, causal graph fused lasso, consistently outperforms competing benchmarks in the majority of cases across Scenarios 1,

Table 4.1: Averaged mean squared errors of estimated treatment effects with standard errors in parenthesis over 500 Monte Carlo simulations

Scenario	n	d	GFL-1	GFL-2	R-	S-	T-	X-	CF	CF-G
1	900	2	0.1169	0.1194	0.1163	0.1170	0.1169	0.1165	**0.1086**	0.1410
			(0.0357)	(0.0343)	(0.0122)	(0.0120)	(0.0129)	(0.0119)	(0.0158)	(0.0155)
1	2500	2	**0.0492**	0.1150	0.1083	0.1085	0.1083	0.1082	0.0497	0.0655
			(0.0147)	(0.0186)	(0.0037)	(0.0039)	(0.0038)	(0.0038)	(0.0096)	(0.0112)
1	10000	2	**0.0114**	0.1046	0.1059	0.1060	0.1059	0.1059	0.0165	0.0179
			(0.0042)	(0.0099)	(0.0010)	(0.0011)	(0.0011)	(0.0011)	(0.0031)	(0.0036)
2	1000	5	**0.0478**	0.0489	0.1439	0.1431	0.1429	0.1430	0.1570	0.1607
			(0.0039)	(0.0044)	(0.0117)	(0.0102)	(0.0103)	(0.0102)	(0.0081)	(0.0080)
2	5000	5	**0.0461**	0.0467	0.1312	0.1313	0.1313	0.1312	0.0987	0.1157
			(0.0010)	(0.0010)	(0.0028)	(0.0029)	(0.0029)	(0.0029)	(0.0069)	(0.0056)
2	10000	5	**0.0459**	0.0461	0.1297	0.1298	0.1298	0.1298	0.0607	0.0842
			(0.0006)	(0.0007)	(0.0018)	(0.0018)	(0.0018)	(0.0018)	(0.0067)	(0.0065)
3	1000	10	**0.0023**	0.0336	0.1359	0.3806	0.4062	0.3740	0.6524	0.7065
			(0.0036)	(0.0334)	(0.1099)	(0.2504)	(0.2593)	(0.2499)	(0.5474)	(0.5918)
3	5000	10	**0.0007**	0.0417	0.1170	0.3281	0.3361	0.3287	0.3223	0.3558
			(0.0012)	(0.0534)	(0.0860)	(0.2142)	(0.2184)	(0.2156)	(0.2880)	(0.3177)
3	10000	10	**0.0004**	0.0556	0.1191	0.3358	0.3434	0.3375	0.2427	0.2469
			(0.0006)	(0.0770)	(0.0805)	(0.2065)	(0.2110)	(0.2086)	(0.2116)	(0.2355)
4	1000	10	0.0566	0.0693	0.0672	0.0663	0.0731	0.0673	0.0623	**0.0422**
			(0.0101)	(0.0150)	(0.0201)	(0.0181)	(0.0212)	(0.0187)	(0.0149)	(0.0042)
4	5000	10	0.0561	0.0633	0.0621	0.0619	0.0636	0.0621	0.0628	**0.0163**
			(0.0052)	(0.0072)	(0.0088)	(0.0086)	(0.0087)	(0.0084)	(0.0084)	(0.0049)
4	10000	10	0.1031	0.1078	0.1072	0.1076	0.1085	0.1074	0.1047	**0.0150**
			(0.0040)	(0.0049)	(0.0061)	(0.0062)	(0.0063)	(0.0060)	(0.0062)	(0.0026)
5	1000	10	0.7203	**0.4743**	0.6258	0.6280	0.6227	0.6239	1.1050	1.1169
			(0.0495)	(0.0328)	(0.0250)	(0.0260)	(0.0248)	(0.0249)	(0.0289)	(0.0295)
5	5000	10	0.5642	**0.3676**	0.6119	0.6137	0.6127	0.6126	0.9151	0.9396
			(0.0208)	(0.0150)	(0.0101)	(0.0104)	(0.0102)	(0.0103)	(0.0139)	(0.0146)
5	10000	10	0.5120	**0.3330**	0.6095	0.6112	0.6108	0.6107	0.8422	0.8710
			(0.0137)	(0.0111)	(0.0070)	(0.0072)	(0.0070)	(0.0071)	(0.0100)	(0.0107)

2, and 3. Particularly noteworthy is its effectiveness in Scenario 2, where treatment effects exhibit piecewise constancy over graph structures, highlighting its capability in leveraging local treatment effect patterns encoded in the graph. Even in scenarios where data does

not explicitly incorporate graph structures, such as Scenario 3, causal graph fused lasso demonstrates competitive performance compared to its peers. However, in Scenario 4, our method shows relatively weaker performance compared to causal forests enhanced with graph statistics, aligning with expectations given the data generation mechanism favoring that approach. Nevertheless, our method consistently outperforms competitors such as meta-learners and causal forests using only covariates.

To further illustrate the effectiveness of our first approach, we present a plot of the true treatment effect alongside the corresponding estimate from the causal graph fused lasso under Scenario 1, using a sample size of 10,000, in Figure 4.1. This visualization demonstrates the accuracy of the method in estimating diverse treatment effects captured by a graph. However, it is important to note that in scenarios with smaller sample sizes, benchmark competitors such as causal forests show better performance than our proposal, and the performance of our proposal improves as the sample size increases. This suggests that our first proposed method offers advantages in larger sample scenarios.

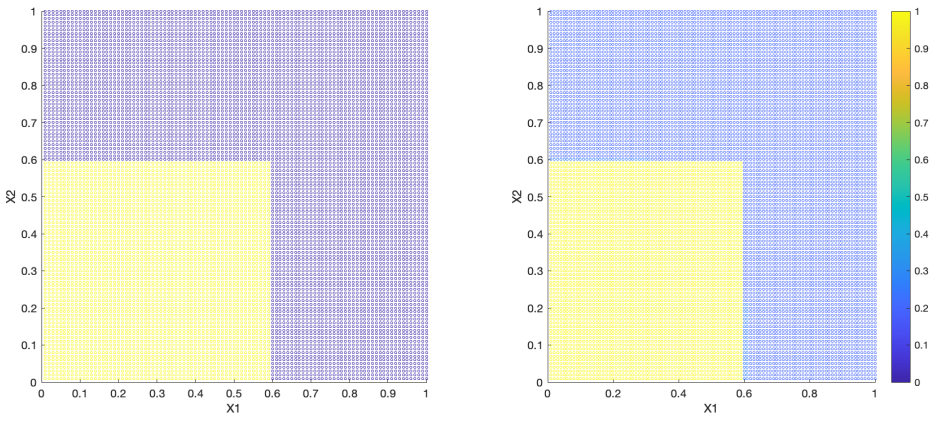

Figure 4.1: A comparison of a single instance between the true treatment effect (*left*) and the estimated treatment effect obtained from causal graph fused lasso (*right*) from Scenario 1. The color bar indicates the values of treatment effects.

If we examine the performance of our second estimator, causal covariate graph fused lasso, from Table 4.1, it becomes apparent that this approach generally underperforms

when compared to the first proposal across most cases. The only exception occurs in Scenario 5, where the main effect is defined by a linear representation, consistent with the underlying statistical model of our second method. Therefore, the superior performance of our second proposal is anticipated in this data scenario. It is important to acknowledge that the motivation behind proposing causal covariate graph fused lasso method is to provide an alternative for researchers interested in understanding the impact of key features on the outcome, particularly in social science contexts. Despite its limitations, this model remains valuable. In examining Scenario 5, we assess the proximity of our estimates on covariate coefficients to the true values. Averaging the covariate coefficients over 500 Monte Carlo simulations with a sample size of 10,000, we obtain an averaged $\hat{\beta} = (0.0978, 0.6942, 0.1926, 0.4010, 0.5984, -0.3043, -0.4987, -0.9048, -0.7950, -1.0050)^\top$. This result closely approximates the true values $\beta^* = (0.1, 0.7, 0.2, 0.4, 0.6, -0.3, -0.5, -0.9, -0.8, -1)^\top$ in the model.

4.6 Application: U.S. Transportation Network Crime Data

4.6.1 Background and Data

Transportation network companies (TNCs), like Uber and Lyft, have become an integral part of modern life in recent years. The convenience and time-saving aspects of these ride-hailing applications revolutionize the way people travel and make them popular among a significant portion of our daily life. Additionally, TNCs have created numerous job opportunities, allowing many individuals to earn a living through flexible work schedules. In the United States, specifically, the use of ride-hailing services has surged, with a survey by [Cen19] revealing that 36 percent of U.S. adults have used services like Uber or Lyft. Meanwhile, recent estimates suggest there are about 1.5 million to 2 million Uber and Lyft drivers in the United States, including both active and part-time drivers.

However, alongside the undeniable convenience offered by TNC services, there has been an unfortunate increase in crimes related to ride-sharing activities. Incidents of violence,

such as rape, assault, child molestation, and theft, as well as accidents resulting in injuries or hit-and-run incidents, have been frequently reported involving TNC drivers. To mitigate these negative consequences, the legislature in the United States has taken several measures. One of the most significant steps has been the introduction of driver background check laws (BCLs). For example, in New York City, Uber and Lyft drivers are required to obtain a license from the NYC Taxi and Limousine Commission (TLC) [TC23]. This ensures that ride-share drivers in NYC are subject to the same regulations as taxi and limousine drivers. To obtain a TLC license, drivers must undergo thorough background and driving history checks. The goal of these measures is to identify and exclude individuals with a history of criminal offenses from participating in ride-sharing services. By implementing these laws, the legislature aims to enhance passenger safety and minimize the risks associated with TNC services.

Previous analyses on different components of background check laws have shown evidence of their effectiveness in reducing crimes. For instance, [SP12] discovered significant effects indicating that checks for mental illness, restraining orders, or fugitive status resulted in a reduction in crimes like homicides. However, there are researchers who question the efficacy of these policies in crime reduction. They argue that these laws may not have the anticipated impact or could even lead to negative effects, such as an increase in the number of property crimes. In a recent study by [RCL24], concerns were raised about the unintended social welfare risks such as an increase in going back to shadier business for alternative income opportunities associated with excluding marginalized individuals from gig employment opportunities like jobs created from TNC services.

Our research aims to investigate the impact of enforcing background check laws on gig workers for transportation network companies (TNCs) on the number of property crimes. The definition of property crimes includes the following categories: robbery, burglary, larceny-theft, and motor vehicle theft. We seek to verify the unintended consequences of implementing background check laws on TNC-related gig workers, hypothesizing an increase in local property crimes after enforcing such legislation. Our study focuses on a dataset collected in

recent years, spanning from 2015 to 2018. The dataset consists of monthly crime counts at different local levels across the United States, publicly accessible through the Federal Bureau of Investigation (FBI) website (`https://cde.ucr.cjis.gov/`). The dataset also includes information on Uber and Lyft's service availability in the regions. Additionally, demographic and macroeconomic factors including total population, unemployment rate, poverty rate, average annual GDP growth rate, proportion of elderly (65+) residents, gender distribution, and racial demographics (white and Asian populations) are included. The detailed data source and descriptions for each variable are presented in Appendix 4.8.2.

To address our research question, our analysis targets analyzing the crime rate, which is calculated as the number of reported crimes per 100,000 individuals within a population as follows.

$$\text{Crime Rate} = \frac{\text{Crime Count}}{\text{Total Population}} \times 100,000.$$

The crime rate serves as the response variable, with macroeconomic factors listed above as the explanatory variables in our analysis. Meanwhile, geographic coordinates, including longitude and latitude, are obtained for each data point.

4.6.2 Nationwide County Level Analysis

We start our analysis by examining the causal effect of introducing gig worker background check laws on property crime rates at the county level across the United States. We exclude certain counties, primarily in Florida, due to the absence of data on the number of crimes. Consequently, our cleaned dataset comprises records from 3015 counties across the United States. The county-level analysis spans from 2015 to 2018, encompassing the initial state-level implementation of background check laws on TNCs in 2015 and the final implementation in 2017 across the United States. Crime counts up to 2018 are aggregated for crime rate analysis. Given the varied timing of law implementation across regions, we employ a cohort-based causal identification approach with staggered treatment adoption. This methodology, as applied in [AAH21], [CS21], and [SA21], aims to mitigate comparison issues by avoiding

the use of early adopters as the control group. In our study, we estimate treatment effects for all counties annually, while considering only those estimates as the final results for counties assigned the treatment within the respective year. Specifically, in 2015, we aggregate monthly property crime counts for the 12 months following December 2015, calculating the crime rate for each county. Counties receiving treatment during 2015 are denoted with a treatment indicator of 1, while others are marked as 0. Then treatment effects are computed for all counties, with only those counties with treatment indicator equal to 1 being recorded in the final result. In 2016, counties previously treated are excluded, and a similar methodology is applied to estimate treatment effects for counties treated in that year. In the final year, 2017, we include all remaining counties from the 2016 analysis and estimate treatment effects collectively alongside the aggregated crime rates up to 2018.

We then proceed to implement both of our proposed methods with the preprocessed data. We first construct a K-nearest neighbor graph with $K = 5$ based on the observed covariates and use the corresponding oriented incidence matrix to estimate heterogeneous treatment effects for both methods. Since propensity scores are unknown at each location in this case, we estimate the propensity scores via graph fused lasso in our first method. The benchmark competitor chosen for this analysis is causal forests, with the same tuning setup as in the simulation experiments.

We begin by comparing the estimated treatment effects derived from three models considered. In Figure 4.2, we present estimations from two of our proposed models alongside the estimation from causal forest. These plots reveal a consistent overall pattern regarding the causal effects of implementing background check laws on TNCs across the United States. We observe positive treatment effects for the majority of the counties in the United States, indicating that these regions suffer from an increase in property crimes after background check enforcement. In contrast, negative treatment effects, equivalent to a decrease in property crimes, are observed mainly in California and New England Regions, which are heavily populated regions. Such effects are also observed in metropolitan regions across different states in the United States. These findings reveal the nuanced impact of background check

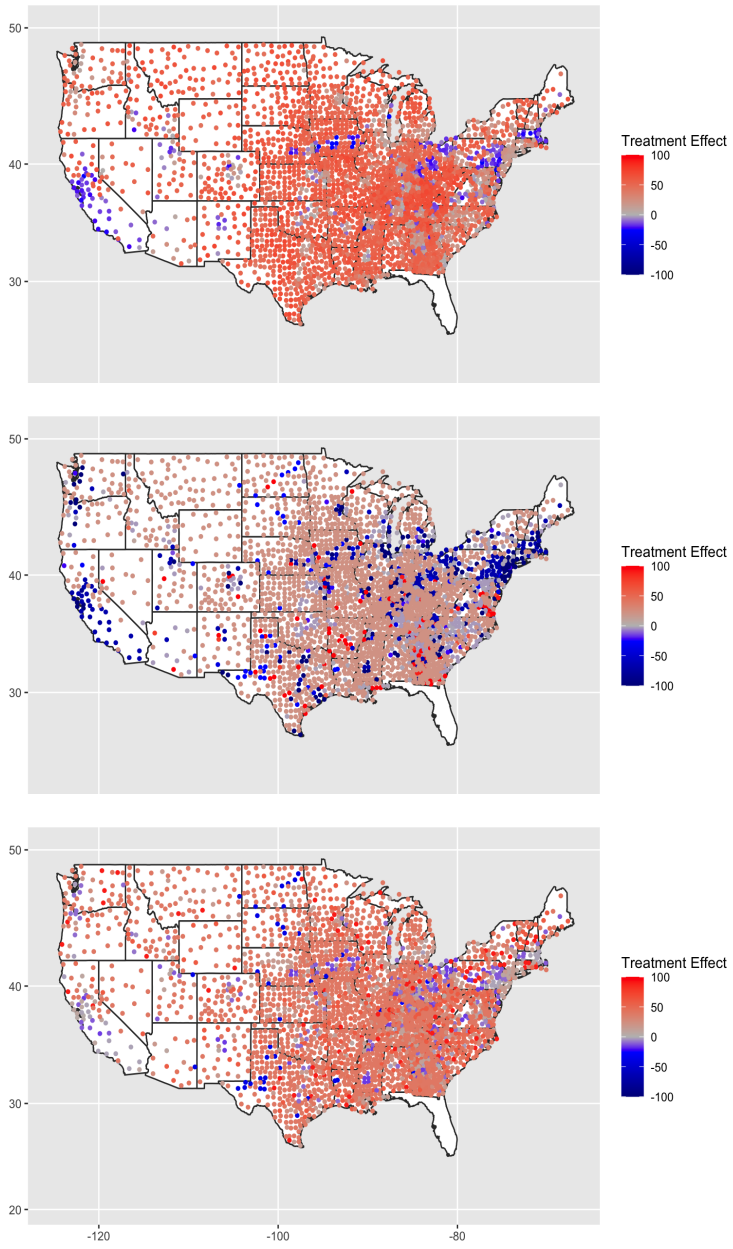

Figure 4.2: Estimated treatment effects of implementing background check laws on TNCs in reducing property crimes across the United States. The horizontal axis represents longitude, and the vertical axis represents latitude. *Top*: estimates from causal graph fused lasso; *Middle*: estimates from causal covairate graph fused lasso; *Bottom*: estimates from causal forests. The color scale displays the unit change in the number of property crimes per 100,000 population.

laws on property crime rates in different regions. While some regions may witness immediate reductions in overall crime rates following such measures, others may face challenges or unintended consequences, including an upsurge in property crimes. Importantly, the positive treatment effects found in our analysis corroborate our hypothesis regarding the unintended negative consequences of implementing background check laws on TNC-related gig workers, and echo concerns raised in studies such as [RCL24], highlighting the need for policymakers to thoroughly evaluate potential social risks when enacting background check laws.

From Figure 4.2, we observe a difference between the two illustrations derived from our proposed methods, which stems from differences in model setups. The causal covariate graph fused lasso assumes a linear representation in the main effect, which may not align perfectly with the real data, potentially leading to some misspecification. Despite this, another notable observation from Figure 4.2 is that both estimations from our proposed methods demonstrate a greater local adaptivity to capture nuanced patterns in regions compared to the estimation from causal forests. This enhanced adaptability highlights the strength of our methods in capturing detailed variations in treatment effects. Such difference stems from the assumptions inherent in our proposed methods, which presume a piecewise constant pattern in treatment effects along a graph structure. By employing a K-NN graph in our estimation, we typically ensure that treatment effects in neighboring counties are similar or identical. This approach may prove beneficial for policymakers seeking to implement consistent treatments across specific regions, rather than adopting more varied measures in individual regions.

4.6.3 Oklahoma Legislation Impact Analysis

In the second part of our analysis, we shift our focus to a specific state, Oklahoma, to investigate whether legislation that eases restrictions on background check laws has an impact on the incidence of property crimes. The Oklahoma Transportation Network Company Services Act, outlined in 47 O.S. Sections 1010-1030, designates the Oklahoma Corporation Commission as the regulatory authority for Transportation Network Companies (TNCs) operating within the state. This legislation came into effect on July 1, 2015, with Section 1019

mandating that TNC companies conduct a criminal background check for each applicant. However, a consequential legislative action followed: Oklahoma's State Question 780 (SQ 780, enacted on July 1, 2017). State Question 780 reclassified certain drug possessions from felonies to misdemeanors, so this law provided some relief on background checks for gig workers, potentially enabling more of them to participate in TNC services.

Now, our objective is to investigate the causal impact of legislation regarding background checks on gig workers on the incidence of property crimes from a more localized perspective, using neighborhood-level data from Oklahoma. Specifically, we focus on zip code-level data, extracting 323 data points representing different neighborhoods in Oklahoma from the nationwide analysis. We use the same set of features as in the previous analysis, and we preprocess the data similarly. We utilize a treatment reversal design to evaluate the causal effects of implementing 47 O.S. and SQ 780 separately. We select treatment indicator time points based on the dates when each law became effective. For each time point, we compute the annual property crime rate by aggregating the total number of property crimes occurring within a 12-month period following the respective time point and then calculating the corresponding crime rate. After data preparation, we apply both of our proposed methods using a K-NN graph with $K = 7$ to estimate treatment effects.

The visualization of estimation results is presented in Figure 4.3. Following the implementation of 47 O.S. in July 2015, mandating background checks for gig drivers, our analysis reveals predominantly positive treatment effects across most neighborhoods in Oklahoma, as displayed in the left column of the figure. However, subsequent to the enactment of SQ 780 two years later, which relaxed the definition of misdemeanors, thereby permitting previously disqualified residents to engage more in gig work, we observe a notable shift in the estimated treatment effects. Specifically, there is a uniform decrease in the value of treatment effects throughout the state after the implementation of SQ 780, as displayed in all three methods applied. Our methods exhibit a greater capacity for highlighting such differences compared to causal forests, demonstrating their superiority in local adaptivity.

Such observations suggest that implementing very strict background check laws for gig

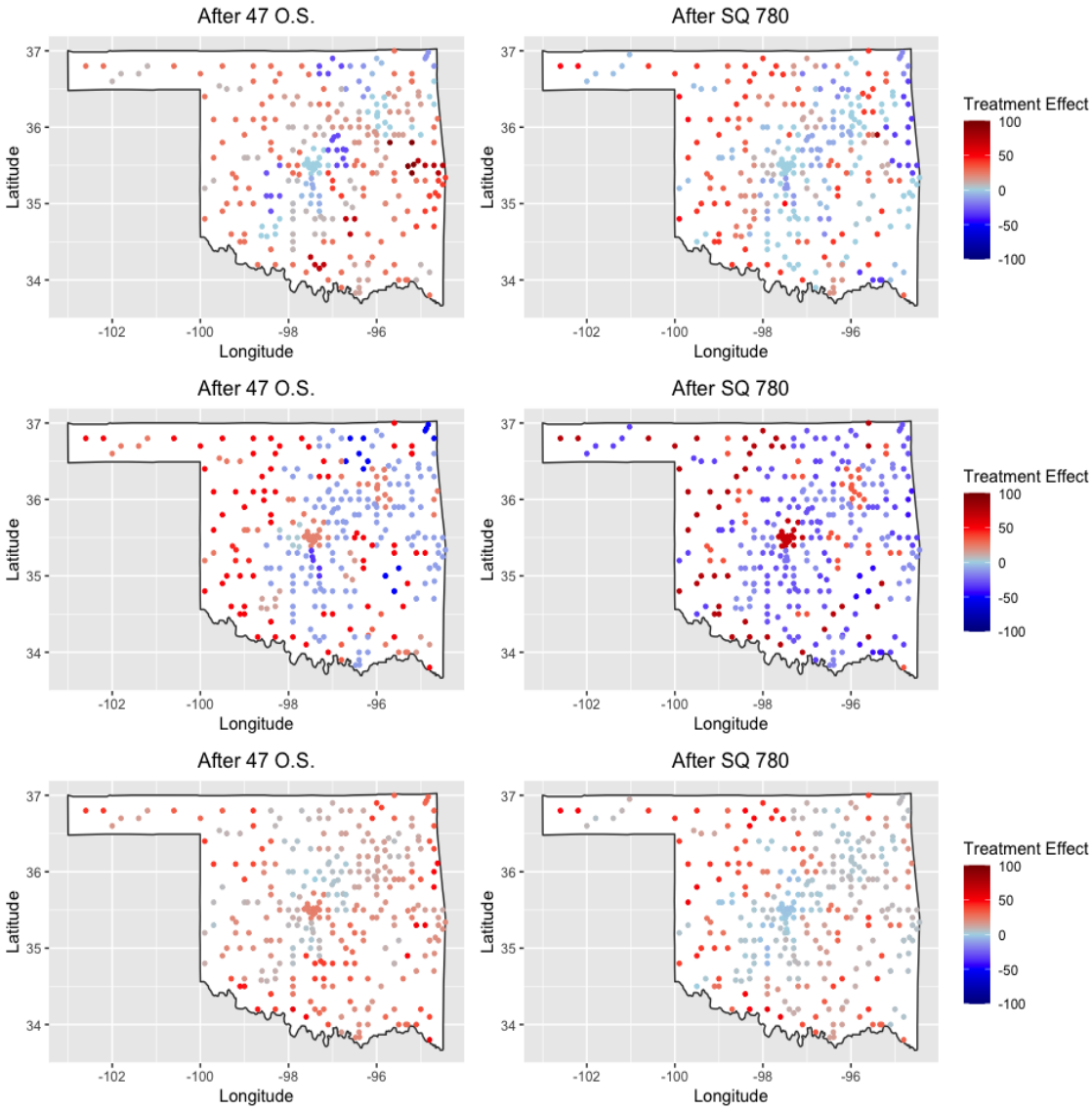

Figure 4.3: Estimated heterogeneous treatment effects of introducing background check laws on Transportation Network Companies (TNC) in Oklahoma. The left column is the estimated treatment effects after the implementation of 47 O.S., and the right column represents the effects after SQ 780 being enforced. *Top:* estimates from causal graph fused lasso; *Middle:* estimates from causal covariate graph fused lasso; *Bottom:* estimates from causal forests. The color scale displays the unit change in the number of property crimes per 100,000 population.

workers in certain regions may lead to unintended negative effects on local crime reduction, particularly arousing concerns about an increase in property crimes. However, with a relaxation in background check regulations, such negative impacts appear to be mitigated, indicating that promoting employment opportunities for gig workers could potentially prevent involvement in criminal activities such as robbery and theft for a living. These findings again support our initial hypothesis that restricting individuals from engaging in gig work through background check laws correlates with an increase in property crimes, and conversely, enabling greater participation in this sector appears to yield the opposite effect.

4.6.4 Summary

In summary, our research provides valuable insights for policymakers aiming to address crime reduction by implementing background check regulations for gig workers in transportation network services. We discover that while metropolitan areas could potentially benefit from a corresponding decrease in property crimes, the majority of regions across the United States might experience a rise in property crimes following the enforcement of background checks. Through a detailed examination of Oklahoma State's data, we identify that easing background check requirements to allow broader participation in TNC services could help mitigate the unintended negative effects of these laws. These findings show the complex dynamics involved in the impact of exclusion-based solutions, such as mandatory background checks, on crime reduction efforts. Policymakers should recognize the evident necessity for further assessment of the unintended social welfare consequences associated with excluding marginalized individuals from opportunities in the gig economy within different regions.

4.7 Conclusion and Discussion

In this chapter, we present two novel methodologies for estimating heterogeneous treatment effects using the fused lasso over graph. The first method explores treatment effect heterogeneity through local patterns captured directly from the graph structure, bypassing

traditional covariate-based approaches. The second method addresses the practical need to assess the impact of individual covariate features while estimating treatment effects concurrently. Our methodologies are well-suited for analyzing data characterized by piecewise constant local structures, commonly observed in real-world datasets. This adaptability is particularly advantageous in causal inference, enabling policymakers to devise targeted interventions based on identified local patterns across geographic or macroeconomic features.

Empirical validation through numerical experiments on simulated data has highlighted the robustness of our estimators, showing superior performance over other popular methods in specific data scenarios. In theoretical analysis on our proposed estimators, we have demonstrated a fast convergence rate of order $n^{-2/3}$ when using a general graph for estimating heterogeneous treatment effects and $n^{-1/d}$ when utilizing a K-nearest-neighbor graph under a multivariate setup. Moreover, our case study on transportation network crime data provides valuable insights for policymakers to reveal how background check laws on TNC gig workers influence changes in property crime rates across different regions and arouse attention on unintended negative consequences of the exclusion-based solution.

While our proposed methodologies demonstrate great performance in estimating heterogeneous treatment effects with a known data structure graph, a worthwhile avenue for future research lies in exploring and extending methods that are flexible and adaptive in detecting local patterns when a graph is not available. The development of such methodologies is important to understand treatment effect heterogeneity in diverse data settings.

4.8 Appendix

4.8.1 The ADMM Algorithm for Solving (4.18)

We reformulate the equation (4.18) as:

$$\begin{aligned} &\underset{\tau\in\mathbb{R}^n, w\in\mathbb{R}^n}{\text{minimize}} \ \frac{1}{2}\sum_{i=1}^{n}(Y_i - X_i^\top\hat{\beta} - Z_i\tau_i)^2 + \lambda_\tau\|Dw\|_1 \\ &\text{s.t.} \qquad\qquad w = \tau, \end{aligned}$$

and we write the corresponding augmented Lagrangian as

$$L_R(\tau, w, u) = \frac{1}{2}\sum_{i=1}^{n}(Y_i - X_i^\top\hat{\beta} - Z_i\tau_i)^2 + \lambda_\tau\|Dw\|_1 + \frac{R}{2}\|\tau - w + u\|_2^2$$

for $R > 0$ with a default setting $R = 0.5$. We then solve this problem by iteratively updating the primal:

$$\tau \leftarrow \underset{\tau\in\mathbb{R}^n}{\arg\min} \left\{\frac{1}{2}\sum_{i=1}^{n}(Y_i - X_i^\top\hat{\beta} - Z_i\tau_i)^2 + \frac{R}{2}\|\tau - w + u\|_2^2\right\},$$

and the dual:

$$w \leftarrow \underset{w\in\mathbb{R}^n}{\arg\min} \left\{\frac{1}{2}\|\tau + u - w\|_2^2 + \frac{\lambda_\tau}{R}\|Dw\|_1\right\}.$$

We solve the primal problem coordinate-wise, following a similar approach as in (4.17), to obtain

$$\tau_i = \frac{R(w_i - u_i) + (Z_i - \hat{e}_i)(Y_i - \hat{m}_i)}{(Z_i - \hat{e}_i)^2 + R},$$

for $i = 1, ..., n$, and we solve the dual problem by the parametric max-flow algorithm [CD09].

4.8.2 TNC Crime Data Source and Descriptions

The data utilized in the analysis of TNC-related crime in Section 4.6 was sourced from various authoritative databases, each contributing specific metrics relevant to our study. The total population was retrieved from the US Census Bureau's data tables. The unemployment rate data came from the US Bureau of Labor Statistics. Poverty rates and other demographic details such as the proportion of elderly residents, male and female populations, and racial

composition (white and Asian residents) were also sourced from the US Census Bureau. Additionally, the annual real GDP growth rates were sourced from the US Bureau of Economic Analysis. For detailed information on the data sources, please see the provided table below.

Variable name	Description	Source	Note
total_pop	total population	https://data.census.gov/table	US Census Bureau
unemployment_r	unemployment rate	https://www.bls.gov/lau/#tables	US Bureau of Labor Statistics
poverty	poverty percentage	https://www.census.gov/data/datasets	US Census Bureau
gdp	annual real GDP growth rate	https://www.bea.gov/taxonomy/term/971	US Bureau of Economic Analysis
p_age_65over	proportion of elderly (65+) residents	https://data.census.gov/table	US Census Bureau
p_male	proportion of male residents	https://data.census.gov/table	US Census Bureau
p_white	proportion of white residents	https://data.census.gov/table	US Census Bureau
p_asian	proportion of Asian residents	https://data.census.gov/table	US Census Bureau

4.8.3 Sub-Gaussian Random Variables

In this section, we provide the detailed definition of sub-Gaussian random variables. A random variable X is said to be *sub-Gaussian* if its tail probabilities satisfied the following bound: there exists a positive parameter $\sigma > 0$ such that for all $t > 0$,

$$\mathbb{P}(|X| \geq t) \leq 2 \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

4.8.4 Proof of Lemma 4.2.1

Proof. Recall the definition of the conditional response surfaces $\mu_i(Z_i) = \mathbb{E}[Y_i \mid G, Z_i]$, then consider the conditional expectation:

$$\begin{aligned}
\mathbb{E}[\epsilon_i \mid G, Z_i] &= \mathbb{E}[Y_i(Z_i) - [\mu_i(0) + Z_i\tau_i^*] \mid G, Z_i] \\
&= \mathbb{E}[Y_i(Z_i) \mid G, Z_i] - \mathbb{E}[\mu_i(0) + Z_i\tau_i^* \mid G, Z_i] \\
&= \mathbb{E}[Y_i \mid G, Z_i] - \mu_i(0) - Z_i\tau_i^* \\
&= \mu_i(Z_i) - \mu_i(0) - Z_i\tau_i^* \\
&= 0,
\end{aligned}$$

where the last equality holds because $\mu_i(Z_i) = \mu_i(0) + Z_i\tau_i^*$ by definition.

Now, following the definition $m_i^* = \mathbb{E}[Y_i \mid G]$, we use the response surface decomposition $Y_i = \mu_i(0) + Z_i\tau_i^* + \epsilon_i$ and take expectations conditional on G to get

$$\begin{aligned} m_i^* &= \mathbb{E}[\mu_i(0) + Z_i\tau_i^* + \epsilon_i \mid G] \\ &= \mu_i(0) + \mathbb{E}[Z_i\tau_i^* \mid G] + \mathbb{E}[\epsilon_i \mid G]. \end{aligned}$$

Since $\mathbb{E}[\epsilon_i \mid G, Z_i] = 0$, it follows that $\mathbb{E}[\epsilon_i \mid G] = 0$, leading to

$$m_i^* = \mu_i(0) + e_i^*\tau_i^*.$$

Then we can rewrite the treatment function τ_i^* in (4.2) in terms of m_i^*, which leads to the decomposition

$$Y_i - m_i^* = (Z_i - e_i^*)\tau_i^* + \epsilon_i.$$

. □

4.8.5 Theorem 4.4.1 and Corollary 4.4.2

4.8.5.1 Proof of Theorem 4.4.1

Proof. From the optimality in the definition of $\hat{\tau}$ in (4.4), we have that

$$\frac{1}{2}\sum_{i=1}^{n}(Y_i - \hat{m}_i - (Z_i - \hat{e}_i)\hat{\tau}_i)^2 + \lambda_\tau\|D\hat{\tau}\|_1 \leq \frac{1}{2}\sum_{i=1}^{n}(Y_i - \hat{m}_i - (Z_i - \hat{e}_i)\tau_i^*)^2 + \lambda_\tau\|D\tau^*\|_1,$$

and so

$$\begin{aligned} \frac{1}{2}\sum_{i=1}^{n}(Z_i - \hat{e}_i)^2(\tau_i^* - \hat{\tau}_i)^2 &\leq \sum_{i=1}^{n}(Y_i - \hat{m}_i - (Z_i - \hat{e}_i)\tau_i^*)(Z_i - \hat{e}_i)(\hat{\tau}_i - \tau_i^*) + \lambda_\tau\|D\tau^*\|_1 - \lambda_\tau\|D\hat{\tau}\|_1 \\ &= T_1 + T_2, \end{aligned} \tag{4.20}$$

where we define

$$T_1 := \sum_{i=1}^{n}\epsilon_i(Z_i - \hat{e}_i)(\hat{\tau}_i - \tau_i^*) + \lambda_\tau\|D\tau^*\|_1 - \lambda_\tau\|D\hat{\tau}\|_1,$$

and

$$T_2 := \sum_{i=1}^{n} (Z_i - \hat{e}_i)(m_i^* - \hat{m}_i - (e_i^* - \hat{e}_i)\tau_i^*)(\hat{\tau}_i - \tau_i^*).$$

Note that there exists a constant $c_\tau > 0$ such that

$$\begin{aligned} T_2 &\leq \|(Z - \hat{e}) \circ (m^* - \hat{m} - (e^* - \hat{e}) \circ \tau^*)\|_2 \|\hat{\tau} - \tau^*\|_2 \\ &\leq \|(Z - \hat{e}) \circ (m^* - \hat{m})\|_2 \|\hat{\tau} - \tau^*\|_2 + \|(Z - \hat{e}) \circ ((e^* - \hat{e}) \circ \tau^*)\|_2 \|\hat{\tau} - \tau^*\|_2 \\ &\leq \|m^* - \hat{m}\|_2 \|\hat{\tau} - \tau^*\|_2 + c_\tau \|e^* - \hat{e}\|_2 \|\hat{\tau} - \tau^*\|_2, \end{aligned}$$

where the first inequality follows from Cauchy-Schwarz inequality, the second from the triangle inequality, and the last from the facts $|Z_i - \hat{e}_i| \leq 1$ for $i = 1, ..., n$, and $\|\tau^*\|_\infty = O_{\mathbb{P}}(1)$.

Then by assumption, there exists some constant $C > 0$ and a sufficiently small $\varepsilon > 0$ such that

$$P(\Omega_1 \cap \Omega_2) \geq 1 - \varepsilon,$$

where

$$\Omega_1 := \left\{ \frac{1}{n} \|\hat{m} - m^*\|_2^2 \leq C r_{m^*} \right\}, \ \Omega_2 := \left\{ \frac{1}{n} c_\tau^2 \|\hat{e} - e^*\|_2^2 \leq C r_{e^*} \right\}.$$

Hence, the following inequality holds:

$$T_2 \leq \sqrt{Cn} \left[\sqrt{r_{e^*}} + \sqrt{r_{m^*}} \right] \cdot \|\hat{\tau} - \tau^*\|_2. \tag{4.21}$$

Furthermore, if we let the constant $C_e = \frac{1}{\min\{e_{\min}, 1 - e_{\max}\}^2}$, we have that

$$\frac{1}{2} \sum_{i=1}^{n} (\hat{\tau}_i - \tau_i^*)^2 \leq \frac{C_e}{2} \sum_{i=1}^{n} (Z_i - e_i^*)^2 (\hat{\tau}_i - \tau_i^*)^2. \tag{4.22}$$

If we assume

$$\frac{1}{4} \|\hat{\tau} - \tau^*\|_2 > C_e \sqrt{Cn} \left[\sqrt{r_{e^*}} + \sqrt{r_{m^*}} \right],$$

then under the events Ω_1 and Ω_2,

$$C_e T_2 \leq \frac{1}{4} \|\hat{\tau} - \tau^*\|_2^2. \tag{4.23}$$

From (4.20), (4.21), (4.22) and (4.23), we obtain that

$$\frac{1}{2} \|\hat{\tau} - \tau^*\|_2^2 \leq C_e T_1 + \frac{1}{4} \|\hat{\tau} - \tau^*\|_2^2,$$

and thereby

$$\|\hat{\tau} - \tau^*\|_2^2 \leq 4C_e T_1.$$

Note that the factor $Z_i - \hat{e}_i$ does not alter the scaling behavior of T_1, therefore $T_1 = \Theta(T_1')$ if we define

$$T_1' := \sum_{i=1}^{n} \epsilon_i(\hat{\tau}_i - \tau_i^*) + \lambda_\tau \|D\tau^*\|_1 - \lambda_\tau \|D\hat{\tau}\|_1.$$

Then from the proof of Theorem 3 in [PSS18], we reach at the following result:

$$\frac{1}{n}\|\hat{\tau} - \tau^*\|_2^2 = O_{\mathbb{P}}\left(\frac{1}{n} + \frac{\|D\tau^*\|_1^{2/3}}{n^{2/3}}\right).$$

On the other hand, if we assume

$$\frac{1}{4}\|\hat{\tau} - \tau^*\|_2 \leq C_e\sqrt{Cn}\left[\sqrt{r_{e^*}} + \sqrt{r_{m^*}}\right],$$

then the inequality $(a+b)^2 \leq 2(a^2+b^2)$ implies

$$\|\hat{\tau} - \tau^*\|_2^2 \leq 32C_e^2 Cn\left[r_{e^*} + r_{m^*}\right],$$

and then the claim follows by combining the above two cases. □

4.8.6 Proof of Corollary 4.4.2

Proof. Following the same reasoning as in the proof of Theorem 4.4.1, with all assumptions in the claim being held, we have that

$$\frac{1}{n}\|\hat{m} - m^*\|_2^2 = O_{\mathbb{P}}\left(\frac{1}{n} + \frac{\|Dm^*\|_1^{2/3}}{n^{2/3}}\right),$$

if $\hat{m}$ is estimated from the optimization problem (2.10), and

$$\frac{1}{n}\|\hat{e} - e^*\|_2^2 = O_{\mathbb{P}}\left(\frac{1}{n} + \frac{\|De^*\|_1^{2/3}}{n^{2/3}}\right),$$

if $\hat{e}$ is estimated from (4.12). Hence, the claim holds by combining the results above and Theorem 4.4.1. □

4.8.7 Definition of Piecewise Lipschitz

To assure that Assumption 17 is valid, we require a piecewise Lipschitz condition on the true function τ^*. In this section, we provide the detailed definition of the class of piecewise Lipschitz functions, followed from Definition 1 in [PSC20]. All notations follow the same from the main context, besides an extra notation on the boundary of a set A, denoted by ∂A.

Definition 4.8.1 *Let $\Omega_\epsilon := [0,1]^d \backslash B_\epsilon(\partial[0,1]^d)$. We say that a bounded function $g : [0,1]^d \to \mathbb{R}$ is piecewise Lipschitz if there exists a set $\mathcal{S} \subset (0,1)^d$ that has the following properties:*

- *The set $\mathcal{S}$ has Lebesgue measure zero.*
- *For some constants $C_\mathcal{S}, \epsilon_0 > 0$, we have that $\mu(h^{-1}\{B_\epsilon(\mathcal{S}) \cap ([0,1]^d \backslash \Omega_\epsilon)\}) \leq C_\mathcal{S}\epsilon$ for all $0 < \epsilon < \epsilon_0$.*
- *There exists a positive constant L_0 such that if z and z' belong to the same connected component of $\Omega_\epsilon \backslash B_\epsilon(\mathcal{S})$, then $|g(z) - g(z')| \leq L_0 \|z - z'\|_2$.*

Roughly speaking, a bounded function g is piecewise Lipschitz if there exists a small set $\mathcal{S}$ that partitions $[0,1]^d$ in such a way that g is Lipschitz within each connected component of the partition. Theorem 2.2.1 in [Zie12] implies that if g is piecewise Lipschitz, then g has bounded variation on any open set within a connected component.

4.8.8 Theorem 4.4.3

4.8.8.1 Auxiliary Lemmas

Lemma 4.8.2 *Let $d_{\min}$ and $d_{\max}$ denote the minimum and maximum degrees, respectively, of a K-nearest neighbor graph G constructed on n nodes. Define the event*

$$\Omega = \{c_1 K \leq d_{\min} \leq d_{\max} \leq c_2 K\}$$

for some positive constants $c_1, c_2 > 0$. Then, as $n \to \infty$, the probability of this event approaches 1:

$$\mathbb{P}(\Omega) \to 1 \quad as \quad n \to \infty.$$

Proof. See the proof of Part 2 of Proposition 30 in [LRH14].

Lemma 4.8.3 *Suppose that Assumptions 10 and 14-16 hold, and inherit all definitions from Lemma 4.8.2. For each data point X_i in the set $\{X_1, \ldots, X_n\}$, define $C_K(X_i)$ as the subset of its K-nearest neighbors that belong to the treated group:*

$$C_K(X_i) = \{j : (i,j) \in E_K, Z_j = 1\}.$$

Let

$$C_{K,\min} = \min_{i=1,\ldots,n} |C_K(X_i)|,$$

then

$$\mathbb{P}\left(C_{K,\min} \leq \frac{1}{2} c_1 K e_{\min}\right) \leq n \exp\left(-\frac{c_1^2 K e_{\min}^2}{2c_2}\right) + \mathbb{P}(\Omega^c).$$

Proof. Let $a > 0$ satisfy the condition

$$a \leq \frac{1}{2} c_1 K e_{\min}. \tag{4.24}$$

Then, the following inequalities hold:

$$\begin{aligned}
\mathbb{P}(C_{K,\min} \leq a) &\leq \mathbb{P}(C_{K,\min} \leq a \mid \Omega) + \mathbb{P}(\Omega^c) \\
&\leq n \max_{i=1,\ldots,n} \mathbb{P}(\{C_{K,\min} \leq a\} \cap \Omega) + \mathbb{P}(\Omega^c) \\
&\leq n \max_{i=1,\ldots,n} \sum_{l=\lceil c_1 K \rceil}^{\lfloor c_2 K \rfloor} \sum_{\substack{i_1,\ldots,i_l \\ \subset \{1,\ldots,n\} \setminus i}} \mathbb{P}\Big(C_{K,\min} \leq a \mid X_{i_1}, \ldots, X_{i_l} \in E_K(X_i), \\
&\qquad X_j \notin E_K(X_i)\ \forall j \notin \{i_1, \ldots, i_l\}\Big) \cdot \mathbb{P}(X_{i_1}, \ldots, X_{i_l} \in E_K(X_i), \\
&\qquad X_j \notin E_K(X_i)\ \forall j \notin \{i_1, \ldots, i_l\}) + \mathbb{P}(\Omega^c),
\end{aligned} \tag{4.25}$$

where the first inequality follows from the law of total probability, and the second from the union bound.

Next, let l be such that $c_1 K \leq l \leq c_2 K$. Conditioning on the event

$$E_{i,i_1,\ldots,i_l} := \{X_{i_1}, \ldots, X_{i_l} \in E_K(X_i) \text{ and } X_j \notin E_K(X_i) \text{ for all } j \notin \{i_1, \ldots, i_l\}\},$$

we obtain

$$|C_K(X_i)| = \sum_{j \in \{i_1,\ldots,i_l\}} Z_j.$$

Since

$$\mathbb{E}\left(\sum_{j \in \{i_1,\ldots,i_l\}} Z_j \mid E_{i,i_1,\ldots,i_l}\right) \geq l e_{\min} \geq c_1 K e_{\min},$$

applying Hoeffding's inequality with the choice of a from (4.24), we get

$$\begin{aligned}
\mathbb{P}(|C_K(X_i)| \leq a \mid E_{i,i_1,\ldots,i_l}) &\leq \mathbb{P}\left(|C_K(X_i)| \leq \mathbb{E}(|C_K(X_i)| \mid E_{i,i_1,\ldots,i_l}) - \frac{1}{2} c_1 K e_{\min} \mid E_{i,i_1,\ldots,i_l}\right) \\
&\leq \exp\left(-\frac{2(c_1 K e_{\min}/2)^2}{l}\right) \\
&\leq \exp\left(-\frac{c_1^2 K e_{\min}^2}{2c_2}\right).
\end{aligned}$$

Combining this result with (4.25), we obtain

$$\begin{aligned}
\mathbb{P}\left(C_{K,\min} \leq \frac{1}{2} c_1 K e_{\min}\right) &\leq n \max_{i=1,\ldots,n} \sum_{l=\lceil c_1 K \rceil}^{\lfloor c_2 K \rfloor} \sum_{\substack{i_1,\ldots,i_l \\ \subset \{1,\ldots,n\} \setminus i}} \exp\left(-\frac{c_1^2 K e_{\min}^2}{2c_2}\right) \cdot \mathbb{P}(E_{i,i_1,\ldots,i_l}) + \mathbb{P}(\Omega^c) \\
&\leq n \exp\left(-\frac{c_1^2 K e_{\min}^2}{2c_2}\right) \mathbb{P}(\Omega) + \mathbb{P}(\Omega^c) \\
&\leq n \exp\left(-\frac{c_1^2 K e_{\min}^2}{2c_2}\right) + \mathbb{P}(\Omega^c). \qquad \square
\end{aligned}$$

Lemmas 4.8.2 and 4.8.3 together imply that if the number of nearest neighbors, K, is chosen such that $K/\log n \to \infty$, then for a randomly selected covariate, the probability that the minimum number of treated neighbors across all data points falls below a threshold of order K approaches zero as the sample size n grows to infinity. Consequently, we conclude that, with high probability, the number of treated neighbors for any given observation scales proportionally to K.

Lemma 4.8.4 *Suppose that Assumptions 10-17 hold. We also assume that* $\max(\|X\beta^*\|_\infty, \|\tau^*\|_\infty) = O_{\mathbb{P}}(1)$, $\max_{i\in\{1,\ldots,n\},j\in\{1,\ldots,d\}} |X_{ij}| = C_X$ *for some positive constant* C_X, *and* $Z_i \perp\!\!\!\perp \epsilon_i$ *for* $i = 1, ..., n$. *Consider the first iteration in Algorithm 4 with the K-NN graph G. If an initialized* $\hat{\beta}^{(0)}$ *satisfies the condition* $\|\hat{\beta}^{(0)}\|_\infty \leq M$ *for some positive constant* M, *then the estimated treatment effect* $\hat{\tau}^{(1)}$ *obeys* $\|\hat{\tau}^{(1)}\|_\infty = O_{\mathbb{P}}(\text{poly}(\log n))$.

Proof. Notice that given the initialized $\hat{\beta}^{(0)}$, $\hat{\tau}^{(1)}$ is the solution to the following optimization problem

$$\hat{\tau}^{(1)} = \underset{\tau\in\mathbb{R}^n}{\arg\min}\ \frac{1}{2}\sum_{i=1}^{n}(Y_i - X_i^\top\hat{\beta}^{(0)} - Z_i\tau_i)^2 + \lambda_\tau\|\nabla_G\tau\|_1. \tag{4.26}$$

The model setup (4.6) implies that

$$\begin{aligned}\|Y\|_\infty &\leq \|X\beta^*\|_\infty + \|\tau^*\|_\infty + \|\epsilon\|_\infty \\ &= O_{\mathbb{P}}(\log^{1/2}(n)),\end{aligned}$$

where $\|\epsilon\|_\infty = O_{\mathbb{P}}(\log^{1/2}(n))$ follows from Gaussian maximal inequality. Then from the triangle inequality, the following claim holds:

$$\begin{aligned}\|Y - X\hat{\beta}^{(0)}\|_\infty &\leq \|Y\|_\infty + \max_{1\leq i\leq n}\sum_{j=1}^{d}|X_{ij}|\cdot\|\hat{\beta}^{(0)}\|_\infty \\ &\leq \|Y\|_\infty + d\cdot\max_{i,j}|X_{ij}|\cdot\|\hat{\beta}^{(0)}\|_\infty \\ &\leq \|Y\|_\infty + d\cdot C_X\cdot M \\ &= O_{\mathbb{P}}(\text{poly}(\log n)).\end{aligned}$$

Suppose that

$$\max_{i=1,\ldots,n}|Y_i - X_i^\top\hat{\beta}^{(0)}| \leq c_1\cdot\text{poly}(\log n)$$

for some constant $c_1 > 0$ with high probability.

Let $\tilde{\tau}_i^{(1)}$ be the projection of $\hat{\tau}_i^{(1)}$ onto $[-c_1\cdot\text{poly}(\log n), c_1\cdot\text{poly}(\log n)]$. Using the properties of the projection in real line, we have that the following two inequalities hold at the same time:

$$\frac{1}{2}\sum_{i=1}^{n}(Y_i - X_i\hat{\beta}^{(0)} - Z_i\tilde{\tau}_i^{(1)})^2 \leq \frac{1}{2}\sum_{i=1}^{n}(Y_i - X_i\hat{\beta}^{(0)} - Z_i\hat{\tau}_i^{(1)})^2, \tag{4.27}$$

$$\sum_{\{i,j\}\in E} |\tilde{\tau}_i^{(1)} - \tilde{\tau}_j^{(1)}| \leq \sum_{\{i,j\}\in E} |\hat{\tau}_i^{(1)} - \hat{\tau}_j^{(1)}|. \tag{4.28}$$

Thereby,

$$\frac{1}{2}\sum_{i=1}^{n}(Y_i - X_i\hat{\beta}^{(0)} - Z_i\tilde{\tau}_i^{(1)})^2 + \lambda_\tau \|\nabla_G \tilde{\tau}^{(1)}\|_1 \leq \frac{1}{2}\sum_{i=1}^{n}(Y_i - X_i\hat{\beta}^{(0)} - Z_i\hat{\tau}_i^{(1)})^2 + \lambda_\tau \|\nabla_G \hat{\tau}^{(1)}\|_1.$$

If $\tilde{\tau}^{(1)}$ does not violate the optimality of $\hat{\tau}^{(1)}$ in (4.26), then both (4.27) and (4.28) must hold with equality. Since the objective function in (4.27) is strictly convex and thus it has a unique minimizer, it follows that $\hat{\tau}^{(1)} = \tilde{\tau}^{(1)}$. Then we reach at the result in the claim since $\hat{\tau}_i^{(1)} \in [-c_1 \cdot \text{poly}(\log n), c_1 \cdot \text{poly}(\log n)]$ for $i = 1, ..., n$. □

Lemma 4.8.5 *With the notation and assumptions inherited from Lemma 4.8.4, let* $\epsilon = (\epsilon_1, ..., \epsilon_n)^\top \in \mathbb{R}^n$ *be a vector of mean-zero sub-Gaussian random variables with parameter* σ*. For the* K*-NN graph* G*, where* K *is chosen such that* $K/\log n \to \infty$*, there exist positive constants* C_3 *depending on* d *and* C_4 *depending on* d*,* $p_{\max}$*, and* $L_{\min}$ *such that the following holds:*

$$\sup_{\Delta\in\mathbb{R}^n : \|\Delta \circ Z\|_2^2 \leq \eta^2, \|\nabla_G \Delta\|_1 \leq 4\|\nabla_G \tau^*\|_1} \sum_{i=1}^{n} \Delta_i \epsilon_i \leq C_3 \sigma K \log^{1/4}(n)\eta + C_4 \sigma K \log(n) \|\nabla_G \tau^*\|_1,$$

with probability at least

$$1 - \frac{2}{\sqrt{\log n}} - \frac{n}{C_6 K} \exp\left(-\frac{C_5}{3}\frac{K}{\log n}\right) - n\exp(-K/3),$$

where η *is chosen such that*

$$\eta^2 = C_d \left\{ K^2 \log^3(n) + K \log^{\frac{7}{2}}(n) + K^2 \log^{\frac{1}{2}}(n) \|\nabla_G \tau^*\|_1 \right\}, \tag{4.29}$$

with a sufficiently large positive constant C_d *depending only on* d*,* $C_5 > 0$ *is a constant depending on* d*, and* $C_6 > 0$ *is a constant depending on* d*,* $p_{\min}$ *and* $L_{\max}$*. Here,* $p_{\max}$*,* $p_{\min}$*,* $L_{\max}$*, and* $L_{\min}$ *are from Assumptions 14 and 16.*

Proof. See the proofs of Lemma 11 in [PSC20] and Lemma 18 in [PMW23].

Lemma 4.8.6 *With the notation and assumptions inherited from Lemma 4.8.4 and 4.8.5, we define*

$$\mathbb{T} := \{\tau \in \mathbb{R}^n : \tau = t\hat{\tau} + (1-t)\tau^*, t \in [0,1]\}.$$

For a sufficiently small $\varepsilon > 0$ and an appropriate choice of λ_τ satisfying $\lambda_\tau = \frac{\eta^2}{4}\|\nabla_G \tau^*\|_1^{-1}$, we have that

$$\mathbb{P}(\mathcal{E}) \leq \frac{\varepsilon}{4} + \mathbb{P}\left(C_X^2 dn\|\hat{\beta}^{(0)} - \beta^*\|_2^2 \geq \frac{\eta^2}{2}\right),$$

where the event $\mathcal{E}$ is defined by

$$\mathcal{E} := \left\{\sup_{\tau\in\mathbb{T}:\|(\tau-\tau^*)\circ Z\|_2^2 \leq \eta^2} \|\nabla_G \tau\|_1 \geq 5\|\nabla_G \tau^*\|_1\right\},$$

and η^2 is chosen as (4.29) in Lemma 4.8.5.

Proof. By the convexity of the constraint (4.7) and the minimizer property, for any $\tau \in \mathbb{T}$, we have that

$$\begin{aligned}\frac{1}{2}\sum_{i=1}^n Z_i^2(\tau_i - \tau_i^*)^2 &\leq \sum_{i=1}^n Z_i(Y_i - X_i^\top\hat{\beta}^{(0)} - Z_i\tau_i^*)(\tau_i - \tau_i^*) + \lambda_\tau\|\nabla_G\tau^*\|_1 - \lambda_\tau\|\nabla_G\tau\|_1 \\ &= \sum_{i=1}^n Z_i(\tau_i - \tau_i^*)\epsilon_i + \sum_{i=1}^n Z_i(X_i^\top\beta^* - X_i^\top\hat{\beta}^{(0)})(\tau_i - \tau_i^*) + \lambda_\tau\|\nabla_G\tau^*\|_1 - \lambda_\tau\|\nabla_G\tau\|_1.\end{aligned} \tag{4.30}$$

Observe that

$$\begin{aligned}\sum_{i=1}^n Z_i(X_i^\top\beta^* - X_i^\top\hat{\beta}^{(0)})(\tau_i - \tau_i^*) &\leq \max_{i=1,\ldots,n}|X_i^\top\beta^* - X_i^\top\hat{\beta}^{(0)}| \cdot \|(\tau-\tau^*)\circ Z\|_1 \\ &\leq \max_{i=1,\ldots,n}\|X_i^\top\|_2 \cdot \|\hat{\beta}^{(0)} - \beta^*\|_2 \cdot \|(\tau-\tau^*)\circ Z\|_1 \\ &\leq C_X\sqrt{d}\|\hat{\beta}^{(0)} - \beta^*\|_2 \cdot \sqrt{n}\|(\tau-\tau^*)\circ Z\|_2,\end{aligned} \tag{4.31}$$

where the first inequality follows from Hölder's inequality, the second inequality from the Cauchy–Schwarz inequality. Then followed by the inequality $2ab \leq a^2 + b^2$, we get

$$\sum_{i=1}^n Z_i(X_i^\top\beta^* - X_i^\top\hat{\beta}^{(0)})(\tau_i - \tau_i^*) \leq \frac{1}{2}C_X^2 dn\|\hat{\beta}^{(0)} - \beta^*\|_2^2 + \frac{1}{2}\sum_{i=1}^n Z_i^2(\tau_i - \tau_i^*)^2. \tag{4.32}$$

Therefore, from the two displays (4.30) and (4.32) above, we have the following inequality:

$$\|\nabla_G \tau\|_1 \leq \frac{1}{\lambda_\tau} \sum_{i=1}^{n} Z_i(\tau_i - \tau_i^*)\epsilon_i + \frac{C_X^2 dn \|\hat{\beta}^{(0)} - \beta^*\|_2^2}{2\lambda_\tau} + \|\nabla_G \tau^*\|_1.$$

Therefore, for any $\tau \in \mathbb{T}$,

$$\|\nabla_G \tau - \nabla_G \tau^*\|_1 \leq \frac{\mathcal{T}_\epsilon(\tau - \tau^*)}{\lambda_\tau} + \frac{C_X^2 dn \|\hat{\beta}^{(0)} - \beta^*\|_2^2}{2\lambda_\tau} + 2\|\nabla_G \tau^*\|_1, \tag{4.33}$$

where

$$\mathcal{T}_\epsilon(\tau - \tau^*) := \sum_{i=1}^{n} Z_i(\tau_i - \tau_i^*)\epsilon_i.$$

Let $\tau \in \mathbb{T}$ be such that $\|(\tau - \tau^*) \circ Z\|_2^2 \leq \eta^2$. In addition, suppose that

$$\|\nabla_G \tau\|_1 \geq 5\|\nabla_G \tau^*\|_1.$$

By the triangle inequality,

$$\|\nabla_G \tau - \nabla_G \tau^*\|_1 \geq \|\nabla_G \tau\|_1 - \|\nabla_G \tau^*\|_1 = 4\|\nabla_G \tau^*\|_1.$$

Define $\bar{\tau} = t\tau + (1-t)\tau^*$, where

$$t = \frac{4\|\nabla_G \tau^*\|_1}{\|\nabla_G \tau - \nabla_G \tau^*\|_1} \in [0, 1].$$

Then it holds that

$$\|\nabla_G \bar{\tau} - \nabla_G \tau^*\|_1 = t\|\nabla_G \tau - \nabla_G \tau^*\|_1 = 4\|\nabla_G \tau^*\|_1. \tag{4.34}$$

The displays (4.33) and (4.34) imply that

$$2\|\nabla_G \tau^*\|_1 \leq \frac{\mathcal{T}_\epsilon(\bar{\tau} - \tau^*)}{\lambda_\tau} + \frac{C_X^2 dn \|\hat{\beta}^{(0)} - \beta^*\|_2^2}{2\lambda_\tau}. \tag{4.35}$$

By selecting the tuning parameter λ_τ satisfying

$$\lambda_\tau = \frac{\eta^2}{4}\|\nabla_G \tau^*\|_1^{-1}, \tag{4.36}$$

the inequality (4.35) leads to

$$\frac{\eta^2}{2} \leq \mathcal{T}_\epsilon(\bar{\tau} - \tau^*) + \frac{1}{2} C_X^2 dn \|\hat{\beta}^{(0)} - \beta^*\|_2^2. \tag{4.37}$$

The property of $\bar{\tau}$ in (4.37), along with the definition of the event $\mathcal{E}$, indicates that

$$\begin{aligned}\mathcal{E} &\subset \left\{ \sup_{\bar{\tau}\in\mathbb{T}:\|(\bar{\tau}-\tau^*)\circ Z\|_2^2\leq\eta^2,\|\nabla_G\bar{\tau}-\nabla_G\tau^*\|_1\leq 4\|\nabla_G\tau^*\|_1} \mathcal{T}_\epsilon(\bar{\tau}-\tau^*) + \frac{1}{2}C_X^2 dn\|\hat{\beta}^{(0)}-\beta^*\|_2^2 \geq \frac{\eta^2}{2} \right\} \\ &\subset \left\{ \sup_{\Delta\in\mathbb{R}^n:\|\Delta\circ Z\|_2^2\leq\eta^2,\|\nabla_G\Delta\|_1\leq 4\|\nabla_G\tau^*\|_1} \mathcal{T}_\epsilon(\Delta) + \frac{1}{2}C_X^2 dn\|\hat{\beta}^{(0)}-\beta^*\|_2^2 \geq \frac{\eta^2}{2} \right\},\end{aligned}$$

where we denote $\Delta := \bar{\tau} - \tau^*$. Therefore, by a union bound,

$$\begin{aligned}\mathbb{P}(\mathcal{E}) &\leq \mathbb{P}\left(\left\{ \sup_{\Delta\in\mathbb{R}^n:\|\Delta\circ Z\|_2^2\leq\eta^2,\|\nabla_G\Delta\|_1\leq 4\|\nabla_G\tau^*\|_1} \mathcal{T}_\epsilon(\Delta) + \frac{1}{2}C_X^2 dn\|\hat{\beta}^{(0)}-\beta^*\|_2^2 \geq \frac{\eta^2}{2} \right\} \right) \\ &\leq \mathbb{P}\left(\sup_{\Delta\in\mathbb{R}^n:\|\Delta\circ Z\|_2^2\leq\eta^2,\|\nabla_G\Delta\|_1\leq 4\|\nabla_G\tau^*\|_1} \mathcal{T}_\epsilon(\Delta) \geq \frac{\eta^2}{4} \right) + \mathbb{P}\left(C_X^2 dn\|\hat{\beta}^{(0)}-\beta^*\|_2^2 \geq \frac{\eta^2}{2} \right).\end{aligned} \tag{4.38}$$

Denote $\tilde{\epsilon}_i = Z_i\epsilon_i$ for $i = 1, ..., n$. Note that $\tilde{\epsilon}$ is also mean zero sub-Gaussian with parameter σ, following from the assumption $Z_i \perp\!\!\!\perp \epsilon_i$ in Lemma 4.8.4 and the fact $Z_i \in \{0, 1\}$. Then using $V_n = 4\|\nabla_G\tau^*\|_1$ and following from Lemma 4.8.5, we have that for some positive constants C_3 depending on d and C_4 depending on d, $p_{\max}$ and $L_{\min}$,

$$\sup_{\Delta\in\mathbb{R}^n:\|\Delta\circ Z\|_2^2\leq\eta^2,\|\nabla_G\Delta\|_1\leq V_n} \sum_{i=1}^n \Delta_i\tilde{\epsilon}_i \leq C_3\sigma K\log^{1/4}(n)\eta + C_4\sigma K\log(n)\|\nabla_G\tau^*\|_1,$$

with probability at least

$$1 - \frac{2}{\sqrt{\log n}} - \frac{n}{C_6 K}\exp\left(-\frac{C_5}{3}\frac{K}{\log n}\right) - n\exp(-K/3),$$

where $C_5 > 0$ is a constant depending on d and $C_6 > 0$ is a constant depending on d, $p_{\min}$ and $L_{\max}$. With the choice of η^2 as in (4.29), it holds that

$$\mathbb{P}\left(\sup_{\Delta\in\mathbb{R}^n:\|\Delta\circ Z\|_2^2\leq\eta^2,\|\nabla_G\Delta\|_1\leq 4\|\nabla_G\tau^*\|_1} \mathcal{T}_\epsilon(\Delta) \geq \frac{\eta^2}{4} \right) \leq \frac{\varepsilon}{4}. \tag{4.39}$$

Joining the inequalities (4.38) and (4.39) yields that

$$\begin{aligned}\mathbb{P}(\mathcal{E}) &\leq \mathbb{P}\left(\sup_{\Delta\in\mathbb{R}^n:\|\Delta\circ Z\|_2^2\leq\eta^2,\|\nabla_G\Delta\|_1\leq 4\|\nabla_G\tau^*\|_1} \mathcal{T}_\epsilon(\Delta) \geq \frac{\eta^2}{4} \right) + \mathbb{P}\left(C_X^2 dn\|\hat{\beta}^{(0)}-\beta^*\|_2^2 \geq \frac{\eta^2}{2} \right) \\ &\leq \frac{\varepsilon}{4} + \mathbb{P}\left(C_X^2 dn\|\hat{\beta}^{(0)}-\beta^*\|_2^2 \geq \frac{\eta^2}{2} \right).\end{aligned}$$

□

Lemma 4.8.7 *Under the same notation and assumptions as in Lemmas 4.8.4-4.8.6, and additionally assuming that the initialized $\hat{\beta}^{(0)}$ satisfies the condition $\|\hat{\beta}^{(0)} - \beta^*\|_2^2 = O_{\mathbb{P}}(n^{-\alpha})$ for $\alpha \geq \frac{1}{d}$, then with η^2 chosen as in (4.29) of Lemma 4.8.5 and ε defined in Lemma 4.8.6, we have that*

$$\mathbb{P}\left(\|(\hat{\tau}^{(1)} - \tau^*) \circ Z\|_2^2 > \eta^2\right) \leq \varepsilon.$$

Proof. Note that η^2 is chosen to be on the order of $O_{\mathbb{P}}(n^{1-1/d}\text{poly}(\log n))$. Since d is assumed to be fixed, by selecting C_X sufficiently large we obtain that

$$\mathbb{P}\left(C_X^2 dn\|\hat{\beta}^{(0)} - \beta^*\|_2^2 \geq \frac{\eta^2}{2}\right) \leq \frac{\varepsilon}{4}, \tag{4.40}$$

following the assumption on the convergence rate of $\|\hat{\beta}^{(0)} - \beta^*\|_2^2$ in the claim. We can then combine this with the result of Lemma 4.8.6 to establish

$$\mathbb{P}\left(\mathcal{E}\right) \leq \frac{\varepsilon}{2}, \tag{4.41}$$

where we denote the event $\mathcal{E}$ as

$$\mathcal{E} := \sup_{\tau \in \mathbb{T}: \|(\tau - \tau^*) \circ Z\|_2^2 \leq \eta^2} \|\nabla_G \tau\|_1 \geq 5\|\nabla_G \tau^*\|_1.$$

Now, we consider the event $\mathcal{E}^c$. Notice that if $\|(\hat{\tau}^{(1)} - \tau^*) \circ Z\|_2^2 > \eta^2$, there exists $\check{\tau} = t\hat{\tau}^{(1)} + (1-t)\tau^* \in \mathbb{T}$ with $t = \frac{\eta}{\|(\hat{\tau}^{(1)} - \tau^*) \circ Z\|_2}$ such that

$$\|(\check{\tau} - \tau^*) \circ Z\|_2^2 = \eta^2. \tag{4.42}$$

Note that (4.31) also implies that

$$\sum_{i=1}^n Z_i(X_i^\top \beta^* - X_i^\top \hat{\beta}^{(0)})(\tau_i - \tau_i^*) \leq 2C_X^2 dn\|\hat{\beta}^{(0)} - \beta^*\|_2^2 + \frac{1}{8}\sum_{i=1}^n Z_i^2(\tau_i - \tau_i^*)^2. \tag{4.43}$$

Then from (4.30), (4.42) and (4.43), we obtain

$$\sum_{i=1}^n Z_i(\check{\tau}_i - \tau_i^*)\epsilon_i + 2C_X^2 dn\|\hat{\beta}^{(0)} - \beta^*\|_2^2 + \lambda_\tau\|\nabla_G \tau^*\|_1 - \lambda_\tau\|\nabla_G \check{\tau}\|_1 \geq \frac{3}{8}\eta^2.$$

Along with the choice of λ_τ satisfying (4.36) and the fact that $-\lambda_\tau\|\nabla_G \check{\tau}\|_1 \leq 0$, we have that

$$\sum_{i=1}^n Z_i(\check{\tau}_i - \tau_i^*)\epsilon_i + 2C_X^2 dn\|\hat{\beta}^{(0)} - \beta^*\|_2^2 \geq \frac{\eta^2}{8}. \tag{4.44}$$

Therefore,

$$
\begin{aligned}
&\left\{\|(\hat{\tau}^{(1)}-\tau^*)\circ Z\|_2^2>\eta^2\right\}\cap\mathcal{E}^c\\
&\subset\left\{\exists\check{\tau}\in\mathbb{T},\|(\check{\tau}-\tau^*)\circ Z\|_2^2=\eta^2,\sum_{i=1}^n Z_i(\check{\tau}_i-\tau_i^*)\epsilon_i+2C_X^2dn\|\hat{\beta}^{(0)}-\beta^*\|_2^2\geq\frac{\eta^2}{8}\right\}\cap\mathcal{E}^c\\
&\subset\left\{\sup_{\|\Delta\circ Z\|_2^2\leq\eta^2,\|\nabla_G\Delta\|_1\leq 6\|\nabla_G\tau^*\|_1}\mathcal{T}_\epsilon(\Delta)+2C_X^2dn\|\hat{\beta}^{(0)}-\beta^*\|_2^2\geq\frac{\eta^2}{8}\right\},
\end{aligned}
$$

where the first containment is directly followed by (4.44), the second follows from the triangle inequality. In consequence,

$$
\begin{aligned}
&\mathbb{P}\left(\left\{\|(\hat{\tau}^{(1)}-\tau^*)\circ Z\|_2^2>\eta^2\right\}\cap\mathcal{E}^c\right)\\
&\leq\mathbb{P}\left(\sup_{\|\Delta\circ Z\|_2^2\leq\eta^2,\|\nabla_G\Delta\|_1\leq 6\|\nabla_G\tau^*\|_1}\mathcal{T}_\epsilon(\Delta)+2C_X^2dn\|\hat{\beta}^{(0)}-\beta^*\|_2^2\geq\frac{\eta^2}{8}\right)\\
&\leq\mathbb{P}\left(\sup_{\|\Delta\circ Z\|_2^2\leq\eta^2,\|\nabla_G\Delta\|_1\leq 6\|\nabla_G\tau^*\|_1}\mathcal{T}_\epsilon(\Delta)\geq\frac{\eta^2}{16}\right)+\mathbb{P}\left(C_X^2dn\|\hat{\beta}^{(0)}-\beta^*\|_2^2\geq\frac{\eta^2}{32}\right).
\end{aligned}
$$

By the same argument that leads to (4.39) and (4.40), it holds that

$$
\mathbb{P}\left(\sup_{\|\Delta\circ Z\|_2^2\leq\eta^2,\|\nabla_G\Delta\|_1\leq 6\|\nabla_G\tau^*\|_1}\mathcal{T}_\epsilon(\Delta)\geq\frac{\eta^2}{16}\right)\leq\frac{\varepsilon}{4},
$$

and

$$
\mathbb{P}\left(C_X^2dn\|\hat{\beta}^{(0)}-\beta^*\|_2^2\geq\frac{\eta^2}{32}\right)\leq\frac{\varepsilon}{4}.
$$

Hence

$$
\mathbb{P}\left(\left\{\|(\hat{\tau}^{(1)}-\tau^*)\circ Z\|_2^2>\eta^2\right\}\cap\mathcal{E}^c\right)\leq\frac{\varepsilon}{2}. \tag{4.45}
$$

Combining (4.41) and (4.45), we get

$$
\mathbb{P}\left(\|(\hat{\tau}^{(1)}-\tau^*)\circ Z\|_2^2>\eta^2\right)\leq\mathbb{P}\left(\left\{\|(\hat{\tau}^{(1)}-\tau^*)\circ Z\|_2^2>\eta^2\right\}\cap\mathcal{E}^c\right)+\mathbb{P}(\mathcal{E})\leq\varepsilon,
$$

thus concluding the claim. □

Lemma 4.8.8 *Under the same notation and assumptions as in Lemmas 4.8.4-4.8.7, the estimated treatment effect $\hat{\tau}^{(1)}$ from (4.26) satisfies*

$$
\|\nabla_G\hat{\tau}^{(1)}\|_1=O_{\mathbb{P}}(n^{1-1/d}\mathrm{poly}(\log n)).
$$

Proof. From (4.41), we obtain that the following holds with probability at least $1 - \frac{\varepsilon}{2}$ for a sufficiently small $\varepsilon > 0$,

$$\sup_{\tau \in \mathbb{T}: \|(\tau - \tau^*) \circ Z\|_2^2 \leq \eta^2} \|\nabla_G \tau\|_1 \leq 5 \|\nabla_G \tau^*\|_1.$$

Also by Assumption 17, $\|\nabla_G \tau^*\|_1 = O_{\mathbb{P}}(n^{1-1/d} \text{poly}(\log n))$. Therefore, for $\tau \in \mathbb{T}$ satisfying $\|(\tau - \tau^*) \circ Z\|_2^2 \leq \eta^2$, we have that

$$\|\nabla_G \tau\|_1 = O_{\mathbb{P}}\left(n^{1-1/d} \text{poly}(\log n)\right).$$

By Lemma 4.8.7, we have that with probability at least $1 - \varepsilon$, the estimated treatment effect $\hat{\tau}^{(1)}$ from (4.26) satisfies

$$\|(\hat{\tau}^{(1)} - \tau^*) \circ Z\|_2^2 \leq \eta^2,$$

thus completing the proof. □

4.8.8.2 Proof of Theorem 4.4.3

Proof. By the optimality of $\hat{\tau}^{(1)}$ in (4.26), we have that

$$\frac{1}{2} \sum_{i=1}^{n} (Y_i - X_i^\top \hat{\beta}^{(0)} - Z_i \hat{\tau}_i^{(1)})^2 + \lambda_\tau \|\nabla_G \hat{\tau}^{(1)}\|_1 \leq \frac{1}{2} \sum_{i=1}^{n} (Y_i - X_i^\top \hat{\beta}^{(0)} - Z_i \tau_i^*)^2 + \lambda_\tau \|\nabla_G \tau^*\|_1.$$

Then,

$$\begin{aligned} \frac{1}{2n} \sum_{i=1}^{n} Z_i^2 (\hat{\tau}_i^{(1)} - \tau_i^*)^2 &\leq \frac{1}{n} \left\{ \sum_{i=1}^{n} Z_i (Y_i - X_i^\top \hat{\beta}^{(0)} - Z_i \tau_i^*)(\hat{\tau}_i^{(1)} - \tau_i^*) + \lambda_\tau \|\nabla_G \tau^*\|_1 - \lambda_\tau \|\nabla_G \hat{\tau}^{(1)}\|_1 \right\} \\ &= T_1 + T_2, \end{aligned} \tag{4.46}$$

where

$$T_1 := \frac{1}{n} \left\{ \sum_{i=1}^{n} \epsilon_i \cdot Z_i (\hat{\tau}_i^{(1)} - \tau_i^*) + \lambda_\tau \|\nabla_G \tau^*\|_1 - \lambda_\tau \|\nabla_G \hat{\tau}^{(1)}\|_1 \right\},$$

and

$$T_2 := \frac{1}{n} \sum_{i=1}^{n} Z_i (X_i^\top \beta^* - X_i^\top \hat{\beta}^{(0)})(\hat{\tau}_i^{(1)} - \tau_i^*).$$

In order to bound T_1, we write $I_n = \nabla_G^\dagger \nabla_G + P_R$, where $P_R = I_n - \nabla_G^\dagger \nabla_G$ is the projection onto the span of $\mathbf{1}_n$, the all-ones vector of $\mathbb{R}^n$. Then by Hölder's inequality,

$$\begin{aligned} T_1 &= \frac{1}{n}\left\{[\epsilon \circ Z]^\top P_R(\hat{\tau}^{(1)} - \tau^*) + [\epsilon \circ Z]^\top \nabla_G^\dagger \nabla_G(\hat{\tau}^{(1)} - \tau^*) + \lambda_\tau(\|\nabla_G \tau^*\|_1 - \|\nabla_G \hat{\tau}^{(1)}\|_1)\right\} \\ &\leq \frac{1}{n}[\epsilon \circ Z]^\top P_R(\hat{\tau}^{(1)} - \tau^*) + \frac{1}{n}\left\{\|(\epsilon \circ Z)^\top \nabla_G^\dagger\|_\infty \|\nabla_G(\hat{\tau}^{(1)} - \tau^*)\|_1 + \lambda_\tau(\|\nabla_G \tau^*\|_1 - \|\nabla_G \hat{\tau}^{(1)}\|_1)\right\}. \end{aligned}$$

Since $Z_i \perp\!\!\!\perp \epsilon_i$ for $i = 1, ..., n$ by the assumption in the claim, and $Z_i \in \{0, 1\}$, if we denote $\tilde{\epsilon} = \epsilon \circ Z$, we have that $\tilde{\epsilon}$ is also mean zero sub-Gaussian with parameter σ. Hence,

$$\frac{1}{n}\sum_{i=1}^n \tilde{\epsilon}_i = O_{\mathbb{P}}(n^{-1/2}). \tag{4.47}$$

Also, note that P_R is idempotent, i.e., $P_R^2 = P_R$, which leads to

$$\begin{aligned} \frac{1}{n}[\epsilon \circ Z]^\top P_R(\hat{\tau}^{(1)} - \tau^*) &= \frac{1}{n}\tilde{\epsilon}^\top P_R P_R(\hat{\tau}^{(1)} - \tau^*) \\ &= \left(\frac{1}{n}\sum_{i=1}^n \tilde{\epsilon}_i\right) \cdot \left(\frac{1}{n}\sum_{i=1}^n (\hat{\tau}_i^{(1)} - \tau_i^*)\right) \\ &\leq \left(\frac{1}{n}\sum_{i=1}^n \tilde{\epsilon}_i\right) \cdot \|\hat{\tau}^{(1)} - \tau^*\|_\infty \\ &= O_{\mathbb{P}}(n^{-1/2}\text{poly}(\log n)), \end{aligned} \tag{4.48}$$

where the last step follows from (4.47) and Lemma 4.8.4.

By the maximal inequality for Gaussian random variables as in the proof of Theorem 2 in [HR16], it yields the following inequality on an event of probability $1 - 2\delta$,

$$\|(\epsilon \circ Z)^\top \nabla_G^\dagger\|_\infty = \|(\nabla_G^\dagger)^\top \tilde{\epsilon}\|_\infty \leq C_\epsilon \log^{1/2}(n) \sqrt{\log\left\{\frac{c_k n}{\delta}\right\}}$$

for a small $\delta > 0$ and some positive constants C_ϵ and c_k.

With the choice of $K = \log^{1+r} n$ in the claim and η^2 as defined in (4.29), we set the tuning parameter λ_τ as

$$\lambda_\tau = \max\left\{C_\epsilon \log^{1/2}(n) \sqrt{\log\left(\frac{c_k n}{\delta}\right)}, C_d r_n \|\nabla_G \tau^*\|_1^{-1}\right\}, \tag{4.49}$$

where $r_n = \log^{5+2r}(n) + n^{1-1/d} \log^{5/2+2r}(n)$. Then we have that

$$\frac{1}{n}\left\{\|(\epsilon \circ Z)^\top \nabla_G^\dagger\|_\infty \|\nabla_G(\hat{\tau}^{(1)} - \tau^*)\|_1 + \lambda_\tau(\|\nabla_G \tau^*\|_1 - \|\nabla_G \hat{\tau}^{(1)}\|_1)\right\} \leq \frac{\lambda_\tau}{n}\Big[\|\nabla_G(\hat{\tau}^{(1)} - \tau^*)\|_1$$

$$
\begin{aligned}
&\quad + \|\nabla_G \tau^*\|_1 - \|\nabla_G \hat{\tau}^{(1)}\|_1 \Big] \\
&\leq \frac{2\lambda_\tau}{n} \|\nabla_G \tau^*\|_1 \\
&= O_{\mathbb{P}}(n^{-1/d} \mathrm{poly}(\log n)),
\end{aligned} \tag{4.50}
$$

with the first inequality following from the choice of λ_τ in (4.49), the second following from the triangle inequality, and the last step from Assumption 17. From (4.48) and (4.50), since the ambient dimension d is assumed to be fixed with $d \geq 2$, we obtain

$$
T_1 = O_{\mathbb{P}}(n^{-1/d} \mathrm{poly}(\log n)). \tag{4.51}
$$

Next we bound T_2. Following the same reasoning as in the proof of Lemma 4.8.6, we have that

$$
\begin{aligned}
T_2 &= \frac{1}{n} \sum_{i=1}^{n} Z_i (X_i^\top \beta^* - X_i^\top \hat{\beta}^{(0)})(\hat{\tau}_i^{(1)} - \tau_i^*) \\
&\leq \frac{1}{n} \max_{i=1,\ldots,n} |X_i^\top \beta^* - X_i^\top \hat{\beta}^{(0)}| \cdot \|(\hat{\tau}^{(1)} - \tau^*) \circ Z\|_1 \\
&\leq \frac{1}{n} \max_{i=1,\ldots,n} \|X_i^\top\|_2 \cdot \|\hat{\beta}^{(0)} - \beta^*\|_2 \cdot \|(\hat{\tau}^{(1)} - \tau^*) \circ Z\|_1 \\
&\leq C_X \sqrt{d} \|\hat{\beta}^{(0)} - \beta^*\|_2 \cdot \frac{1}{\sqrt{n}} \|(\hat{\tau}^{(1)} - \tau^*) \circ Z\|_2 \\
&\leq C_X^2 d \|\hat{\beta}^{(0)} - \beta^*\|_2^2 + \frac{1}{4n} \sum_{i=1}^{n} Z_i^2 (\hat{\tau}_i^{(1)} - \tau_i^*)^2,
\end{aligned} \tag{4.52}
$$

where the first inequality follows from Hölder's inequality, the second from the Cauchy–Schwarz inequality, the third from the basic norm inequality, and the final inequality uses the inequality $(a+b)^2 \leq 2(a^2+b^2)$.

Therefore, from (4.46), (4.51), and (4.52), we have that

$$
\frac{1}{n} \|(\hat{\tau}^{(1)} - \tau^*) \circ Z\|_2^2 = O_{\mathbb{P}}\left(n^{-1/d} \mathrm{poly}(\log n)\right). \tag{4.53}
$$

Note that

$$
\frac{1}{n} \|\hat{\tau}^{(1)} - \tau^*\|_2^2 = \frac{1}{n} \sum_{Z_i = 1} Z_i^2 (\hat{\tau}_i^{(1)} - \tau_i^*)^2 + \frac{1}{n} \sum_{Z_i = 0} (\hat{\tau}_i^{(1)} - \tau_i^*)^2. \tag{4.54}
$$

From Lemma 4.8.3 and by our choice of $K = \log^{1+r} n$ in the claim, for each data i, with high probability, we can find at least one corresponding data j_i such that $(i, j_i) \in E_K$ and $Z_{j_i} = 1$. Then for such data pair (i, j_i), we have that

$$\begin{aligned}(\hat{\tau}_i^{(1)} - \tau_i^*)^2 &\le 2(\hat{\tau}_{j_i}^{(1)} - \tau_i^*)^2 + 2(\hat{\tau}_i^{(1)} - \hat{\tau}_{j_i}^{(1)})^2 \\ &\le 4(\hat{\tau}_{j_i}^{(1)} - \tau_{j_i}^*)^2 + 4(\tau_i^* - \tau_{j_i}^*)^2 + 2(\hat{\tau}_i^{(1)} - \hat{\tau}_{j_i}^{(1)})^2,\end{aligned}$$

and thus

$$\begin{aligned}\frac{1}{n}\sum_{Z_i=0}(\hat{\tau}_i^{(1)} - \tau_i^*)^2 \le \frac{4}{n} &\sum_{Z_i=0, Z_{j_i}=1, (i,j_i)\in E_K} Z_{j_i}^2(\hat{\tau}_{j_i}^{(1)} - \tau_{j_i}^*)^2 \\ &+ \frac{4}{n}\sum_{Z_i=0, Z_{j_i}=1, (i,j_i)\in E_K} (\tau_i^* - \tau_{j_i}^*)^2 \\ &+ \frac{2}{n}\sum_{Z_i=0, Z_{j_i}=1, (i,j_i)\in E_K} (\hat{\tau}_i^{(1)} - \hat{\tau}_{j_i}^{(1)})^2.\end{aligned} \tag{4.55}$$

Since

$$\sum_{(i,j_i)\in E_K}(\tau_i^* - \tau_{j_i}^*)^2 \le \max_{(i,j_i)\in E_K}|\tau_i^* - \tau_{j_i}^*| \cdot \sum_{(i,j_i)\in E_K}|\tau_i^* - \tau_{j_i}^*| = \max_{(i,j_i)\in E_K}|\tau_i^* - \tau_{j_i}^*| \cdot \|\nabla_G \tau^*\|_1,$$

then from Assumption 17 and the assumption $\|\tau^*\|_\infty = O_{\mathbb{P}}(1)$, we have that

$$\sum_{(i,j_i)\in E_K}(\tau_i^* - \tau_{j_i}^*)^2 = O_{\mathbb{P}}\left(n^{1-1/d}\mathrm{poly}(\log n)\right). \tag{4.56}$$

Similarly, we obtain

$$\sum_{(i,j_i)\in E_K}(\hat{\tau}_i^{(1)} - \hat{\tau}_{j_i}^{(1)})^2 \le \max_{(i,j_i)\in E_K}|\hat{\tau}_i^{(1)} - \hat{\tau}_{j_i}^{(1)}| \cdot \sum_{(i,j_i)\in E_K}|\hat{\tau}_i^{(1)} - \hat{\tau}_{j_i}^{(1)}| = \max_{(i,j_i)\in E_K}|\hat{\tau}_i^{(1)} - \hat{\tau}_{j_i}^{(1)}| \cdot \|\nabla_G \hat{\tau}^{(1)}\|_1,$$

and by the results of Lemmas 4.8.4 and 4.8.8, it yields

$$\sum_{(i,j_i)\in E_K}(\hat{\tau}_i^{(1)} - \hat{\tau}_{j_i}^{(1)})^2 = O_{\mathbb{P}}\left(n^{1-1/d}\mathrm{poly}(\log n)\right). \tag{4.57}$$

The displays (4.55), (4.56) and (4.57) lead to

$$\frac{1}{n}\sum_{Z_i=0}(\hat{\tau}_i^{(1)} - \tau_i^*)^2 = O_{\mathbb{P}}\left(n^{-1/d}\mathrm{poly}(\log n)\right). \tag{4.58}$$

Combining (4.53), (4.54) and (4.58), we arrive at the claim in the theorem. □

CHAPTER 5

Conclusion

5.1 Research Summary

This dissertation has made contributions to the statistical research area of locally adaptive estimation, introducing novel methodologies using the ideas of fused lasso and nearest neighbor algorithms. The work has focused extensively on two key areas of application: quantile regression and causal inference, offering innovative approaches to address complex data challenges in these fields. Below, we succinctly summarize the key contributions of three pivotal chapters:

Non-parametric Quantile Regression via the K-NN Fused Lasso

In this chapter, we introduce a novel adaptive quantile estimator within a non-parametric framework, which incorporates the fused lasso penalty over a K-nearest neighbors graph. This estimator is designed to adapt to the underlying structure of the data. Under mild assumptions on the data generation process for d-dimensional data, we prove that the proposed estimator achieves the optimal convergence rate of $n^{-1/d}$, up to a logarithmic factor. We further present efficient algorithms for computing the estimator and explore methodologies for model selection. Through a series of numerical experiments on both simulated and real datasets, we demonstrate the clear advantages of our estimator over existing state-of-the-art methods.

2D Score Based Estimation of Heterogeneous Treatment Effects

In this chapter, we propose a score-based framework to estimate the conditional average treatment effect (CATE) function. The framework integrates two components: (i) leverage the joint use of propensity and prognostic scores in a matching algorithm to obtain a proxy of the heterogeneous treatment effects for each observation, (ii) utilize non-parametric regression trees to construct an estimator for the CATE function conditioning on the two scores. The method naturally stratifies treatment effects into subgroups over a 2d grid whose axis is propensity and prognostic scores. We conduct benchmark experiments on multiple simulated data and demonstrate clear advantages of the proposed estimator over state-of-the-art methods. We also evaluate empirical performance in real-life settings, using two observational data from a clinical trial and a complex social survey, and interpret policy implications following the numerical results.

Causal Effect Estimation via Fused Lasso over Graph

In this chapter, we present new methods for estimating heterogeneous treatment effects within connected graphs. Our approaches leverage fused lasso over graph, a framework that adapts locally to patterns over graph structures. These methods are not only straightforward but also computationally efficient, and the corresponding estimators achieve rapid convergence rates under mild data assumptions. Through a number of simulation studies, we demonstrate the superior numerical performance of our proposed methods compared to conventional causal effect estimation methods. We then apply our methods to analyze the impact of enforcing background checks on gig workers for transportation network companies (TNCs) on property crime rates. Our findings provide valuable policy insights supported by empirical evidence.

5.2 Discussion on Comparison with Parametric Models

This dissertation primarily focuses on developing new statistical methods that can adapt to underlying local patterns in data. While parametric methods, such as linear models [HTF01, MPV12, IR15], often perform well in terms of estimation accuracy across a wide range of scenarios, they are less effective in recovering intricate local data patterns that are not fully explained by a set of covariates.

In contrast, the methods proposed in this dissertation offer more flexible, data-driven approaches for capturing intrinsic local data patterns without sacrificing estimation accuracy. For instance, in Chapter 2 and Chapter 4, we introduce novel models that leverage graph structures to address challenges in quantile regression and causal inference, respectively. Graph structures serve as a robust alternative to covariate sets, in capturing complex, localized patterns in data. In Chapter 3, we propose a score-based framework that generates a low-dimensional summary of treatment effect heterogeneity. This framework offers advantages over parametric methods in providing a simplest representation of local patterns in treatment effects, which would otherwise be challenging to characterize using a multivariate feature set.

5.3 Future Work

A promising extension of the work presented in this dissertation is the analysis of heterogeneous quantile treatment effects (HQTE). The heterogeneous quantile treatment effect (HQTE) function is defined as the difference between the quantiles of the conditional distributions of treatment and control group. It provides information on the treatment effect at every quantile level. Unlike heterogeneous treatment effects estimated on the mean, HQTE provides a more nuanced view by examining the treatment effects at different points in the distribution. Such an analysis has wide applications in fields including biomedical science [BLW11] and political science [XGO22].

Existing studies, including [Fir07, HKZ19, CLM24], have examined quantile treatment effects (QTE) under different model setups. Recent work by [GW23] introduced a high-dimensional estimator for HQTE using an l_1-penalized regression adjustment combined with a quantile-specific bias correction scheme based on rank scores. However, there is a notable gap in the literature on locally adaptive HQTE estimation methods that can be flexibly adjusted to the underlying structure of the data.

Building on the work presented in this dissertation, future research will focus on developing data-adaptive estimators for HQTE. Specifically, we aim to extend the models of the quantile K-NN fused lasso estimator proposed in Chapter 2 and the causal graph fused lasso estimator in Chapter 4 into HQTE estimation. Proposing valid HQTE estimators that effectively leverage local adaptivity and analyzing the theoretical properties of the estimators, including their consistency and robustness, are critical directions. In addition, the practical efficacy of these methods should be evaluated through extensive simulation studies and applications to real-world datasets.

Bibliography

[AAH21] Dmitry Arkhangelsky, Susan Athey, David A. Hirshberg, Guido W. Imbens, and Stefan Wager. "Synthetic difference-in-differences." *American Economic Review*, **111**(12):4088–4118, 2021.

[AB22] Rachael C. Aikens and Michael Baiocchi. "Assignment-control plots: a visual companion for causal inference study design." *The American Statistician*, pp. 1–26, 2022.

[ACP18] Joseph Antonelli, Matthew Cefalu, Nathan Palmer, and Denis Agniel. "Doubly robust matching estimators for high dimensional confounding adjustment." *Biometrics*, **74**(4):1171–1179, 2018.

[ACW18] Alberto Abadie, Matthew M. Chingos, and Martin R. West. "Endogenous stratification in randomized experiments." *The Review of Economics and Statistics*, **100**(4):567–580, 2018.

[AFS15] Barak Ariel, William A. Farrar, and Alex Sutherland. "The effect of police body-worn cameras on use of force and citizens' complaints against the police: a randomized controlled trial." *Journal of Quantitative Criminology*, **31**(3):509–535, 2015.

[AGB20] Rachael C. Aikens, Dylan Greaves, and Michael Baiocchi. "A pilot design for observational studies: using abundant data thoughtfully." *Statistics in Medicine*, **39**(30):4821–4840, 2020.

[AI06] Alberto Abadie and Guido W. Imbens. "Large sample properties of matching estimators for average treatment effects." *Econometrica*, **74**(1):235–267, 2006.

[AI16] Susan Athey and Guido W. Imbens. "Recursive partitioning for heterogeneous causal effects." *Proceedings of the National Academy of Sciences*, **113**(27):7353–7360, 2016.

[AI17] Susan Athey and Guido W. Imbens. "The state of applied econometrics: causality and policy evaluation." *Journal of Economic Perspectives*, **31**(2):3–32, 2017.

[AJC18] Peter C. Austin, Nathaniel Jembere, and Maria Chiu. "Propensity score matching and complex surveys." *Statistical Methods in Medical Research*, **27**(4):1240–1257, 2018.

[ALL14] Morteza Alamgir, Gábor Lugosi, and Ulrike von Luxburg. "Density-preserving quantization with application to graph downsampling." In *Proceedings of the 27th Conference on Learning Theory*, volume 35, pp. 543–559, 2014.

[And11] Martin A. Andresen. "Estimating the probability of local crime clusters: the impact of immediate spatial neighbors." *Journal of Criminal Justice*, **39**(5):394–404, 2011.

[APE00] 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.

[AS16] Peter C. Austin and Tibor Schuster. "The performance of different propensity score methods for estimating absolute effects of treatments on survival outcomes: a simulation study." *Statistical Methods in Medical Research*, **25**(5):2214–2237, 2016.

[ATW19] Susan Athey, Julie Tibshirani, and Stefan Wager. "Generalized random forests." *The Annals of Statistics*, **47**(2):1148–1178, 2019.

[AW19] Susan Athey and Stefan Wager. "Estimating treatment effects with causal forests: an application." *Observational Studies*, **19**(5):37–51, 2019.

[BB08] Anthony A. Braga and Brenda J. Bond. "Policing crime and disorder hot spots: a randomized controlled trial." *Criminology*, **46**(3):577–607, 2008.

[BBP04] Alain Barrat, Marc Barthélemy, Romualdo Pastor-Satorras, and Alessandro Vespignani. "The architecture of complex weighted networks." *Proceedings of the National Academy of Sciences*, **101**(11):3747–3752, 2004.

[BC11] Alexandre Belloni and Victor Chernozhukov. "l_1-penalized quantile regression in high-dimensional sparse models." *The Annals of Statistics*, **39**(1):82–130, 2011.

[BCC19] Alexandre Belloni, Victor Chernozhukov, Denis Chetverikov, and Iván Fernández-Val. "Conditional quantile processes based on series or many regressors." *Journal of Econometrics*, **213**(1):4–29, 2019.

[BCQ97] M. R. Brito, E. L. Chávez, A. J. Quiroz, and J. E. Yukich. "Connectivity of the mutual k-nearest-neighbor graph in clustering and outlier detection." *Statistics & Probability Letters*, **35**(1):33–42, 1997.

[BD17] Dimitris Bertsimas and Jack William Dunn. "Optimal classification trees." *Machine Learning*, **106**(7):1039–1082, 2017.

[Ben80] Jon Louis Bentley. "Multidimensional divide-and-conquer." *Communications of the ACM*, **23**(4):214–229, 1980.

[Ber87] Richard A. Berk. "Causal inference as a prediction problem." *Crime and Justice*, **9**:183–200, 1987.

[BFS84] Leo Breiman, Jerome H. Friedman, Charles J. Stone, and R. A. Olshen. *Classification and regression trees*. Taylor & Francis, 1984.

[BGC20] Halley L. Brantley, Joseph Guinness, and Eric C. Chi. "Baseline drift estimation for air quality data using quantile trend filtering." *The Annals of Applied Statistics*, **14**(2):585–604, 2020.

[Bla13] Matthew Blackwell. "A framework for dynamic causal inference in political science." *American Journal of Political Science*, **57**(2):504–520, 2013.

[BLW11] Katherine Bowers, Gongshu Liu, Ping Wang, Tao Ye, Zhen Tian, Enqing Liu, Zhijie Yu, Xilin Yang, Mark Klebanoff, Edwina Yeung, Gang Hu, and Cuilin Zhang. "Birth weight, postnatal weight change, and risk for high blood pressure among Chinese children." *Pediatrics*, **127**(5):1272–1279, 2011.

[BLZ16] Adam Bloniarz, Hanzhong Liu, Cun-Hui Zhang, Jasjeet S. Sekhon, and Bin Yu. "Lasso adjustments of treatment effect estimates in randomized experiments." *Proceedings of the National Academy of Sciences*, **113**(27):7383–7390, 2016.

[BML22] David Buil-Gil, Angelo Moretti, and Samuel H. Langton. "The accuracy of crime statistics: assessing the impact of police data bias on geographic crime analysis." *Journal of Experimental Criminology*, **18**(3):515–541, 2022.

[BPC11] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. "Distributed optimization and statistical learning via the alternating direction method of multipliers." *Foundations and Trends in Machine Learning*, **3**(1):1–122, 2011.

[Bre01] Leo Breiman. "Random forests." *Machine Learning*, **45**:5–32, 2001.

[BS18] Álvaro Barbero and Suvrit Sra. "Modular proximal optimization for multidimensional total-variation regularization." *Journal of Machine Learning Research*, **19**:1–82, 2018.

[BTP19] Anthony A. Braga, Brandon S. Turchan, Andrew V. Papachristos, and David M. Hureau. "Hot spots policing and crime reduction: an update of an ongoing systematic review and meta-analysis." *Journal of Experimental Criminology*, **15**(3):289–311, 2019.

[CCD18] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. "Double/debiased machine learning for treatment and structural parameters." *The Econometrics Journal*, **21**(1):1–68, 2018.

[CD09] Antonin Chambolle and Jérome Darbon. "On total variation minimization and surface evolution using parametric maximum flows." *International Journal of Computer Vision*, **84**(3):288–307, 2009.

[Cen19] Pew Research Center. "Methodology and topline for ride-hailing services." *Pew Research Center*, pp. 1–4, 2019.

[CFM15] Anthony C. Constantinou, Martin Freestone, William Marsh, and Jeremy Coid. "Causal inference for violence risk management and decision support in forensic psychiatry." *Decision Support Systems*, **80**(4):42–55, 2015.

[CFS09] Jie Chen, Haw ren Fang, and Yousef Saad. "Fast approximate kNN graph construction for high dimensional data via recursive Lanczos bisection." *Journal of Machine Learning Research*, **10**:1989–2012, 2009.

[CGM10] Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. "BART: Bayesian additive regression trees." *The Annals of Applied Statistics*, **4**(1):266–298, 2010.

[CH67] Thomas M. Cover and Peter E. Hart. "Nearest neighbor pattern classification." *IEEE Transactions on Information Theory*, **13**(1):21–27, 1967.

[CHI08] Richard K. Crump, V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik. "Nonparametric tests for treatment effect heterogeneity." *Review of Economics and Statistics*, **90**(3):389–405, 2008.

[CHI09] Richard K. Crump, V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik. "Dealing with limited overlap in estimation of average treatment effects." *Biometrika*, **96**(1):187–199, 2009.

[CKS23] Yifan Cui, Michael R. Kosorok, Erik Sverdrup, Stefan Wager, and Ruoqing Zhu. "Estimating heterogeneous treatment effects with right-censored data via causal

survival forests." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **85**(2):179–211, 2023.

[CLM24] Xiaohong Chen, Ying Liu, Shujie Ma, and Zheng Zhang. "Causal inference of general treatment effects using neural networks with a diverging number of confounders." *Journal of Econometrics*, **238**(1):808–843, 2024.

[Coc68] William G. Cochran. "The effectiveness of adjustment by subclassification in removing bias in observational studies." *Biometrics*, **24**(2):295–313, 1968.

[Cox83] Dennis D. Cox. "Asymptotics for M-type smoothing splines." *The Annals of Statistics*, **11**(2):530–551, 1983.

[Cox84] David R. Cox. "Interaction (with discussion)." *International Statistical Review*, **52**(1):1–24, 1984.

[CP11] Antonin Chambolle and Thomas Pock. "A first-order primal-dual algorithm for convex problems with applications to imaging." *Journal of Mathematical Imaging and Vision*, **40**:120–145, 2011.

[CS21] Brantly Callaway and Pedro H. C. Sant'Anna. "Difference-in-differences with multiple time periods." *Journal of Econometrics*, **225**(2):200–230, 2021.

[CSD96] Alfred F. Connors Jr., Theodore Speroff, Neal V. Dawson, Charles Thomas, Frank E. Harrell Jr, Douglas Wagner, Norman Desbiens, Lee Goldman, Albert W. Wu, Robert M. Califf, William J. Fulkerson Jr., Humberto Vidaillet, Steven Broste, Paul Bellamy, Joanne Lynn, and William A. Knaus. "The effectiveness of right heart catheterization in the initial care of critically ill patients." *Journal of the American Medical Association*, **276**(11):889–897, 1996.

[CX08] Zongwu Cai and Xiaoping Xu. "Nonparametric quantile estimations for dynamic smooth coefficient models." *Journal of the American Statistical Association*, **103**(484):1595–1608, 2008.

[Das91] Belur V. Dasarathy. *Nearest Neighbor (NN) norms: NN pattern classification techniques*. IEEE Computer Society Press, 1991.

[Daw00] Alexander Philip Dawid. "Causal inference without counterfactuals." *Journal of the American Statistical Association*, **95**(450):407–424, 2000.

[DDF21] Alexander D'Amour, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet S. Sekhon. "Overlap in observational studies with high-dimensional covariates." *Journal of Econometrics*, **221**:644–654, 2021.

[DFM19] Peng Ding, Avi Feller, and Luke Miratrix. "Decomposing treatment effect variation." *Journal of the American Statistical Association*, **114**(525):304–317, 2019.

[DHK16] Issa J. Dahabreh, Rodney Hayward, and David M. Kent. "Using group data to treat individuals: understanding heterogeneous treatment effects in the age of precision medicine and patient-centred evidence." *International Journal of Epidemiology*, **45**(6):2184–2193, 2016.

[Die93] Dietrich Wettschereck andThomas Dietterich. "Locally adaptive nearest neighbor algorithms." In *Advances in Neural Information Processing Systems 9*, pp. 184–191. Morgan Kaufmann Publishers, Inc., 1993.

[DK01] P. Laurie Davies and Arne Kovac. "Local extremes, runs, strings and multiresolution." *The Annals of Statistics*, **32**(1):1–65, 2001.

[DL18] Peng Ding and Fan Li. "Causal inference: a missing data perspective." *Statistical Science*, **33**(2):214–237, 2018.

[Don97] David L. Donoho. "CART and best-ortho-basis: a connection." *The Annals of Statistics*, **25**(5):1870–1911, 1997.

[DSS14] Eva H. Dugoff, Megan Schuler, and Elizabeth A. Stuart. "Generalizing observational study results: applying propensity score methods to complex surveys." *Health Service Research*, **49**(1):284–303, 2014.

[DW02] Rajeev H. Dehejia and Sadek Wahba. "Propensity score matching methods for nonexperimental causal studies." *Review of Economics and Statistics*, **84**(1):151–161, 2002.

[Efr79] Bradley Efron. "Bootstrap methods: another look at the Jackknife." *The Annals of Statistics*, **7**(1):1–26, 1979.

[Eub88] Randall L. Eubank. *Spline smoothing and nonparametric regression*, volume 90. M. Dekker New York, 1988.

[FFB14] Jianqing Fan, Yingying Fan, and Emre Barut. "Adaptive robust variable selection." *The Annals of Statistics*, **42**(1):324–351, 2014.

[FH51] E. Fix and J. L. Hodges. "Discriminatory analysis, nonparametric discrimination: consistency properties." Technical Report 4, USAF School of Aviation Medicine, Randolph Field, 1951.

[FHT10] Jerome H. Friedman, Trevor Hastie, and Robert Tibshirani. "Regularization paths for generalized linear models via coordinate descent." *Journal of Statistical Software*, **33**(1):1–22, 2010.

[Fir07] Sergio Firpo. "Efficient semiparametric estimation of quantile treatment effects." *Econometrica*, **75**(1):259–276, 2007.

[FMW22] Laura Forastiere, Fabrizia Mealli, Albert Wu, and Edoardo M. Airoldi. "Estimating causal effects under network interference with Bayesian generalized propensity scores." *Journal of Machine Learning Research*, **23**(289):1–61, 2022.

[FWW11] Michele Jonsson Funk, Daniel Westreich, Chris Wiesen, Til Stürmer, M. Alan Brookhart, and Marie Davidian. "Doubly robust estimation of causal effects." *American Journal of Epidemiology*, **173**(7):761–767, 2011.

[GC07] Amy Berrington de González and David R. Cox. "Interpretation of interaction: a review." *The Annals of Applied Statistics*, **1**(2):371–385, 2007.

[GD23] Mengsi Gao and Peng Ding. "Causal inference in network experiments: regression-based analysis and design-based properties." *arXiv eprint*, p. 2309.07476, 2023.

[GK11] Brian J. Gaines and James H. Kuklinski. "Estimation of heterogeneous treatment effects related to self-selection." *American Journal of Political Science*, **55**(3):724–736, 2011.

[GK12] Donald P. Green and Holger L. Kern. "Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees." *The Public Opinion Quarterly*, **76**(3):491–511, 2012.

[GKK06] László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

[GR93] Xing Sam Gu and Paul R. Rosenbaum. "Comparison of multivariate matching methods: structures, distances, and algorithms." *Journal of Computational and Graphical Statistics*, **2**(4):405–420, 1993.

[GW23] Alexander Giessing and Jingshen Wang. "Debiased inference on heterogeneous quantile treatment effects with regression rank scores." *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, **85**(5):1561–1588, 2023.

[GWB03] Gongde Guo, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer. "KNN model-based approach in classification." In *On The Move to Meaningful Internet Systems*, pp. 986–996. Springer, 2003.

[Hah98] Jinyong Hahn. "On the role of the propensity score in efficient semiparametric estimation of average treatment effects." *Econometrica*, **66**(2):315–331, 1998.

[Han04] Ben B. Hansen. "Full matching in an observational study of coaching for the SAT." *Journal of the American Statistical Association*, **99**(467):609–618, 2004.

[Han06] Ben B. Hansen. "Bias reduction in observational studies via prognosis scores." *Statistics Department, University of Michigan; Ann Arbor, Michigan: Technical Report*, **441**, 2006.

[Han08] Ben B. Hansen. "The prognostic analogue of the propensity score." *Biometrika*, **95**(2):481–488, 2008.

[HHC88] Henrick J. Harwood, Robert L. Hubbard, John J. Collins, and John V. Rachal. "The costs of crime and the benefits of drug abuse treatment: a cost-benefit analysis using TOPS data." *NIDA Research Monograph*, **86**:209–235, 1988.

[HI01] Keisuke Hirano and Guido W. Imbens. "Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization." *Health Services & Outcomes Research Methodology*, **2**(3):259–278, 2001.

[Hil11] Jennifer L. Hill. "Bayesian nonparametric modeling for causal inference." *Journal of Computational and Graphical Statistics*, **20**(1):217–240, 2011.

[HKK01] Euihong Han, George Karypis, and Vipin Kumar. "Text categorization using weight adjusted k-Nearest neighbor classification." In *Advances in Knowledge Discovery and Data Mining 2035*, pp. 53–65. Springer, 2001.

[HKZ19] Peisong Han, Linglong Kong, Jiwei Zhao, and Xingcai Zhou. "A general framework for quantile estimation with incomplete data." *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, **81**(2):305–333, 2019.

[HL00] David R. Hunter and Kenneth Lange. "Quantile regression via an MM algorithm." *Journal of Computational and Graphical Statistics*, **9**(1):60–77, 2000.

[HL17] Dorit S. Hochbaum and Cheng Lu. "A faster algorithm for solving a generalization of isotonic median regression and a class of fused lasso problems." *SIAM Journal on Optimization*, **27**(4):2563–2596, 2017.

[HLP14] James J. Heckman, Hedibert F. Lopes, and Remi Piatek. "Treatment effects: a Bayesian perspective." *Econometric Reviews*, **33**(1-4):36–67, 2014.

[HMC20] P. Richard Hahn, Jared S. Murray, and Carlos M. Carvalho. "Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects." *Bayesian Analysis*, **15**(3):965–1056, 2020.

[HNP98] Xuming He, Pin Ng, and Stephen Portnoy. "Bivariate quantile smoothing splines." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **60**(3):537–550, 1998.

[Hoe10] Holger Hoefling. "A path algorithm for the fused lasso signal approximator." *Journal of Computational and Graphical Statistics*, **19**(4):984–1006, 2010.

[Hol86] Paul W. Holland. "Statistics and causal inference." *Journal of the American Statistical Association*, **81**(396):945–960, 1986.

[HR16] Jan-Christian Hütter and Philippe Rigollet. "Optimal rates for total variation denoising." In *Proceedings of the 29th Annual Conference on Learning Theory*, volume 49, pp. 1115–1146, 2016.

[HR20] Miguel A. Hernán and James M. Robins. *Causal inference: what if*. Chapman & Hall/CRC, 2020.

[HS94] Xuming He and Peide Shi. "Convergence rate of b-spline estimators of nonparametric conditional quantile functions." *Journal of Nonparametric Statistics*, **3**(3–4):299–308, 1994.

[HS13] Jennifer L. 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*, **7**(3):1386–1420, 2013.

[HTF01] Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. New York: Springer, 2001.

[ID04] Kosuke Imai and David A. van Dyk. "Causal inference with general treatment regimes." *Journal of the American Statistical Association*, **99**(467):854–866, 2004.

[Imb04] Guido W. Imbens. "Nonparametric estimation of average treatment effects under exogeneity: a review." *Review of Economics and Statistics*, **86**:4–29, 2004.

[IR13] Kosuke Imai and Marc Ratkovic. "Estimating treatment effect heterogeneity in randomized program evaluation." *The Annals of Applied Statistics*, **7**(1):443–470, 2013.

[IR15] Guido W. Imbens and Donald B. Rubin. *Causal inference for statistics, social, and biomedical sciences: an introduction*. Cambridge University Press, 2015.

[IS11] Kosuke Imai and Aaron Strauss. "Estimation of heterogeneous treatment effects from randomized experiments, with application to the optimal planning of the get-out-the-vote campaign." *Political Analysis*, **19**:1–19, 2011.

[JDG03] Elizabeth Johnson, Francesca Dominici, Michael Griswold, and Scott L. Zeger. "Disease cases and their medical costs attributable to smoking: an analysis of the national medical expenditure survey." *Journal of Econometrics*, **112**(1):135–151, 2003.

[Joh13] Nicholas Johnson. "A dynamic programming algorithm for the fused lasso and l_0-segmentation." *Journal of Computational and Graphical Statistics*, **22**(2):246–260, 2013.

[Kat11] Kengo Kato. "Group lasso for high dimensional sparse quantile regression models.", 2011.

[KB78] Roger Koenker and Gilbert Bassett Jr. "Regression quantiles." *Econometrica*, **46**(1):33–50, 1978.

[Kee15] Luke Keele. "The statistics of causal inference: a view from political methodology." *Political Analysis*, **23**(3):313–335, 2015.

[KH07] David M. Kent and Rodney A. Hayward. "Limitations of applying summary results of clinical trials to individual patients: the need for risk stratification." *Journal of the American Medical Association*, **298**(10):1209–1212, 2007.

[KKB09] Seung-Jean Kim, Kwangmoo Koh, Stephen Boyd, and Dimitry Gorinevsky. "l_1 trend filtering." *SIAM Review*, **51**(2):339–360, 2009.

[KNP94] Roger Koenker, Pin Ng, and Stephen Portnoy. "Quantile smoothing splines." *Biometrika*, **81**(4):673–680, 1994.

[Koe05] Roger Koenker. *Quantile regression*. Cambridge University Press, 2005.

[KP18] Noureddine El Karoui and Elizabeth Purdom. "Can we trust the bootstrap in high-dimensions? The case of linear models." *Journal of Machine Learning Research*, **19**(5):1–66, 2018.

[Kpo11] Samory Kpotufe. "k-NN regression adapts to local intrinsic dimension." In *Advances in Neural Information Processing Systems 24*, pp. 729–737. Curran Associates, Inc., 2011.

[KSB19] 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.

[KVW18] Brandon Koch, David M. Vock, and Julian Wolfson. "Covariate selection with group lasso and doubly robust estimation of causal effects." *Biometrics*, **74**(1):8–17, 2018.

[LD04] Jared K. Lunceford and Marie Davidian. "Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study." *Statistics in Medicine*, **23**(19):2937–2960, 2004.

[LDH23] Zhexiao Lin, Peng Ding, and Fang Han. "Estimation based on nearest neighbor matching: from density ratio to average treatment effect." *Econometrica*, **91**(6):2187–2217, 2023.

[LL16] Alexander R. Luedtke and Mark J. van der Laan. "Super-learning of an optimal dynamic treatment rule." *International Journal of Biostatistics*, **12**(1):305–332, 2016.

[LPH07] Mark J. van der Laan, Eric C. Polley, and Alan E. Hubbard. "Super learner." *Statistical Applications in Genetics and Molecular Biology*, **6**(1):1–23, 2007.

[LPZ24] Hangjian Li, Oscar Hernan Madrid Padilla, and Qing Zhou. "Learning Gaussian DAGs from network data." *Journal of Machine Learning Research*, **25**(377):1–52, 2024.

[LRH14] Ulrike von Luxburg, Agnes Radl, and Matthias Hein. "Hitting and commute times in large random neighborhood graphs." *Journal of Machine Learning Research*, **15**(52):1751–1798, 2014.

[LS14] Finbarr P. Leacy and Elizabeth A. Stuart. "On the joint use of propensity and prognostic scores in estimation of the average treatment effect on the treated: a simulation study." *Statistics in Medicine*, **33**(20):3488–3508, 2014.

[LT13] Michel Ledoux and Michel Talagrand. *Probability in Banach spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

[Lux07] Ulrike von Luxburg. "A tutorial on spectral clustering." *Statistics and Computing*, **17**(4):395–416, 2007.

[LV21] Wen Wei Loh and Stijn Vansteelandt. "Confounder selection strategies targeting stable treatment effect estimators." *Statistics in Medicine*, **40**(3):607–630, 2021.

[LW22] Shuangning Li and Stefan Wager. "Random graph asymptotics for treatment effect estimation under network interference." *The Annals of Statistics*, **50**(4):2334–2358, 2022.

[LZ08] Youjuan Li and Ji Zhu. "l_1-norm quantile regression." *Journal of Computational and Graphical Statistics*, **17**(1):163–185, 2008.

[Mei06] Nicolai Meinshausen. "Quantile random forests." *Journal of Machine Learning Research*, **7**:983–999, 2006.

[MG97] Enno Mammen and Sara van de Geer. "Locally adaptive regression splines." *The Annals of Statistics*, **25**(1):387–413, 1997.

[MPV12] Douglas C. Montgomery, Elizabeth A. Peck, and G. Geoffrey Vining. *Introduction to Linear Regression Analysis*. John Wiley & Sons, 2012.

[MR01a] Kewei Ming and Paul R. Rosenbaum. "A note on optimal matching with variable controls using the assignment algorithm." *Journal of Computational and Graphical Statistics*, **10**(3):455–463, 2001.

[MR01b] Kewei Ming and Paul R. Rosenbaum. "Substantial gains in bias reduction from matching with a variable number of controls." *Biometrics*, **56**(1):118–124, 2001.

[MR09] Eric S. McCord and Jerry H. Ratcliffe. "Intensity value analysis and the criminogenic effects of land use features on local crime patterns." *Crime Patterns and Analysis*, **2**(1):17–30, 2009.

[MW14] Stephen L. Morgan and Christopher Winship. *Counterfactuals and causal inference: methods and principles for social research*. Cambridge University Press, 2014.

[MWE12] Ojmarrh Mitchell, David B. Wilson, Amy Eggers, and Doris L. MacKenzie. "Assessing the effectiveness of drug courts on recidivism: a meta-analytic review of traditional and non-traditional drug courts." *Journal of Criminal Justice*, **40**(1):60–71, 2012.

[MZL22] Haixu Ma, Donglin Zeng, and Yufeng Liu. "Learning individualized treatment rules with many treatments: a supervised clustering approach using adaptive fusion." In *Advances in Neural Information Processing Systems*, volume 35, pp. 15956–15969, 2022.

[MZL23] Haixu Ma, Donglin Zeng, and Yufeng Liu. "Learning optimal group-structured individualized treatment rules with many treatments." *Journal of Machine Learning Research*, **24**(102):1–48, 2023.

[NG04] Mark E. J. Newman and Michelle Girvan. "Finding and evaluating community structure in networks." *Physics Review E*, **69**:026113, 2004.

[NW21] Xinkun Nie and Stefan Wager. "Quasi-oracle estimation of heterogeneous treatment effects." *Biometrika*, **108**(2):299–319, 2021.

[OE16] Ziad Obermeyer and Ezekiel J. Emanuel. "Predicting the future: big data, machine learning, and clinical medicine." *The New England Journal of Medicine*, **375**(13):1216–1219, 2016.

[OG18] Francesco Ortelli and Sara van de Geer. "On the total variation regularized estimator over a class of tree graphs." *Electronic Journal of Statistics*, **12**(2):4517–4570, 2018.

[OR21] Arman Oganisian and Jason A. Roy. "A practical introduction to Bayesian estimation of causal effects: parametric and nonparametric approaches." *Statistics in Medicine*, **40**(2):518–551, 2021.

[OSL20] Elizabeth L. Ogburn, Ilya Shpitser, and Youjin Lee. "Causal inference, social networks and chain graphs." *Journal of the Royal Statistical Society Series A: Statistics in Society*, **183**(4):1659–1676, 2020.

[PB97] Kelley Pace and Ronald Barry. "Sparse spatial autoregressions." *Statistics & Probability Letters*, **33**(3):291–297, 1997.

[PBF22] David Puelz, Guillaume Basse, Avi Feller, and Panos Toulis. "A graph-theoretic approach to randomization tests of causal effects under general interference." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **84**(1):174–204, 2022.

[PC22] Oscar Hernan Madrid Padilla and Sabyasachi Chatterjee. "Risk bounds for quantile trend filtering." *Biometrika*, **109**(3):751–768, 2022.

[PCR21] Oscar Hernan Madrid Padilla, Yanzhen Chen, and Gabriel Ruiz. "A causal fused lasso for interpretable heterogeneous treatment effects estimation." *arXiv preprint*, 2021.

[Pea95] Judea Pearl. "Causal diagrams for empirical research." *Biometrika*, **82**(4):669–710, 1995.

[Pea09] Judea Pearl. *Causality: models, reasoning, and inference*. Cambridge University Press, 2009.

[PGK17] Matthew Pietrosanu, Jueyu Gao, Linglong Kong, Bei Jiang, and Di Niu. "Advanced algorithms for penalized quantile and composite quantile regression." *Computational Statistics*, pp. 1–14, 2017.

[PMW23] Carlos Misael Madrid Padilla, Oscar Hernan Madrid, and Daren Wang. "Temporal-spatial model via trend filtering." *arXiv eprint*, 2023.

[PQJ18] Scott Powers, Junyang Qian, Kenneth Jung, Alejandro Schuler, Nigam H. Shah, Trevor Hastie, and Robert Tibshirani. "Some methods for heterogeneous treat-

ment effect estimation in high dimensions." *Statistics in Medicine*, **37**(11):1767–1787, 2018.

[PSC20] Oscar Hernan Madrid Padilla, James Sharpnack, Yanzhen Chen, and Daniela M. Witten. "Adaptive nonparametric regression with the K-nearest neighbour fused lasso." *Biometrika*, **107**(2):293–310, 2020.

[PSS18] Oscar Hernan Madrid Padilla, James Sharpnack, James Scott, and Robert J. Tibshirani. "The DFS fused lasso: linear-time denoising over general graphs." *Journal of Machine Learning Research*, **18**(176):1–36, 2018.

[PSW16] Ashley Petersen, Noah Simon, and Daniela M. Witten. "Convex regression with interpretable sharp partitions." *Journal of Machine Learning Research*, **17**(94):1–31, 2016.

[PTC22] Oscar Hernan Madrid Padilla, Wesley Tansey, and Yanzhen Chen. "Quantile regression with ReLU networks: estimators and minimax rates." *Journal of Machine Learning Research*, **23**(247):1–42, 2022.

[QM11] Min Qian and Susan A. Murphy. "Performance guarantees for individualized treatment rules." *The Annals of Statistics*, **39**(2):1180–1210, 2011.

[RCL24] Arun Rai, Yanzhen Chen, and Yatang Lin. "Exclusion for public safety or inclusion for gig employment: managing the tension with a trilogy of guardians." *MIS Quarterly*, 2024.

[Rob88] Peter M. Robinson. "Root-n-consistent semiparametric regression." *Econometrica*, **56**(4):931–954, 1988.

[Rob04] James M. Robins. "Optimal structural nested models for optimal sequential decisions." In *Proceedings of the Second Seattle Symposium in Biostatistics*, pp. 189–326. Springer, 2004.

[ROF92] 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.

[Ros89] Paul R. Rosenbaum. "Optimal matching for observational studies." *Journal of the American Statistical Association*, **84**(408):1024–1032, 1989.

[Ros91] Paul R. Rosenbaum. "A characterization of optimal designs for observational studies." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **53**(3):597–610, 1991.

[RPR20] Alexandros Rekkas, Jessica K. Paulus, Gowri Raman, John B. Wong, Ewout W. Steyerberg, Peter R. Rijnbeek, David M. Kent, and David van Klaveren. "Predictive approaches to heterogeneous treatment effects: a scoping review." *BMC Medical Research Methodology*, **20**(1):264, 2020.

[RR83] 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.

[RR85] Paul R. Rosenbaum and Donald B. Rubin. "Constructing a control group using multivariate matched sampling methods that incorporate the propensity score." *The American Statistician*, **39**(1):33–38, 1985.

[RT96] Donald B. Rubin and Neal Thomas. "Matching using estimated propensity scores: relating theory to practice." *Biometrics*, **52**(1):249–264, 1996.

[RT00] Donald B. Rubin and Neal Thomas. "Combining propensity score matching with additional adjustments for prognostic covariates." *Journal of the American Statistical Association*, **95**(450):573–585, 2000.

[Rub73] Donald B. Rubin. "Matching to remove bias in observational studies." *Biometrics*, **29**(1):159–183, 1973.

[Rub74] Donald B. Rubin. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology*, **66**(5):688–701, 1974.

[Rub01] Donald B. Rubin. "Using propensity scores to help design observational studies: application to the tobacco litigation." *Health Services and Outcomes Research Methodology*, **2**:169–188, 2001.

[SA21] Liyang Sun and Sarah Abraham. "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects." *Journal of Econometrics*, **225**(2):175–199, 2021.

[Sch78] Gideon Schwarz. "Estimating the dimension of a model." *The Annals of Statistics*, **6**(2):461–464, 1978.

[SG08] Elizabeth A. Stuart and Kerry M. Green. "Using full matching to estimate causal effects in non-experimental studies: examining the relationship between adolescent marijuana use and adult outcomes." *Developmental Psychology*, **44**(2):395–406, 2008.

[Sha91] Juliet Popper Shaffer. "The Gauss-Markov theorem and random regressors." *The American Statistician*, **45**(4):269–273, 1991.

[SJS17] 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*, pp. 3076–3085, 2017.

[SLO19] 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 32*, pp. 15167–15176. Curran Associates, Inc., 2019.

[SM15] I. Manjula Schou and Ian C. Marschner. "Methods for exploring treatment ef-

fect heterogeneity in subgroup analysis: an application to global clinical trials." *Pharmaceutical Statistics*, **14**(1):44–55, 2015.

[Smi97] Herbert L. Smith. "Matching with multiple controls to estimate treatment effects in observational studies." *Sociological Methodology*, **27**(1):325–353, 1997.

[SMR21] Matthew J. Smith, Camille Maringe, Bernard Rachet, Mohammad A. Mansournia, Paul N. Zivich, Stephen R. Cole, and Miguel Angel Luque-Fernandez. "Introduction to computational causal inference using reproducible Stata, R, and Python code: a tutorial." *Statistics in Medicine*, 2021.

[SP12] Bisakha Sen and Anantachai Panjamapirom. "State background checks for gun purchase and firearm deaths: an exploratory study." *Preventive Medicine*, **55**(4):346–350, 2012.

[Spl23] Jerzy Splawa-Neyman. "On the application of probability theory to agricultural experiments." *The Annals of Agricultural Sciences*, **10**:1–51, 1923.

[Stu10] Elizabeth A. Stuart. "Matching methods for causal inference: a review and a look forward." *Statistical Science: a Review Journal of the Institute of Mathematical Statistics*, **25**(1):1–21, 2010.

[STW09] Xiaogang Su, Chih-Ling Tsai, Hansheng Wang, David M. Nickerson, and Bogong Li. "Subgroup analysis via recursive partitioning." *Journal of Machine Learning Research*, **10**:141–158, 2009.

[SWH13] Vladimir Spokoiny, Weining Wang, and Wolfgang Karl Härdle. "Local quantile regression." *Journal of Statistical Planning and Inference*, **143**(7):1109–1129, 2013.

[SWT16] Veeranjaneyulu Sadhanala, Yu-Xiang Wang, and Ryan J. Tibshirani. "Total variation classes beyond 1d: minimax rates, and the limitations of linear smoothers."

In *Advances in Neural Information Processing Systems 30*, pp. 3513–3521. Curran Associates, Inc., 2016.

[Tal05] Michel Talagrand. *The Generic Chaining*. Springer, 2005.

[TC23] New York City Taxi and Limousine Commission. *Chapter 59 for hire service*, 2023.

[TGC16] Matt Taddy, Matt Gardner, Liyun Chen, and David Draper. "A nonparametric Bayesian analysis of heterogeneous treatment effects in digital experimentation." *Journal of Business & Economic Statistics*, **34**(4):661–672, 2016.

[Tib96] Robert Tibshirani. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **58**(1):267–288, 1996.

[Tib14] Ryan J. Tibshirani. "Adaptive piecewise polynomial estimation via trend filtering." *The Annals of Statistics*, **42**(1):285–323, 2014.

[Tse01] Paul Tseng. "Convergence of a block coordinate descent method for nondifferentiable minimization." *Journal of Optimization Theory and Applications*, **109**(3):475–494, 2001.

[TSR05] 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.

[TT11] Ryan J. Tibshirani and Jonathan Taylor. "The solution path of the generalized lasso." *The Annals of Statistics*, **39**(3):1335–1371, 2011.

[TT12] Ryan J. Tibshirani and Jonathan Taylor. "Degrees of freedom in lasso problems." *The Annals of Statistics*, **40**(2):1198–1232, 2012.

[TTS18] Wesley Tansey, Jesse Thomason, and James Scott. "Maximum-variance total variation denoising for interpretable spatial smoothing." In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 2018.

[TTT17] Julien Tanniou, Ingeborg van der Tweel, Steven Teerenstra, and Kit C. B. Roes. "Estimates of subgroup treatment effects in overall nonsignificant trials: to what extent should we believe in them?" *Pharmaceutical Statistics*, **16**(4):280–295, 2017.

[Utr81] Florencio I. Utreras. "On computing robust splines and applications." *SIAM Journal on Scientific and Statistical Computing*, **2**(2):153–163, 1981.

[Van15] Tyler J. VanderWeele. *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press, 2015.

[Var16] Hal R. Varian. "Causal inference in economics and marketing." *Proceedings of the National Academy of Sciences*, **113**(27):7310–7315, 2016.

[Ver18] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

[WA18] 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.

[Wah02] Grace Wahba. "Soft and hard classification by reproducing kernel Hilbert space methods." *Proceedings of the National Academy of Sciences*, **99**(26):16524–16530, 2002.

[Wai18] Martin J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2018.

[WCA23] Richard A. Watson, Hengrui Cai, Xinming An, Samuel Mclean, and Rui Song. "On heterogeneous treatment effects in heterogeneous causal graphs." In *Proceed-*

ings of the 40th International Conference on Machine Learning, pp. 36714–36747, 2023.

[WL08] Tong Tong Wu and Kenneth Lange. "Coordinate descent algorithms for lasso penalized regression." *The Annals of Applied Statistics*, **2**(1):224–244, 2008.

[WLT07] Hansheng Wang, Runze Li, and Chih-Ling Tsai. "Tuning parameter selectors for the smoothly clipped absolute deviation method." *Biometrika*, **94**(3):553–568, 2007.

[WNC06] Rebecca Willett, Robert Nowak, and Rui M. Castro. "Faster rates in regression via active learning." In *Advances in Neural Information Processing Systems 18*, pp. 179–186. MIT Press, 2006.

[WS98] Duncan J. Watts and Steven H. Strogatz. "Collective dynamics of 'small-world' networks." *Nature*, **393**:440–442, 1998.

[WSS16] Yu-Xiang Wang, James Sharpnack, Alex Smola, and Ryan J. Tibshirani. "Trend filtering on graphs." *Journal of Machine Learning Research*, **17**(105):1–41, 2016.

[XGO22] Ruofan Xu, Jiti Gao, Tatsushi Oka, and Yoon-Jae Whang. "Estimation of heterogeneous treatment effects using quantile regression with interactive fixed effects.", 2022.

[XST17] Veeranjaneyulu Sadhanala andYu Xiang Wang, James Sharpnack, and Ryan J Tibshirani. "Higher-order total variation classes on grids: minimax theory and trend filtering methods." In *Advances in Neural Information Processing Systems 31*, pp. 5802–5812. Curran Associates, Inc., 2017.

[YCP23] Siwei Ye, Yanzhen Chen, and Oscar Hernan Madrid Padilla. "2D score based estimation of heterogeneous treatment effects." *Journal of Causal Inference*, **11**(1):1–26, 2023.

[YHL12] Guo-Xun Yuan, Chia-Hua Ho, and Chih-Jen Lin. "An improved GLMNET for l1-regularized logistic regression." *Journal of Machine Learning Research*, **13**:1999–2030, 2012.

[YIC16] Shu Yang, Guido W. Imbens, Zhanglin Cui, Douglas E. Faries, and Zbigniew Kadziola. "Propensity score matching and subclassification in observational studies with multi-level treatments." *Biometrics*, **72**(4):1055–1065, 2016.

[YJ98] Keming Yu and M. C. Jones. "Local linear quantile regression." *Journal of the American Statistical Association*, **93**(441):228–237, 1998.

[YM01] Keming Yu and Rana A. Moyeed. "Bayesian quantile regression." *Statistics & Probability Letters*, **54**(4):437–447, 2001.

[YP21] Siwei Ye and Oscar Hernan Madrid Padilla. "Non-parametric quantile regression via the K-NN fused lasso." *Journal of Machine Learning Research*, **22**(111):1–38, 2021.

[ZDI20] Shandong Zhao, David A. van Dyk, and Kosuke Imai. "Propensity score-based methods for causal inference in observational studies with non-binary treatments." *Statistical Methods in Medical Research*, **29**(3):709–727, 2020.

[Zie12] William P. Ziemer. *Weakly differentiable functions: Sobolev spaces and functions of bounded variation*, volume 120. Springer Science & Business Media, 2012.

[ZK17] Jose R. Zubizarreta and Luke Keele. "Optimal multilevel matching in clustered observational studies: a case study of the effectiveness of private schools under a large-scale voucher system." *Journal of the American Statistical Association*, **112**(518):547–560, 2017.

[ZLL17] Weijia Zhang, Thuc Duy Le, Lin Liu, Zhi-Hua Zhou, and Jiuyong Li. "Mining heterogeneous causal effects for personalized cancer treatment." *Bioinformatics*, **33**(15):2372–2378, 2017.

[Zub12] Jose R. Zubizarreta. "Using mixed integer programming for matching in an observational study of kidney failure after surgery." *Journal of the American Statistical Association*, **107**(500):1360–1371, 2012.

[ZZL15] Ying-Qi Zhao, Donglin Zeng, Eric B. Laber, Rui Song, Ming Yuan, and Michael Rene Kosorok. "Doubly robust learning for estimating individualized treatment with censored data." *Biometrika*, **102**(1):151–168, 2015.