Master's Thesis

Estimating Continuous Treatment Effects via Kernel Methods: How Bandwidth Choice Shapes the Results

Ryoto Kawahara

Student number: 15284050

Date of final version: July 15, 2025
Master's programme: Econometrics
Specialisation: Data Analytics
Supervisor: Dr. A. Juodis
Second reader: Dr. E.S. Zwiers

Faculty of Economics and Business

University of Amsterdam
Economics & Business

Statement of Originality

This document is written by Ryoto Kawahara who declares to take full responsibility for the contents of this document. I declare that the text and the work presented in this document is original and that no sources other than those mentioned in the text and its references have been used in creating it. I have not used generative AI (such as ChatGPT) to generate or rewrite text. UvA Economics and Business is responsible solely for the supervision of completion of the work and submission, not for the contents.

Contents

Chapter 1

Introduction

Causal inference has played a key role in evaluating policy and various types of treatment. For panel data and repeated cross-section data, Differences-in-differences (DiD) has been one of the most popular causal inference methods when some units receive treatment while others remain untreated. Under some assumptions, DiD can identify the causal effect by comparing the change in outcome in the treatment group with the change in outcome in the control group. An important assumption of DiD is that each unit in the treatment group receives a binary treatment at the same time and that the treatment lasts afterwards. Another important assumption for DiD to be valid is parallel trends, which states that the potential outcomes of treated and control units would have followed the same trend in the absence of treatment. This excludes the influence of other factors and ensures comparability between the treatment and control groups. In recent years, a growing body of literature has extended the canonical DiD by relaxing these assumptions (Roth et al., 2023).

Among recent advancements in the literature, one strand of literature extends DiD to settings with a continuous treatment and high-dimensional covariates, using a kernel function and a bandwidth. Based on a method proposed by Haddad et al. (2024), this paper investigates how the bandwidth choice affects estimation performance, using Monte Carlo simulations. It aims to provide insights not only for DiD, but also more broadly for kernel-based estimation of continuous treatment effects.

The method proposed by Haddad et al. (2024) builds on a growing body of literature on DiD with covariates, assuming conditional parallel trends. The parallel trends assumption may fail if outcome-related covariates are imbalanced across groups (Abadie, 2005). Instead, the conditional parallel trends assumption imposes parallel trends only conditional on covariates. Under this assumption, Abadie (2005) proposes a semi-parametric DiD that can address the violation of the parallel trends, caused by compositional differences in covariates between the treatment and control groups. Furthermore, Chang (2020) introduces the double/debiased machine learning (DML) method by Chernozhukov et al. (2018), based on the method by Abadie (2005). While machine learning methods provide flexible nonparametric estimation and can handle high-dimensional covariates, they can result in bias, slower convergence (Chernozhukov

et al., 2018). By applying the DML framework to DiD, it helps prevent such problems while enjoying the benefits of machine learning methods.

Recent advancements in continuous DiD, which extends the canonical DiD to settings with a continuous treatment, provide another foundation for the method by Haddad et al. (2024). In many practical settings, treatment variables are often continuous rather than binary, yet the canonical DiD has been applied to continuous outcomes in empirical studies—for example, by Ananat et al. (2024). However, its theoretical properties under continuous treatments had remained underexplored. Callaway et al. (2025) shows that, when a treatment variable is continuous, the canonical DiD can yield counterintuitive causal parameters with negative weights, even when all treatment effects are non-negative. Callaway et al. (2025); D'Haultfœuille et al. (2023); de Chaisemartin et al. (2025) propose alternative estimation methods that yield interpretable causal parameters.

Building on the literature on continuous DiD and DiD with covariates, Zhang (2025) proposes a kernel-based continuous DiD model using DML, which can identify the average treatment effect on the treated (ATT) for a continuous treatment, even when a large number of covariates may affect treatment assignment. Based on Abadie (2005), Zhang (2025) introduces a kernel function with a bandwidth and assigns non-negative weights to local neighborhoods around a treatment level to estimate the ATT for a continuous treatment. Furthermore, Haddad et al. (2024) extends this framework to settings where every unit in both the treatment and control groups receives a non-zero treatment, and where treatment intensities can vary over time.

Kernel-based methods require a bandwidth to control the amount of smoothing, and the estimates can be highly sensitive to this choice in terms of bias and variance. In the context of kernel density estimation and kernel regression, there exists a large body of research on optimal bandwidth selection. However, these approaches are not directly applicable to kernel-based causal inference with a continuous treatment, where the goal is to estimate causal effects rather than densities or outcomes, and the true treatment effects are not observed in the data. Both Zhang (2025) and Haddad et al. (2024) propose undersmoothing the bandwidth so that the bias vanishes asymptotically. However, how to determine the bandwidth and how it affects the estimates remain underexplored in the literature. As a result, selecting the bandwidth in this context is a nontrivial problem, and there is currently no established guidance on how it should be done.

While previous studies propose heuristic bandwidth values for kernel-based continuous DiD estimation, they do not provide any justification for the specific bandwidth choices. This paper is the first to examine the sensitivity of bandwidth in continuous treatment effect estimation and to offer implications for bandwidth selection, with a focus on the kernel-based continuous DiD with DML proposed by Haddad et al. (2024). To illustrate bandwidth sensitivity, I conduct a Monte Carlo simulation study that varies the bandwidth and evaluates its impact on the estimation in terms of bias, standard deviation, root mean squared error (RMSE), and coverage

rate. In addition, Zhang (2025) and Haddad et al. (2024) consider a data generating process with normal-like distributed treatment intensity and uniformly distributed errors, which may not reflect realistic settings. My simulation assumes bimodality in treatment intensity, where each unit is assigned to either a higher or lower treatment group, reflecting more plausible treatment assignment mechanisms. Furthermore, I consider multiple combinations of treatment and control levels to examine how estimator performance depends on data sparsity around those levels. This study not only contributes to the literature on kernel-based continuous DiD, but also offers insights for other kernel-based continuous treatment estimators, including approaches like Huber et al. (2020). The results can help guide empirical researchers in choosing appropriate bandwidths when applying such methods in practice.

This study finds that the choice of bandwidth plays an important role in continuous treatment effect estimation. The results confirm the bias–variance tradeoff and show that the optimal bandwidth varies across treatment levels and sample sizes. When the treatment level is close to the tails of distribution, the estimator may remain biased even with very narrow bandwidths, leading to inconsistency. In addition, the best bandwidth for minimizing RMSE is often different from the one that optimizes coverage. These results highlight the complexity of bandwidth selection and motivate a more detailed investigation through simulation.

The structure of the paper is as follows. Section 2 introduces the methodology of kernel-based continuous DiD with DML by Haddad et al. (2024) and highlights its challenges in bandwidth selection. Section 3 describes the simulation setup, presents the results, and discusses their implications as well as practical issues. Section 4 concludes the paper.

Chapter 2

Methodology

2.1 Binary DiD Methods

2.1.1 Canonical DiD

First, I briefly introduce Difference-in-Differences (DiD) in the two-period binary treatment setting, where some units receive treatment in the post-treatment period and others remain untreated in both periods. I adopt the potential outcome framework (Rubin, 1974), and define $Y_{i,t}(d)$ as the potential outcome of unit i at time $t \in \{0,1\}$ under treatment status $d \in \{0,1\}$. For each unit i, I observe only the realized outcome $Y_{i,t} = Y_{i,t}(D_i)$ and the binary treatment assignment $D_i \in \{0,1\}$, which is assumed to be received only at $t = 1$ if treated. Then, the average treatment effect on the treated (ATT) is defined as:

$$\theta_0 \coloneqq ATT = \mathbb{E}\left[Y_{i,1}(1) - Y_{i,1}(0) \mid D_i = 1\right]. \tag{2.1}$$

To identify the ATT by canonical DiD, the following assumptions need to hold:

Assumption 1 (Parallel Trends):

$$\mathbb{E}\left[Y_{i,1}(0) - Y_{i,0}(0) \mid D_i = 1\right] = \mathbb{E}\left[Y_{i,1}(0) - Y_{i,0}(0) \mid D_i = 0\right]. \tag{2.2}$$

Assumption 2 (No Anticipation):

$$\forall i, \text{ s.t. } D_i = 1,\ Y_{i,0}(0) = Y_{i,0}(1). \tag{2.3}$$

Assumption 1 ensures that the expected outcomes of the treatment and control groups would evolve in parallel if both remain untreated. This isolates contamination by confounders, which can disproportionally affect the outcome over time. Assumption 2 ensures the outcome of treatment units in the pre-treatment period do not depend on the treatment status. This implies treatment units do not change their behavior in the pre-treatment period due to the fact that they will receive the treatment. Under these assumptions, the ATT can be identified

as the difference-in-differences of the observed outcomes:

$$\begin{aligned}
\theta_0 &= \mathbb{E}\left[Y_{i,1}(1) - Y_{i,1}(0) \mid D_i = 1\right] \\
&= \left(\mathbb{E}\left[Y_{i,1}(1) \mid D_i = 1\right] - \mathbb{E}\left[Y_{i,0}(0) \mid D_i = 1\right]\right) - \left(\mathbb{E}\left[Y_{i,1}(0) \mid D_i = 1\right] - \mathbb{E}\left[Y_{i,0}(0) \mid D_i = 1\right]\right) \\
&= \mathbb{E}\left[Y_{i,1}(1) - Y_{i,0}(1) \mid D_i = 1\right] - \mathbb{E}\left[Y_{i,1}(0) - Y_{i,0}(0) \mid D_i = 1\right] \quad \text{(by Assumption 2)} \\
&= \mathbb{E}\left[Y_{i,1}(1) - Y_{i,0}(1) \mid D_i = 1\right] - \mathbb{E}\left[Y_{i,1}(0) - Y_{i,0}(0) \mid D_i = 0\right] \quad \text{(by Assumption 1)} \\
&= \mathbb{E}\left[Y_{i,1} - Y_{i,0} \mid D_i = 1\right] - \mathbb{E}\left[Y_{i,1} - Y_{i,0} \mid D_i = 0\right].
\end{aligned}$$

This implies that the ATT is the difference in expected outcome changes between the treatment and control groups. Figure 2.1 provides a graphical representation of the parallel trends assumption and the resulting ATT.

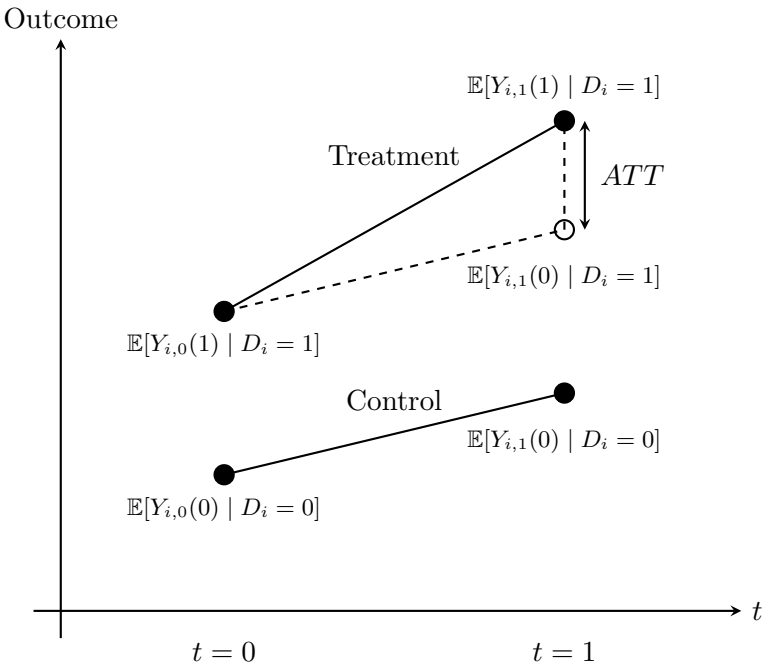

Figure 2.1: Illustration of parallel trends and *ATT*

Although the ATT can be estimated by taking a difference of sample means, the most popular way is to use a two-way fixed effects (TWFE) estimator:

$$Y_{i,t} = \alpha_i + \phi_t + \beta\left(D_i \cdot \mathbb{1}\left[t = 1\right]\right) + \epsilon_{i,t}, \tag{2.4}$$

where α_i and ϕ_t denote unit- and time-specific fixed effects, respectively, and $\epsilon_{i,t}$ denotes an error term. Under the TWFE specification above, the ordinary least squares (OLS) estimate $\hat{\beta}$ is equivalent to the ATT estimate $\hat{\theta}$.

2.1.2 Semi-parametric DiD with Conditional Parallel Trends

Recently, a growing body of literature has extended the canonical DiD by relaxing these assumptions (Roth et al., 2023). One strand of the literature relaxes the parallel trends by

accommodating covariates, which affect both outcomes and the treatment status. When covariates are not balanced between the treatment and control groups, then the parallel trends assumption may be implausible (Abadie, 2005). Whereas the presence of covariates itself is not problematic if their effects on the outcome are time-invariant, such an imbalance in covariates can lead to time-varying confounders. For example, states with Democratic preferences tended to implement Medicaid expansions more frequently and may have experienced distinct macroeconomic shocks that vary over time (Roth et al., 2023). In such cases, it is more plausible to assume the parallel trends only conditional on covariates X_i:

Assumption 3 (Conditional Parallel Trends for Binary Treatment):

$$\mathbb{E}\left[Y_{i,1}\left(0\right) - Y_{i,0}\left(0\right) \mid D_i = 1, X_i\right] = \mathbb{E}\left[Y_{i,1}\left(0\right) - Y_{i,0}\left(0\right) \mid D_i = 0, X_i\right]. \tag{2.5}$$

Under Assumption 3, Abadie (2005) proposes a semi-parametric DiD, which uses inverse probability weighting (IPW) based on the propensity score $P(D = 1 \mid X)$. The estimator addresses the violation of the parallel trends, caused by compositional differences in covariates between the treatment and control groups. In addition, it can allow for heterogeneity in treatment effects characterized by covariates. For panel data, Abadie (2005) identifies the ATT as follows:

$$\theta_0 = \mathbb{E}\left[\frac{Y_{i,1} - Y_{i,0}}{P\left(D_i = 1\right)} \cdot \frac{D_i - P\left(D_i = 1 \mid X_i\right)}{1 - P\left(D_i = 1 \mid X_i\right)}\right]. \tag{2.6}$$

where the propensity score $P(D = 1 \mid X)$ can be estimated by parametric or non-parametric methods as the first step.

2.1.3 Integration with DML

Although standard non-parametric methods can be applied for the first-step estimation, naive plug-in estimates by machine learning methods can cause regularization bias and slower convergence, while they can handle a large number of covariates and provide flexible estimation (Chernozhukov et al., 2018). Chang (2020) extends the semi-parametric DiD by introducing the double/debiased machine learning (DML). Based on (2.6), Chang (2020) defines the following score function:

$$\psi\left(\theta_0, p_0, \eta_0\right) = \frac{D - g\left(X\right)}{p_0\left(1 - g\left(X\right)\right)}\left(Y_1 - Y_0 - l\left(X\right)\right) - \theta_0, \tag{2.7}$$

where $p_0 = P\left(D = 1\right)$, and η_0 denotes the infinite-dimensional nuisance parameter, consisting of the following components:

$$\eta_0 = \left(P(D = 1 \mid X), \mathbb{E}[Y_1 - Y_0 \mid X, D = 0]\right) \equiv \left(g(X), l(X)\right).$$

This score function satisfies the Neyman orthogonality, which ensures that the Gateaux derivative of the score function is zero with respect to infinite-dimensional nuisance parameters (Chang, 2020). Chang (2020) combines this orthogonal score with k-fold cross-fitting, where the data is randomly partitioned into k subsets, and the parameters for the k-th fold are estimated

using the remaining $k-1$ auxiliary subsets. Let I_k denote the index set of observations in the k-th fold. Then, for each fold k, the estimator is constructed as follows:

$$\hat{\theta}_k = \frac{1}{n} \sum_{i \in I_k} \frac{D_i - \hat{g}_k(X_i)}{\hat{p}_k (1 - \hat{g}_k(X_i))} \left(Y_{i,1} - Y_{i,0} - \hat{l}_k(X_i) \right), \tag{2.8}$$

where $\hat{p}_k$ is estimated by the following sample analogue $\frac{1}{n} \sum_{i \in I_k^c} D_i$, and $\left(\hat{g}_k, \hat{l}_k \right)$ can be estimated by any machine learning methods using auxiliary sample I_k^c. Finally, the ATT estimator $\hat{\theta}$ is obtained by averaging over the k estimates $\hat{\theta}_k$. As a result, Neyman orthogonality and cross-fitting can avoid regularization bias and slower convergence, resulting in $\sqrt{N}$-consistent and asymptotically normal estimators, while preserving the flexibility of machine learning methods (Chang, 2020).

2.2 Continuous DiD Methods

2.2.1 Continuous DiD with DML

Another strand of the literature extends the canonical DiD to settings where a treatment variable is continuous. In the two-period continuous treatment setting, the ATT is defined as:

$$\theta_0(d) \coloneqq ATT(d) = \mathbb{E}\left[Y_{i,1}(d) - Y_{i,1}(0) \mid D_i = d \right]. \tag{2.9}$$

In many practical settings, treatment variables are often continuous rather than binary, yet canonical DiD specifications are commonly applied. For example, Ananat et al. (2024) analyze the labor supply effects of the temporary expansion of the Child Tax Credit (CTC) in 2021, using canonical DiD specifications that incorporate continuous treatment measures such as the increase in CTC benefits and the change in return to work. Although the canonical DiD has been applied to continuous outcomes in empirical studies, its theoretical properties under continuous treatments had remained underexplored. Callaway et al. (2025) shows that the two-way fixed effects (TWFE) estimator can yield counterintuitive causal parameters with negative weights, even when all treatment effects are non-negative.[1] Among recent advancements in the literature on continuous DiD such as Callaway et al. (2025); D'Haultfœuille et al. (2023); de Chaisemartin et al. (2025), Zhang (2025) proposes continuous DiD with DML under the following conditional parallel trends assumption for continuous treatment:

Assumption 4 (Conditional Parallel Trends for Continuous Treatment):

$$\mathbb{E}\left[Y_{i,1}(0) - Y_{i,0}(0) \mid D_i = d, X_i \right] = \mathbb{E}\left[Y_{i,1}(0) - Y_{i,0}(0) \mid D_i = 0, X_i \right]. \tag{2.10}$$

Assumption 4 ensures that the expected change in outcome is identical between units with treatment intensity d and those in the control group, assuming that neither group receives any

[1] Negative weights can also arise in staggered binary treatment settings (Athey and Imbens, 2022; Callaway and Sant'Anna, 2021; Goodman-Bacon, 2021; Sun and Abraham, 2021; Borusyak et al., 2024; de Chaisemartin and D'Haultfœuille, 2020). Callaway et al. (2025) shows this problem can also arise in non-staggered continuous treatment settings.

treatment. Under Assumption 4, Zhang (2025) shows that, for panel data, the ATT is identified as:

$$\theta_0(d) = \mathbb{E}[Y_{i,1} - Y_{i,0} \mid D_i = d] - \mathbb{E}\left[(Y_{i,1} - Y_{i,0})\,\mathbb{1}[D_i = 0]\,\frac{f_{D|X}(d)}{f_D(d)\,P(D_i = 0 \mid X_i)}\right]. \quad (2.11)$$

Zhang (2025) uses the following observation to estimate the conditional density:

$$f_{D|X}(d \mid x) = \lim_{h \to 0} \mathbb{E}[K_h(D - d) \mid X = x], \quad K_h(z) := \frac{1}{h} K\left(\frac{z}{h}\right). \quad (2.12)$$

where $K_h(z)$ is a kernel function and h is a bandwidth that controls the amount of smoothing.

In the continuous treatment setting, observations with exactly the same treatment level as a given value rarely exist. Therefore, causal effects must be evaluated using local neighborhoods of the treatment level, which motivates the use of kernel weighting. The kernel function assigns larger weights to units whose treatment intensity is closer to the treatment level, and smaller weights to those further away. This enables estimation of ATT at a particular continuous treatment level by taking a weighted average over nearby observations, analogous to comparing treated and untreated groups in binary settings. As a result, the performance of the estimator is highly sensitive to the choice of bandwidth, which is examined in detail through a Monte Carlo simulation in the following sections.

Using kernel weighting, Zhang (2025) shows that ATT can be expressed as:

$$\theta_0(d) = \lim_{h \to 0} \mathbb{E}\left[(Y_{i,1} - Y_{i,0})\,\frac{K_h(D_i - d)\,P(D_i = 0 \mid X_i) - \mathbb{1}[D_i = 0]\,\mathbb{E}[K_h(D_i - d) \mid X_i]}{f_D(d)\,P(D_i = 0 \mid X_i)}\right]. \quad (2.13)$$

Based on (2.13), Zhang (2025) constructs a Neyman orthogonal score function and proposes an estimation procedure using cross-fitting, which involves the kernel function and the bandwidth. Zhang (2025) shows that the resulting estimator is asymptotically normal under some regularity conditions.

2.2.2 Extension to Non-zero Treatment Intensities

Whereas the method by Zhang (2025) assumes that all units in the control group receive no treatment, it is often possible that every unit receives positive treatment, and there are no "pure" control units with treatment intensity 0. For example, all vaccination rates per region are non-zero, and one may want to estimate the causal effect of increasing the vaccination rate from 50% to 70%. In such cases, the ATT comparing treatment intensities d and d' is defined as:

$$\theta_0(d, d') := ATT(d, d') = \mathbb{E}\left[Y_{i,1}(d) - Y_{i,1}(d') \mid D_i = d\right]. \quad (2.14)$$

The $ATT(d, d')$ captures the causal effect of increasing the treatment intensity from d' to d for treatment units receiving d. Building on Abadie (2005); Zhang (2025), Haddad et al. (2024) extends it to the settings where treatment intensities are non-zero under a stronger parallel

trends assumption for non-zero treatment intensities, in the two-period setting[2]:

Assumption 5 (Conditional Parallel Trends for Non-zero Continuous Treatment):

$$\mathbb{E}[Y_{i,1}(d') - Y_{i,0}(d'') \mid D_{i,1} = d, D_{i,0} = d'', X_i] = \mathbb{E}[Y_{i,1}(d') - Y_{i,0}(d'') \mid D_{i,1} = d', D_{i,0} = d'', X_i]. \tag{2.15}$$

When both groups receive treatment with intensity d'' in the pre-treatment period, and, in the post-treatment period, the treatment and control groups receive treatment intensities d, d', respectively, then Assumption 5 imposes conditional parallel trends, stating that the expected change in outcome when they receive treatment d'' in the pre-treatment period and d' in the post-treatment period is identical between the treatment and control groups. Figure 2.2 provides a graphical representation of the conditional parallel trends assumption and the resulting ATT. For notational simplicity, conditional expectations are abbreviated by omitting expressions such as $D_{i,1} = d', D_{i,0} = d''$.

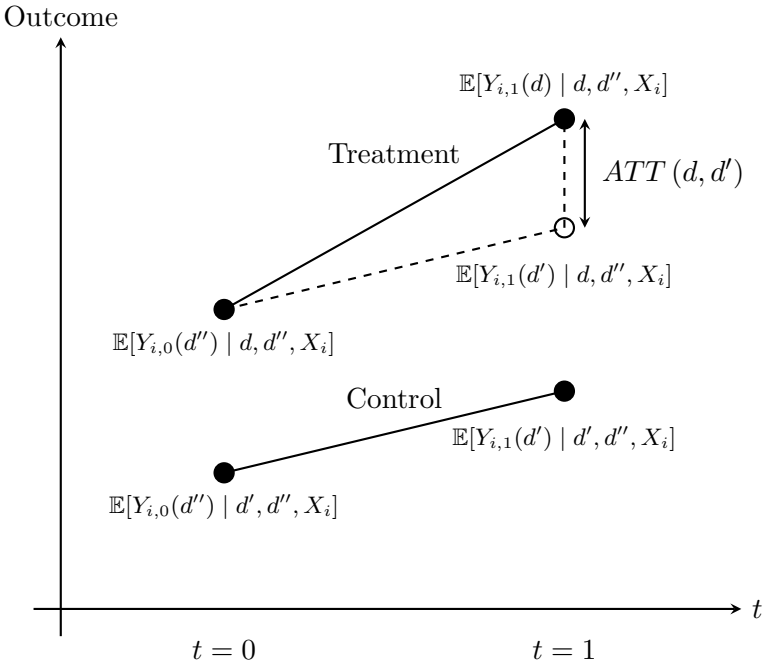

Figure 2.2: Illustration of conditional parallel trends and $ATT\,(d, d')$

Under Assumption 5, Haddad et al. (2024) proposes continuous DiD with DML model that can identify the ATT, and satisfies asymptotic normality under some regularity conditions.[3] As in Zhang (2025), the resulting estimator involves kernel weighting functions, which give non-negative weights to local neighborhoods of the treatment and control levels. Therefore, the estimator depends on the kernel weighting, especially the choice of the bandwidth. This

[2] Haddad et al. (2024) does not restrict it to the two-period setting. For simplicity, I restrict attention to the two-period setting.

[3] I omit the exact expression of the ATT formula by Haddad et al. (2024) due to complexity. See Haddad et al. (2024) for details.

motivates the need to investigate how the bandwidth selection influences the performance of the estimator, which is the focus of the simulation study in the following sections.

2.3 Bandwidth Selection

2.3.1 Bandwidth in Nonparametric Kernel Estimators

Before discussing the bandwidth selection in the continuous treatment estimation, I briefly review how the bandwidth is selected in the literature on non-parametric kernel estimators, including kernel density estimation and kernel regression. In these methods, the bandwidth choice is the most influential factor in estimation performance rather than the kernel choice. The bandwidth choice controls the tradeoff between bias and variance, and, as a result, strongly affects the estimation error. Accordingly, there has been a large body of research on bandwidth selection in both contexts.

In kernel density estimation, the goal is to estimate the density using a bandwidth that minimizes the mean integrated squared error (MISE) of the density. Silverman (1986) proposes a plug-in estimate that determines a bandwidth h, according to Cameron and Trivedi (2005):

$$h = 1.3643 \cdot \delta \cdot N^{-1/5} \cdot \min(s, iqr/1.349), \tag{2.16}$$

where δ is $(\int K(z)^2 dz/(\int z^2 K(z)^2 dz)^2)^{1/5}$, s is the sample standard deviation of the density, and iqr is the sample interquartile range. The $iqr/1.349$ avoids excessive influence from outliers, which would otherwise increase s and lead to a larger h. For the kernel function $K(z)$, Haddad et al. (2024) and also the following simulation adopts the Epanechnikov kernel:

$$K(z) = \frac{3}{4}(1 - z^2)\mathbb{1}[|z| < 1] \tag{2.17}$$

Then, according to Cameron and Trivedi (2005), the Silverman's plug-in estimate for the Epanechnikov kernel is:

$$h = 2.345 \cdot N^{-1/5} \cdot \min(s, iqr/1.349). \tag{2.18}$$

It assumes normality of the density, and derives the optimal bandwidth in terms of MISE. Whereas the true density is not available and may not be normal, Cameron and Trivedi (2005) notes that the plug-in estimate works well on symmetric unimodal densities even if the true distribution is not normal. Note that the assumption regarding the density is only required to determine the plug-in estimate, and the optimal bandwidth can be derived by minimizing the error between the estimated density and the true (but assumed normal) density, which is more straightforward than in kernel regression and continuous treatment estimation.

In kernel regression, the goal is to predict the outcome values using a bandwidth that minimizes the mean squared error (MSE) of the outcome values. Instead of the plug-in estimate, cross-validation has been widely used in practice, because minimizing the MSE analytically requires assumptions about the outcome model and an initial bandwidth, both of which are

unknown (Cameron and Trivedi, 2005). In cross-validation, after splitting the whole dataset into k subsets, one can evaluate the out-of-sample MSE of the k-th subset, using the other $k-1$ subsets for fitting, and estimate the out-of-sample performance of each candidate bandwidth by aggregating the resulting k different MSEs. Varying the bandwidth and repeating the cross-validation procedure reveals the optimal bandwidth among the candidates. Note that cross-validation requires the true outcome values to implement, but does not need assumptions on the outcome model.

In continuous treatment estimation, the goal is neither estimating density of treatment nor predicting outcome values, but estimating the ATT for treatment. Determining a plug-in estimate by minimizing the estimation error of ATT must require assumptions on the treatment effect, the treatment intensity, and the outcome model, which poses more complex challenges and may be less plausible than in kernel density estimation. In addition, unlike kernel regression, none of the treatment effects is available in the dataset, which makes the cross-validation infeasible. Whereas the bandwidth selection in kernel density and kernel regression has been well studied, such challenges in continuous treatment estimation motivate the need to explore how bandwidth should be chosen in this context.

2.3.2 Bandwidth in Continuous DiD with DML

As seen in the previous section, the bandwidth selection in continuous treatment estimation poses a different challenge compared to kernel density estimation and kernel regression. This section presents assumptions on the bandwidth in continuous DiD, and reviews how the bandwidth in continuous treatment effect estimation is selected in the literature.

Zhang (2025); Haddad et al. (2024) impose the following assumptions regarding the rate of convergence of the bandwidth:

Assumption 6 (Rates of Convergence):

(a) The kernel bandwidth h is a deterministic sequence that depends on n and satisfies $nh \to \infty$ and $\sqrt{nh^5} = o(1)$;

(b) There exists a sequence $\epsilon_n \to 0$, such that $h^{-1}\epsilon_n^2 = o(1)$;

(c) There exists a sequence $\epsilon_n \to 0$, such that $h^{-2}\epsilon_n^2 = o(1)$;

These assumptions jointly imply the bandwidth should be neither too small nor too large. Assuming the rate of convergence of the estimators of the nuisance parameters is $\epsilon_N = o\left(N^{-1/4}\right)$, which is standard in the DML literature (Chernozhukov et al., 2018), then the rate of convergence of the bandwidth must satisfy:

$$N^{-1/4} \leq h \leq N^{-1/5}. \tag{2.19}$$

In addition, Zhang (2025); Haddad et al. (2024) propose undersmoothing the bandwidth so that the bias vanishes asymptotically. However, the literature provides no clear guidance on how much undersmoothing is appropriate in practice.

As a result, the bandwidth in continuous treatment estimation is typically chosen heuristically. Zhang (2025) adopts $0.5N^{-1/4}$ as the bandwidth of the continuous DiD with DML.[4] Unlike Zhang (2025), Haddad et al. (2024) defines the bandwidth as $0.5 \cdot 2.34N^{-1/4}$, which is referred to as "Silverman-type" rule of thumb for Epanechnikov kernels.[5] Similarly, the default option in the *causalweight* package is $0.7 \cdot 2.34N^{-1/4} \cdot s$, where s is a sample standard deviation of treatment (Bodory et al., 2025). Unlike other kernel estimation methods, all of the examples adopt order $N^{-1/4}$ instead of $N^{-1/5}$ for undersmoothing. In addition, a constant such as 0.5 or 0.7 is multiplied to control the amount of undersmoothing. However, none of them provide clear justification for why these specific constants are chosen.

Accordingly, there has been no clear guidance on how to determine the bandwidth in the literature. Therefore, this motivates the need to investigate how the bandwidth choice affects estimation performance, which is addressed in the simulation study in the next chapter.

[4] Zhang (2025) defines the bandwidth h as $0.5N^{-4}$. Given Assumption 6, it must be a typo and should be $0.5N^{-1/4}$.

[5] As seen in the previous subsection, it is not exactly the same as the original Silverman's rule because it uses $N^{-1/4}$ instead of $N^{-1/5}$, and does not consider the sample standard deviation.

Chapter 3

Simulation Study

3.1 Simulation Design

3.1.1 Data Generating Process

I introduce the data generating process (DGP) used in the simulation.[1] First, I define a time indicator and divide units into two groups:

$$t_i \sim \text{Bernoulli}\,(0.5)\,. \tag{3.1}$$

Here, t_i indicates whether unit i belongs to the pre-treatment period ($t = 0$) or the post-treatment period ($t = 1$), and the number of units from each period is expected to be equal. While the method by Haddad et al. (2024) accommodates multiple time periods and time-varying treatments, I restrict attention to a two-period setting for simplicity.

Next, I define covariates X_i that influence both treatment and outcome:

$$X_i \sim \mathcal{N}\,(0_k, I_k)\,. \tag{3.2}$$

Here, X_i denotes a $k \times 1$ vector of covariates of unit i, affecting both the treatment D_i and the outcome y_i. The number of covariates k is set to 50 to reflect the curse of dimensionality, and each covariate is independently drawn from a standard normal distribution.

Unlike the simulation settings in Zhang (2025); Haddad et al. (2024), I assume a bimodal distribution of treatment doses, which practitioners may encounter in empirical settings.[2] Bimodality in treatment intensity is observed in Acemoglu and Finkelstein (2008) and revisited by Callaway et al. (2025); Zhang (2025), who investigate the causal effects of the Medicare Prospective Payment System (PPS) introduced in 1983 by defining treatment as the Medicare inpatient

[1] Whereas Haddad et al. (2024) proposes methods for both repeated cross-section and panel data, I only focus on the repeated cross-section setting due to computational constraints. I expect the results for the panel data model to be similar to those for the repeated cross-section model, as shown in Haddad et al. (2024).

[2] For example, some patients with higher levels of risk receive more treatment, and others with lower risk receive less. Other examples include variation in advertising exposure, where individuals may be exposed to a high or low number of ads based on their predicted engagement, and pollution levels, which can differ greatly across regions depending on environmental conditions.

share of each hospital. To reflect such bimodality, I introduce a probability for treatment group assignment:

$$p_i = \frac{1}{1 + \exp\left(-X_i'\gamma\right)}. \tag{3.3}$$

Let p_i denote the probability of receiving higher treatment. When $p_i \geq \frac{1}{2}$, unit i belongs to the higher treatment group; otherwise, to the lower treatment group. Given the group assignment, treatment doses D_i are generated as non-negative continuous values with group-specific location shifts:

$$D_i = \begin{cases} \left|\delta_h + X_i'\alpha + u_i\right|, & \text{if } p_i \geq \frac{1}{2}, \quad u_i \sim \mathcal{N}\left(0,1\right), \\ \left|\delta_l + X_i'\alpha + v_i\right|, & \text{if } p_i < \frac{1}{2}, \quad v_i \sim \mathcal{N}\left(0,1\right), \end{cases} \tag{3.4}$$

where $\delta_h = 5$ and $\delta_l = 2$ for the high and low treatment groups, respectively.[3] The coefficients α and γ are diminishing such that $\alpha_j = \gamma_j = 0.5/j^2$ for $j = 1, \ldots, k$. One example of the treatment distribution is shown in Figure 3.1.

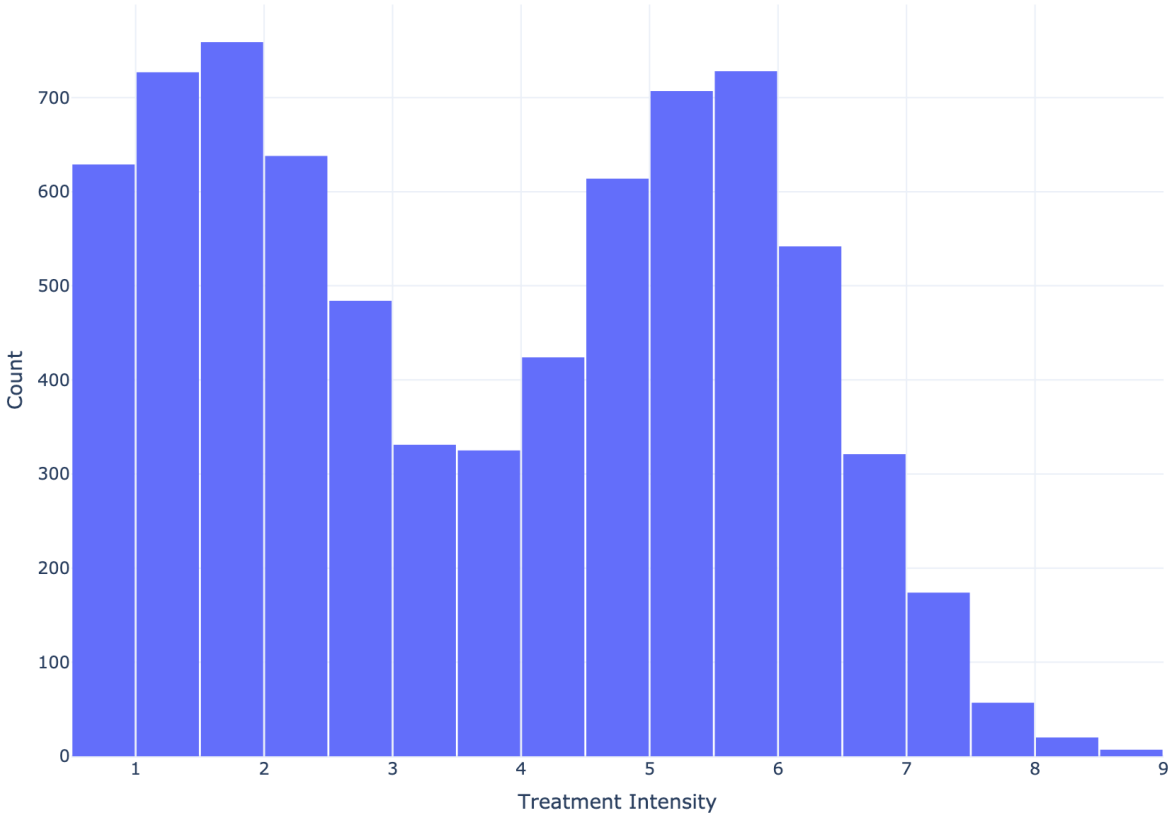

Figure 3.1: Histogram of treatment intensity

Finally, I introduce an outcome equation:

$$y_i = X_i'\beta + \left(1 + D_i^2\right) \cdot t_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}\left(0,1\right). \tag{3.5}$$

[3]Since D_i is defined as the absolute value of a normally distributed variable, the resulting distribution is folded normal. Therefore, the population means slightly deviate from 5 and 2.

The outcome y_i is affected by a nonlinear function of treatment intensity in the post-treatment period ($t = 1$). For example, the ATT comparing treatment levels $d = 5$ and $d' = 2$ equals $25 - 4 = 21$. The coefficients β are diminishing such that $\beta_j = 0.5/j^2$ for $j = 1, \ldots, k$. The relationships among the variables are summarized in Figure 3.2. Covariates X_i influence the treatment assignment probability p_i, the continuous treatment dose D_i, and the outcome y_i. The treatment level is affected by p_i and influences the outcome y_i in the post-treatment period ($t = 1$).

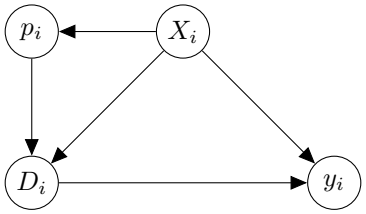

Figure 3.2: Directed acyclic graph (DAG)

3.1.2 Parameter Grid

To assess the bandwidth sensitivity under various scenarios, I define the following parameter grid:

$$h = c \cdot N^{-1/4}, \quad c \in \{0.25, 0.5, 0.75, 1.0, 1.25, 1.5, 1.75, 2.0\}, \tag{3.6}$$

$$N \in \{2000, 4000, 8000\}, \tag{3.7}$$

$$d \in \{3, 4, 5, 6, 7\}, \tag{3.8}$$

where h denotes the bandwidth value. The constant c scales the bandwidth h and affects the level of smoothing. N denotes the number of observations, and d denotes the treatment level. The goal of the simulation is to identify the best bandwidth factor c for each parameter combination and to investigate how it varies across different combinations. I vary the number of observations N to capture how data density affects the bandwidth choice. Furthermore, I consider multiple combinations of treatment and control levels to examine how estimator performance depends on data sparsity around those levels, because the kernel-based estimator heavily relies on local neighborhoods around the reference levels. For simplicity and comparability, I fix the control level d' at 2, and vary the treatment level d from 3 to 7.

For each combination of parameters (N, c, d, d'), I implement Monte Carlo simulations and aggregate 4,000 simulations to ensure reliable estimates.[4] The bandwidth selection is evaluated in terms of bias, standard deviation, root mean squared error (RMSE), and coverage rate of the 95% confidence interval. A bandwidth is considered to be the best when the bias and standard

[4]When both the sample size N and the constant c are too small, the estimation sometimes fails due to a lack of local neighborhoods around the treatment level and control level. To ensure every combination has 4,000 results, I select the first 4,000 successful simulations for each combination. The error rate was highest at 8% when $(N, c, d, d') = (2000, 0.25, 7, 2)$.

deviation are negligibly small, the RMSE is minimized, and the coverage rate is around the nominal rate of 95%. These criteria reflect the bias–variance tradeoffs in finite samples.

3.2 Simulation Results

Building on the simulation setup in the previous section, this section presents results of the simulation study and investigates the relationship between the bandwidth and bias, standard deviation, RMSE, and coverage rate. Figures 3.3, 3.4, 3.5, 3.6, and 3.7 show how the bandwidth affects bias, standard deviation, RMSE, and coverage rate when the treatment level is 3, 4, 5, 6, or 7 and the control level is fixed at 2. Figure A.1 in Appendix A shows distributions of the ATT estimates across the treatment levels and the bandwidth factors.

3.2.1 Bias–Variance Tradeoff

The upper left panels in Figures 3.3, 3.4, 3.5, 3.6, and 3.7 focus on the bias across different treatment levels. For the treatment levels from 3 to 5, the bias increases as the bandwidth gets wider, which is expected in the standard theory of kernel estimators. The larger value of the bandwidth accommodates more observations around the treatment level, thus the variance of the estimator decreases at the expense of the bias. However, for the treatment levels 6 and 7, the bias values shrink as the bandwidth factors get smaller, but converge to positive values after crossing zero, which means the estimator is positively biased and inconsistent. Figure A.2 in Appendix A illustrates how the bandwidth factors affect the absolute bias and compares it across the treatment levels and N. For the treatment levels 6 and 7, the absolute bias plots exhibit convex shapes, whereas those of all other levels are monotonically increasing.

The upper right panels in Figures 3.3, 3.4, 3.5, 3.6, and 3.7 focus on the standard deviation across different treatment levels. As expected, when the bandwidth gets narrower, the standard deviation inflates, and it shrinks as the bandwidth gets wider. A narrower bandwidth implies that the estimator relies on a smaller number of local neighborhoods around the treatment level, which makes the estimate highly sensitive to the data and leads to higher variance. Figure A.3 in Appendix A shows how the bandwidth factors affect the standard deviation across the treatment levels and N. Higher treatment levels are associated with higher standard deviations. In particular, the density around treatment level 7 is the lowest among all levels, resulting in especially high standard deviations at that level.

3.2.2 RMSE-minimizing Bandwidth Choices

The lower left panels in Figures 3.3, 3.4, 3.5, 3.6, and 3.7 focus on the RMSE across different treatment levels. Table 3.1 shows the best bandwidth factor c and the corresponding bandwidth values $c \cdot N^{-1/4}$ that minimize RMSE across treatment level d and the number of observations N. The best choice of the bandwidth factor c in terms of RMSE varies across the treatment levels. When the treatment level is from 3 to 5, the best bandwidth factors are all the same across N,

which are 1.25, 0.75, and 0.75, respectively. This implies the bandwidth values get narrower as N increases, which is consistent with theoretical predictions. As the number of observations increases, the number of local neighborhoods around the treatment level also increases; therefore, the optimal bandwidth in terms of RMSE gets narrower. On the other hand, the best bandwidth factors for the treatment level 6 and 7 increase in N, and the corresponding bandwidth values are almost the same across N, which contradicts the theory. For the treatment level 6, all bandwidth values are slightly below 0.19. For the treatment level 7, the bandwidth values are around 0.12, and even increase as N increases.

Table 3.1: RMSE-minimizing bandwidth factors and corresponding values.

N	$d=3$		$d=4$		$d=5$		$d=6$		$d=7$	
	Factor	Value	Factor	Value	Factor	Value	Factor	Value	Factor	Value
2000	1.25	0.187	0.75	0.112	0.75	0.112	1.25	0.187	0.75	0.112
4000	1.25	0.158	0.75	0.094	0.75	0.094	1.50	0.189	1.00	0.126
8000	1.25	0.132	0.75	0.079	0.75	0.079	1.75	0.185	1.25	0.132

Notes: The column "Factor" reports the RMSE-minimizing bandwidth factor c, and the column "Value" reports the corresponding bandwidth h computed as $h = c \cdot N^{-1/4}$.

The unexpected result observed in the treatment levels 6 and 7 is caused by inconsistency of the estimator. As seen in Figures 3.6, 3.7, and A.2 in Appendix A, the bias in the treatment levels 6 and 7 does not approach zero when the bandwidth factor gets smaller, unlike in the treatment levels from 3 to 5. As a result, shrinking the bandwidth as N increases does not always improve the RMSE, which is the sum of the squared bias and the variance. This explains the distinctive behavior in the treatment levels 6 and 7.

The inconsistency is possibly caused by asymmetry in the local neighborhoods of the treatment levels. It is well known in the literature on the kernel estimator that all the kernel estimators can be subject to boundary bias, which arises near the tails or the boundary of the support (Müller and Stadtmüller, 2002). Cid and von Davier (2015) explains this occurs when kernel weights assigned to observations become asymmetric at the boundary or when observations are not equally spaced. In my simulation study, there is no boundary in the right tail; however, treatment levels 6 and 7 are close to the right tails and the gradients of the density at those levels are steeper than at the other levels as seen in Figure 3.1.

3.2.3 Coverage-optimizing Bandwidth Choices

The lower right panels in Figures 3.3, 3.4, 3.5, 3.6, and 3.7 focus on the coverage rate across different treatment levels. Overall, the coverage rates do not converge to the nominal rate of 95% and are rarely close to it in almost all cases, highlighting the challenge of selecting an appropriate bandwidth for inference. Table 3.2 summarizes the best bandwidth in terms of the coverage rate by the treatment levels d and N. Compared with Table 3.1, the best bandwidth

in terms of the coverage rate is rarely the same as the bandwidth that minimizes RMSE. Except for the treatment levels 3 and 6, the bandwidth factor around 1.0 seems a relatively good choice in practice, although it is not always the best.

Table 3.2: Coverage-optimizing bandwidth factor and corresponding bandwidth values.

N	$d = 3$		$d = 4$		$d = 5$		$d = 6$		$d = 7$	
	Factor	Value	Factor	Value	Factor	Value	Factor	Value	Factor	Value
2000	2.00	0.299	1.25	0.187	1.00	0.150	1.75	0.262	1.00	0.150
4000	2.00	0.252	1.00	0.126	0.75	0.094	1.75	0.220	1.25	0.157
8000	2.00	0.212	0.75	0.079	0.50	0.053	1.00	0.106	1.50	0.159

Notes: The column "Factor" reports the coverage-optimizing bandwidth factor c, and the column "Value" reports the corresponding bandwidth h computed as $h = c \cdot N^{-1/4}$.

3.2.4 Findings and Implications

Overall, as expected, the bias–variance tradeoff is observed in most cases, whereas inconsistency is observed in some cases. In regular cases, when the bandwidth gets narrower, the bias approaches zero while the standard deviation increases, and vice versa. The resulting RMSE, defined as the sum of the squared bias and the variance, exhibits convex shapes, which attain minimum values in the middle of the interval of the bandwidth factor c. However, the best choice of the bandwidth factor c varies across the treatment levels, depending on peripheral densities and gradients.

When the peripheral density around the treatment is dense and not so steeply changing, the bias goes to zero when the bandwidth gets narrower. In such cases, the best choice of the bandwidth factor is considered to be around 1.0 across all patterns of N, as seen with 0.75 in the treatment levels 4 and 5, and 1.25 in the treatment level 3. However, when the gradient of the density is steep or when the treatment level is close to the tails, such as the treatment levels 6 or 7, the bias does not approach zero even when the bandwidth gets narrower. As a result, the best choice of the bandwidth factor is not constant across N, which complicates the bandwidth selection. Moreover, even when the peripheral density is dense and the gradient of the density is not so steep, the coverage rate still fails to reach the nominal rate of 95%. In addition, the best choice in terms of the coverage rate is often not the same as the best choice in terms of RMSE, implying a tradeoff in bandwidth selection depending on the evaluation criteria. These findings highlight the difficulty in selecting an appropriate bandwidth in continuous treatment estimation, and contribute to the literature not only on kernel-based continuous DiD estimators but also on a broader class of continuous treatment effect estimators that rely on kernel weighting.

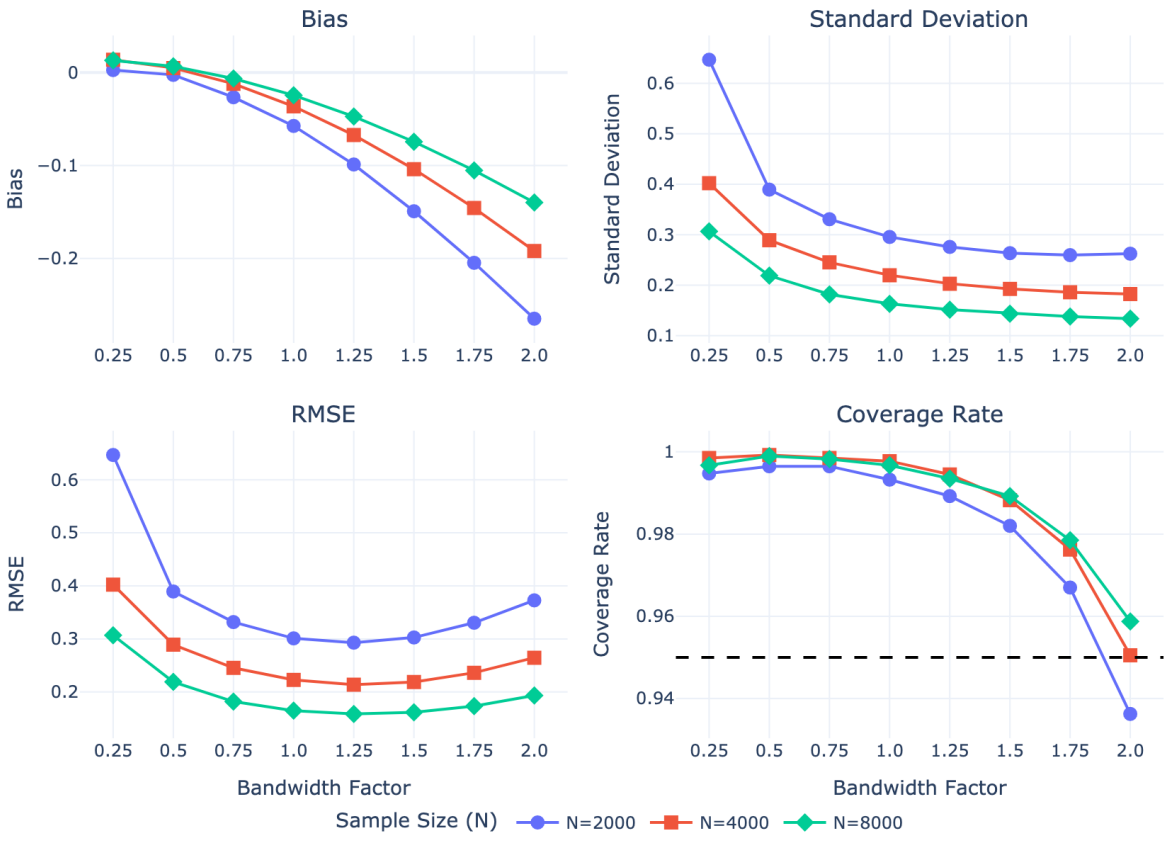

Figure 3.3: Summary of simulations with treatment=3, control=2

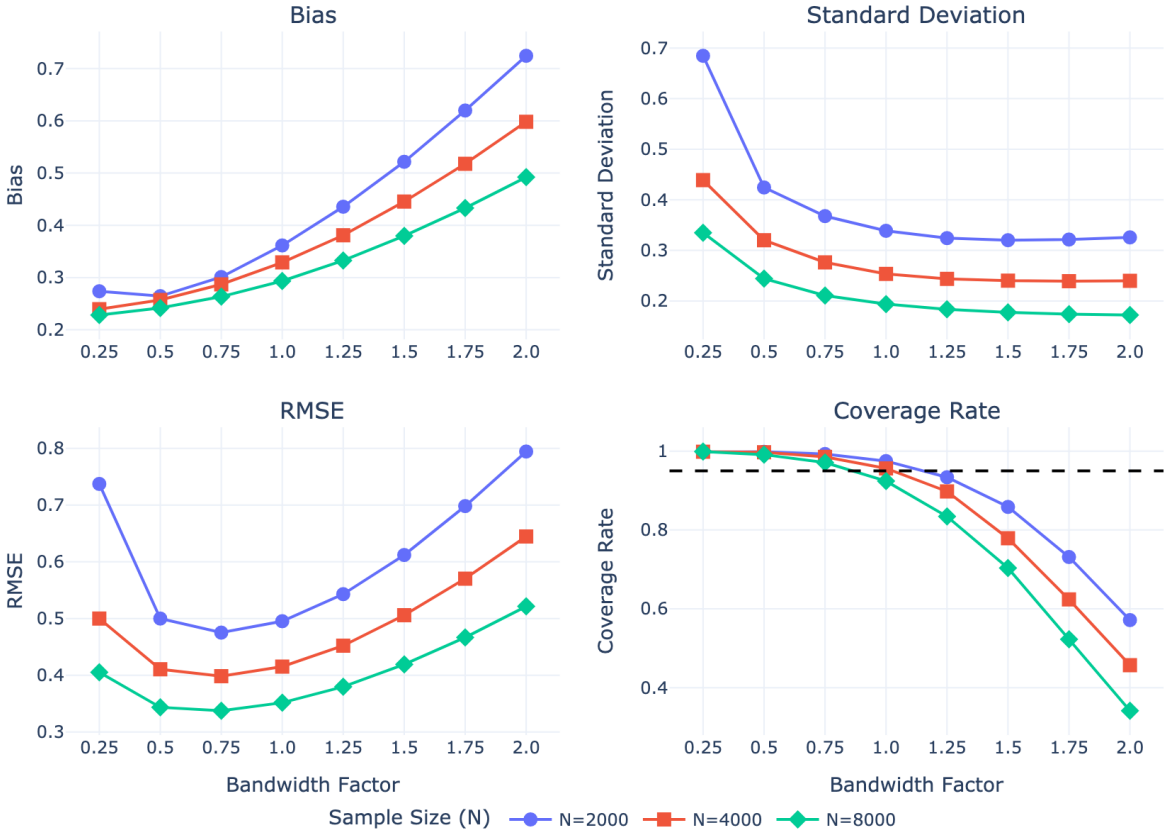

Figure 3.4: Summary of simulations with treatment=4, control=2

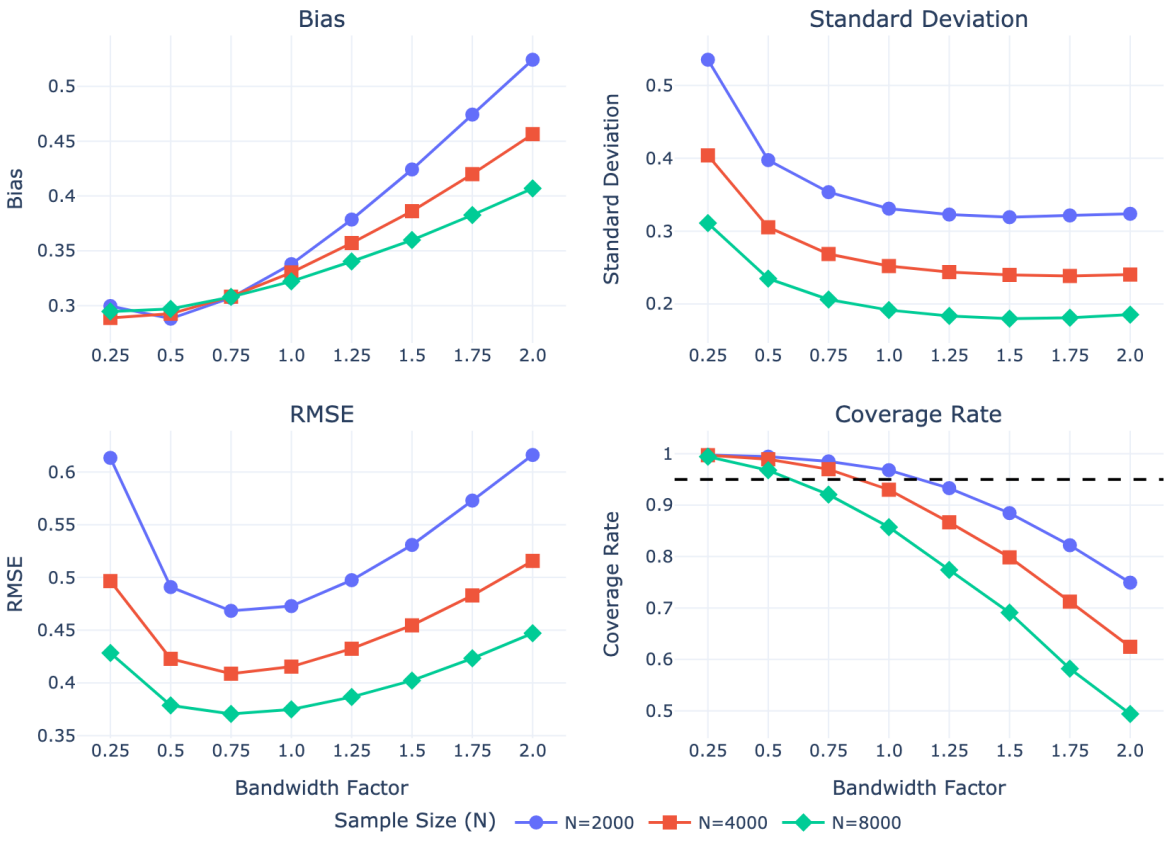

Figure 3.5: Summary of simulations with treatment=5, control=2

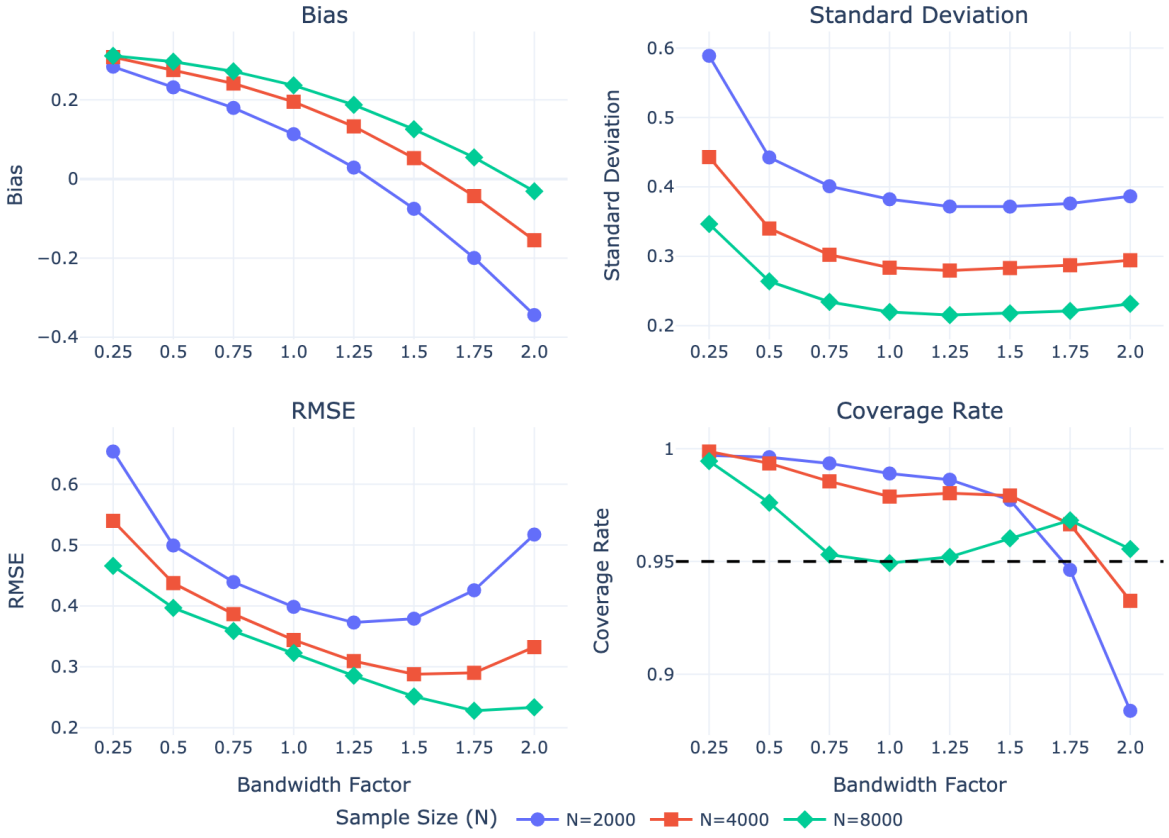

Figure 3.6: Summary of simulations with treatment=6, control=2

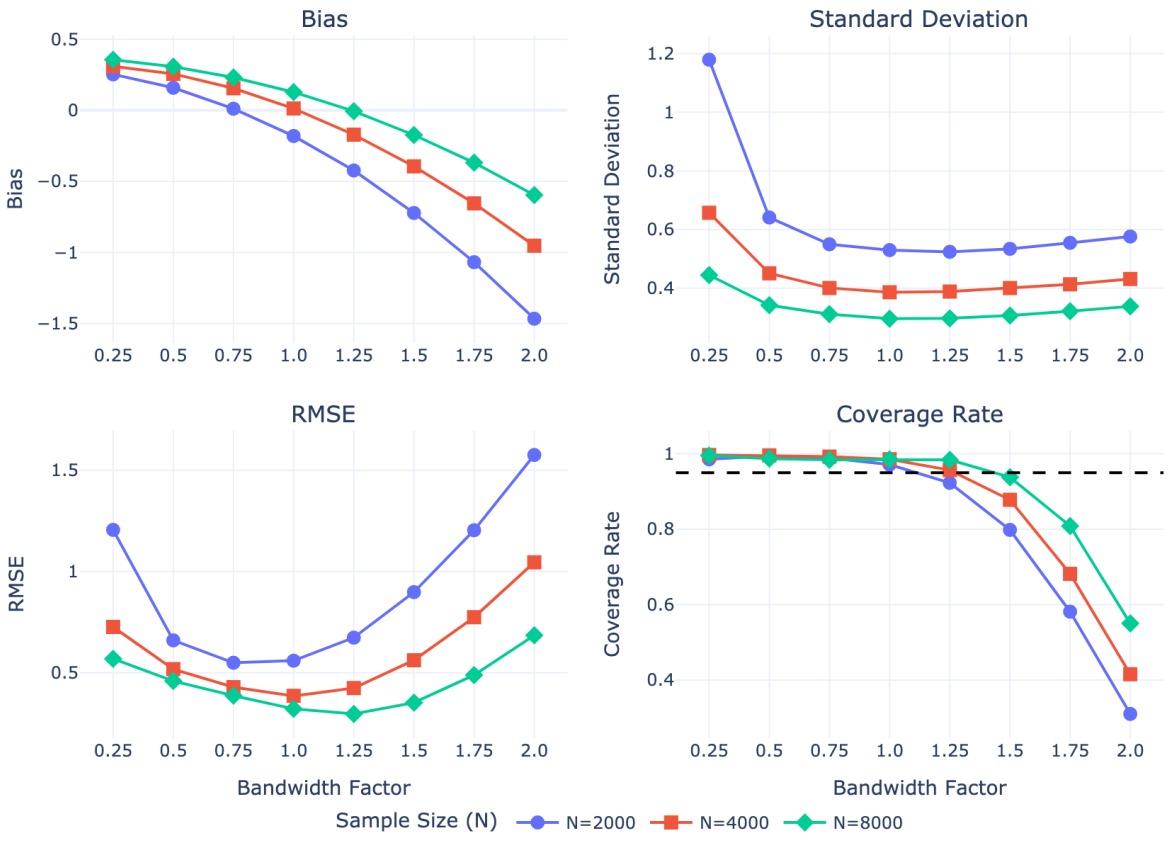

Figure 3.7: Summary of simulations with treatment=7, control=2

3.3 Discussion

3.3.1 Validity of Package Default Option

The default option of the R package `causalweight` by Bodory et al. (2025) is $0.7 \cdot 2.34 n^{-1/4} \cdot s$. As the sample standard deviation s in my data generating process (DGP) is approximately 2.12, the bandwidth suggested by the default option is $3.4756 \cdot N^{-1/4}$, resulting in bandwidth values of 0.52, 0.44, and 0.37 for $N = 2000, 4000, 8000$, respectively. Comparing these with the best bandwidth values in Table 3.1 and 3.2, the default values are considerably wider than the best values, which suggests that narrowing the bandwidth improves both RMSE and coverage. Although the default option may work well for unimodal treatment intensity distributions, such as the example of Bodory et al. (2025) or the simulation study in Haddad et al. (2024), this study demonstrates that the default setting may largely deviate from the best performance, especially when the treatment density is irregular or multimodal. A practical recommendation from this simulation study is to start with a bandwidth factor around 1, that is, $h = c \cdot N^{-1/4}, c \in \{0.75, 1.0, 1.25\}$, which seems to work well, if not best, when the peripheral density around the treatment is dense and not so steeply changing. However, as the computational burden prevents trying another DGP, this observation is also based on the one specific DGP, which leads to a limitation of this study. Therefore, the rule from this study may not be generally applicable,

especially if the variance of the treatment intensity is much larger. For example, the number of local neighborhoods must differ for the treatments with variances of 1 and 50 even if the total number of observations N is the same for both cases.

3.3.2 Bandwidth Rules for Continuous Treatment Effects

It may be possible to create a rule of thumb or plug-in estimate to determine the bandwidth in continuous treatment effect estimation. In kernel density estimation (KDE), Silverman's plug-in estimate can help choose an appropriate bandwidth, using both the number of observations N and the sample standard deviation s. As the entire density is estimated and evaluated in KDE, the total number of observations N and the overall sample variance s matter. In contrast to Silverman's rule, bandwidth selection in continuous treatment effect estimation must consider the local density, because only the local neighborhoods around the treatment and control levels affect estimation performance. As seen in the simulation study, not only the number of local neighborhoods but also the symmetry of the density around the evaluation point plays an important role. Developing a rule-based approach that considers both the density and symmetry of local neighborhoods will be an important topic for future research.

3.3.3 Adaptive Bandwidth

It is also possible to use different bandwidth values for the treatment level and the control level. Haddad et al. (2024) proposes a kernel-based continuous DiD using a common bandwidth for both levels. However, the local behavior around these levels is not always the same. For example, if the density around the treatment level is sparse while that around the control level is dense, applying a wider bandwidth to the treatment level and a narrower one to the control level may lead to better estimation performance. Varying the bandwidth depending on local density has been studied in kernel density estimation (KDE); see, for example, Terrell and Scott (1992). However, in continuous treatment effect estimation, using two different bandwidths complicates the problem even further. It doubles the number of parameters to select, and how their joint choice affects estimation is a non-trivial issue. Although this approach poses additional challenges, adaptive bandwidth selection remains an important direction for future research, especially in settings where the distribution of treatment intensity varies sharply across the region.

Chapter 4

Conclusion

This study investigates how the choice of bandwidth affects estimation performance in continuous treatment effect estimation, based on the kernel-based continuous DiD with high-dimensional covariates proposed by Haddad et al. (2024). Using Monte Carlo simulations, I evaluate how the bandwidth influences the bias, standard deviation, RMSE, and coverage rate across different treatment levels and sample sizes.

The results confirm the standard bias–variance tradeoff relationships in most settings. When the bandwidth becomes narrower, the bias tends to decrease while the standard deviation increases. However, for treatment levels close to the tails of the distribution, the bias does not converge to zero, leading to inconsistency of the estimator. In addition, the best bandwidth for minimizing RMSE is not always the same as the bandwidth that maximizes coverage. These findings highlight the complexity of bandwidth selection in continuous treatment effect estimation.

The results suggest that the default bandwidth in the `causalweight` package may be too wide, and narrowing it improves both RMSE and coverage in many cases. A practical recommendation from this study is to consider a bandwidth factor around 1, i.e., $c \in \{0.75, 1.0, 1.25\}$. This range seems to perform reasonably well when the peripheral density around the treatment level is not sparse or rapidly changing.

There are several limitations in this study. The analysis is based on one specific DGP due to computational constraints, and the results may not generalize to other settings with different treatment intensity distributions. Future research may explore rule-based bandwidth selection, such as plug-in methods that incorporate local density information. Moreover, the study only considers a common bandwidth for the treatment and control levels. Adaptive bandwidth choices, where different bandwidths are used for the treatment and control levels, deserve further investigation, especially in settings where the density sharply varies across the region.

Appendix A

Figures

Figure A.1: Box plot of centered ATT estimates

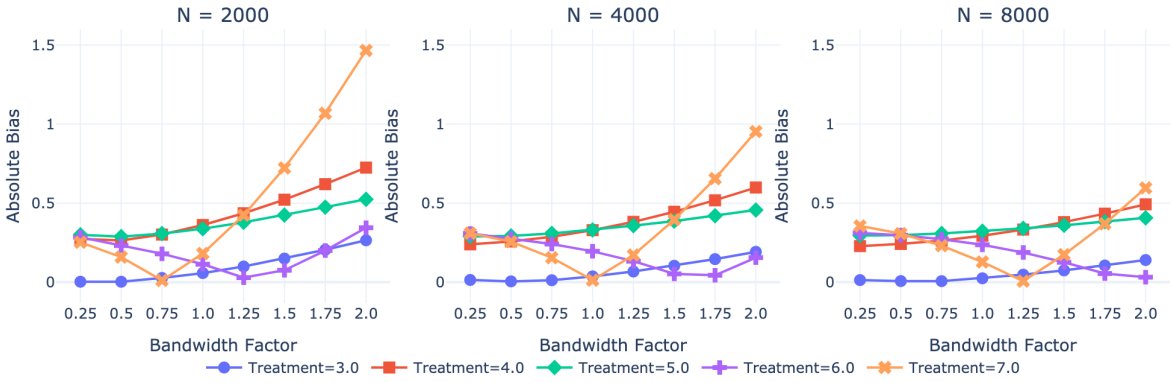

Figure A.2: Absolute bias

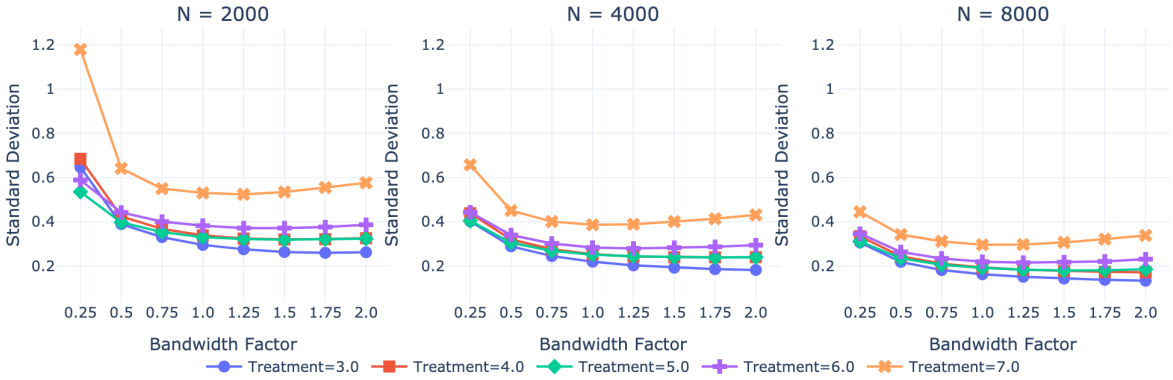

Figure A.3: Standard deviation

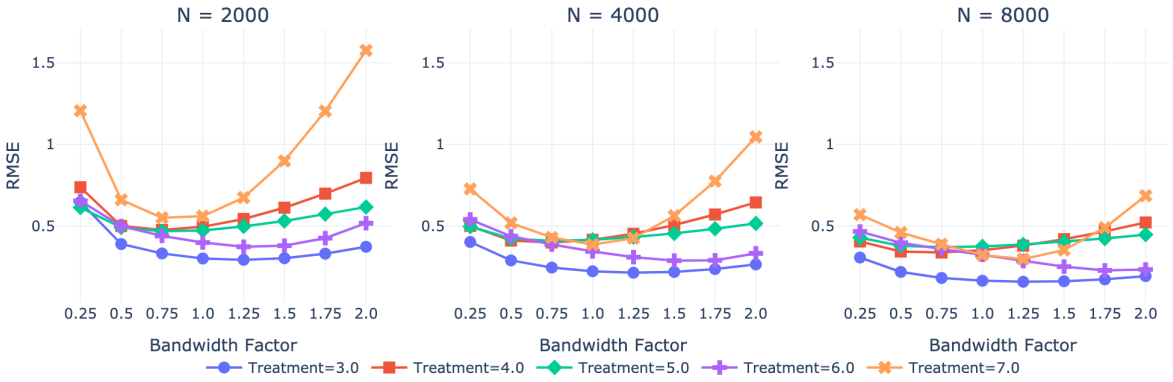

Figure A.4: RMSE

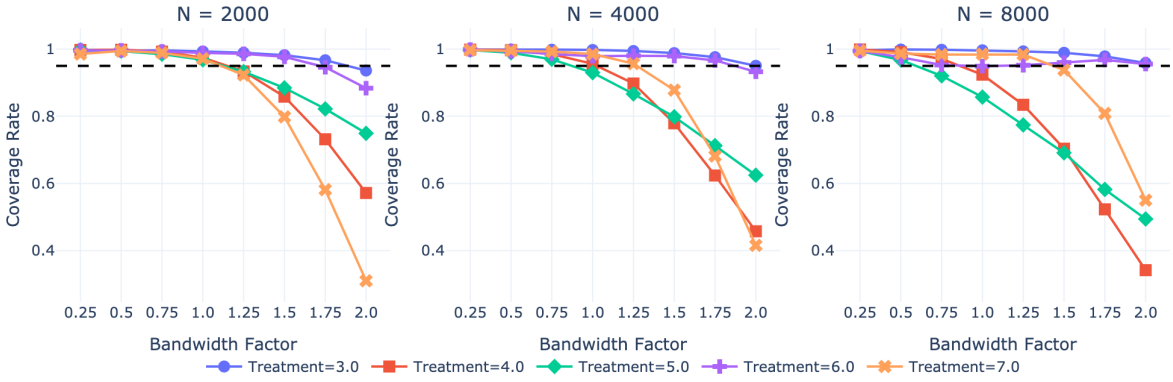

Figure A.5: Coverage rate

Appendix B

Implementation

B.1 Estimation Setup

The simulation was executed in R, using the `didcontDML` function in the `causalweight` package (Bodory et al., 2025), and the results were summarized using SQL. The corresponding codes are provided in Listing B.1 in Appendix B.2.1 and Listing B.2 in Appendix B.2.2.

B.1.1 Model Configuration

Except for the dataset and the treatment and control levels, the other options in the function are set to their default values, which follow the same settings as in Haddad et al. (2024). For DML, I adopt lasso regression, which can handle high-dimensional covariates using an L1 regularization term, and perform 3-fold cross-fitting to estimate nuisance parameters while avoiding overfitting. Although other machine learning methods—such as random forest, XGBoost, and SVM—are available in the package and may achieve higher accuracy, they are far more computationally intensive; therefore, my simulation builds on lasso regression. For the same reason, I restrict the number of folds in cross-fitting to 3. To avoid excessive influence from observations with extreme weights in inverse probability weighting (IPW), I discard those receiving a weight greater than 10% in the IPW computation of any conditional mean outcome, given the treatment and time period (Haddad et al., 2024).

B.1.2 Computational Environment

All simulations were executed on Google Cloud Platform (GCP), using five preemptible virtual machines (VMs) running in parallel. Each VM was either a `c2d-highcpu-32` instance with 32 vCPUs and 64 GB of memory or a `c2d-highcpu-16` instance with 16 vCPUs and 32 GB of memory. The total computation time across all machines was approximately 545 hours. This computational burden limited the feasibility of exploring more data generating processes, additional combinations of bandwidth and control levels, larger sample sizes, and other machine learning methods.

B.2 Programs

B.2.1 R Code for Simulation

```r
library(causalweight)
library(doParallel)
library(foreach)

# Function to generate data according to a specified DGP
generate_data <- function(N, seed = NULL) {
  if (!is.null(seed)) set.seed(seed)  # Set seed for reproducibility

  # Assign pre-treatment and post-treatment periods randomly
  t <- rbinom(N, 1, 0.5)

  # Generate high-dimensional covariates
  k <- 50
  X <- matrix(rnorm(N * k, 0, 1), nrow = N, ncol = k)

  # Generate treatment group assignment probabilities using a logistic
     function
  gamma <- 0.5 / ((1:k)^2)
  p <- 1 / (1 + exp(- X %*% gamma))

  # Define two different treatment intensities based on probability threshold
  delta_h <- 5
  delta_l <- 2
  alpha <- 0.5 / ((1:k)^2)
  u <- rnorm(N, 0, 1)
  v <- rnorm(N, 0, 1)
  d_h <- abs(delta_h + X %*% alpha + u)  # High intensity
  d_l <- abs(delta_l + X %*% alpha + v)  # Low intensity
  d <- ifelse(p >= 0.5, d_h, d_l)        # Assign intensity

  # Generate the outcome variable as a function of covariates and treatment
  beta <- 0.5 / ((1:k)^2)
  e <- rnorm(N, 0, 1)
  y <- X %*% beta + (1 + d^2) * t + e

  return(list(y = y, d = d, t = t, X = X))
}

# Function to run the simulation study
run_simulation <- function(dgp_func, S, treatment, control, N_list,
    bwfactor_list, seed) {
  start_time <- Sys.time()
  cat(sprintf("[START] Simulation started at %s\n", start_time))

  # Set up parallel backend using all available cores
```

```
cl <- makeCluster(parallel::detectCores())
registerDoParallel(cl)

# Create a grid of all combinations of N and bwfactor
combo_grid <- expand.grid(
  N = N_list,
  bwfactor = bwfactor_list,
  stringsAsFactors = FALSE
)
K <- nrow(combo_grid)

# Run simulations in parallel for each combination
results_list <- foreach(j = 1:K, .packages = c("causalweight")) %dopar% {
  N <- combo_grid$N[j]
  bwfactor <- combo_grid$bwfactor[j]

  atet_vec <- numeric(S)
  se_vec <- numeric(S)

  # Repeat simulation S times for each combination
  for (i in 1:S) {
    current_seed <- seed + i

    result <- tryCatch({
      data <- dgp_func(N, seed = current_seed)

      # Estimate ATET using didcontDML
      res <- didcontDML(
        y = data$y, d = data$d, t = data$t, controls = data$X,
        dtreat = treatment, dcontrol = control,
        MLmethod = "lasso",
        bw = N^(-1/4),
        bwfactor = bwfactor
      )

      list(atet = res$ATET, se = res$se)

    }, error = function(e) {
      # The estimation can fail due to lack of the local neighborhood when
          the bandwidth is too small.
      # In such cases, we return NA for atet and se.
      list(atet = NA, se = NA)
    })

    atet_vec[i] <- result$atet
    se_vec[i] <- result$se
  }

  list(atet = atet_vec, se = se_vec)
```

```
}

stopCluster(cl)  # Stop parallel backend

# Combine results into matrices
atet_matrix <- do.call(cbind, lapply(results_list, function(x) x$atet))
se_matrix <- do.call(cbind, lapply(results_list, function(x) x$se))

# Label columns by parameter combinations
labels <- paste0("N_", combo_grid$N, "_bwfactor_", combo_grid$bwfactor)
colnames(atet_matrix) <- labels
colnames(se_matrix) <- labels

# Prepare long format results
repeated_combo <- combo_grid[rep(1:K, each = S), ]
seed_vec <- rep(seed + seq_len(S), times = K)

long_results <- data.frame(
  treatment = treatment,
  control = control,
  true_effect = treatment^2 - control^2,
  N = repeated_combo$N,
  bwfactor = repeated_combo$bwfactor,
  atet = as.vector(atet_matrix),
  se = as.vector(se_matrix),
  seed = seed_vec
)

# Calculate confidence intervals
long_results$ci_lower <- long_results$atet - qnorm(1 - 0.05 / 2) *
    long_results$se
long_results$ci_upper <- long_results$atet + qnorm(1 - 0.05 / 2) *
    long_results$se

# Add coverage indicator
long_results$coverage <- as.integer(
  long_results$true_effect >= long_results$ci_lower &
      long_results$true_effect <= long_results$ci_upper
)

# Write results to a single CSV file
long_results <- long_results[, c(
  "treatment", "control", "true_effect", "N", "bwfactor",
  "atet", "se", "ci_lower", "ci_upper", "coverage", "seed"
)]
write.csv(long_results, file = "simulation_results.csv", row.names = FALSE,
     na = "")

end_time <- Sys.time()
```

```r
  elapsed <- difftime(end_time, start_time, units = "mins")
  cat(sprintf("[END] Simulation completed at %s (Elapsed: %.2f min)\n",
      end_time, as.numeric(elapsed)))
}

# Run the simulation with specified parameters
run_simulation(
  dgp_func = generate_data,
  S = 5000,
  treatment = 3,
  control = 2,
  N_list = c(2000, 4000, 8000),
  bwfactor_list = c(0.25, 0.5, 0.75, 1.0, 1.25, 1.5, 1.75, 2.0),
  seed = 0
)
```

Listing B.1: R code for simulation

B.2.2 SQL Query for Summary Statistics

```sql
WITH non_null_simulation_results AS (
  SELECT
    * EXCEPT(group_row_number)
  FROM (
    SELECT
      *,
      ROW_NUMBER() OVER (
        PARTITION BY
          CAST(treatment AS STRING),
          CAST(control AS STRING),
          CAST(true_effect AS STRING),
          CAST(N AS STRING),
          CAST(bwfactor AS STRING)
        ORDER BY
          seed
      ) AS group_row_number
    FROM
      `continuous-did-dml.results.simulation_results`
    WHERE
      atet IS NOT NULL
      AND
      se IS NOT NULL
      AND
      coverage IS NOT NULL
  )
  WHERE
    group_row_number <= 4000
)
```

```
SELECT
  treatment,
  control,
  true_effect,
  N,
  bwfactor,
  AVG(atet) - true_effect AS bias,
  STDDEV(atet) AS se,
  SQRT(AVG(POW(atet - true_effect, 2))) AS rmse,
  AVG(se) AS avse,
  AVG(coverage) AS coverage_rate
FROM
  non_null_simulation_results
GROUP BY
  treatment, control, true_effect, N, bwfactor
ORDER BY
  treatment, control, N, bwfactor
;
```

Listing B.2: SQL query for summary statistics

Bibliography

Abadie, A. (2005). Semiparametric difference-in-differences estimators. *The Review of Economic Studies*, 72(1):1–19.

Acemoglu, D. and Finkelstein, A. (2008). Input and technology choices in regulated industries: Evidence from the health care sector. *Journal of Political Economy*, 116(5):837–880.

Ananat, E., Glasner, B., Hamilton, C., Parolin, Z., and Pignatti, C. (2024). Effects of the expanded child tax credit on employment outcomes. *Journal of Public Economics*, 238:105168.

Athey, S. and Imbens, G. W. (2022). Design-based analysis in difference-in-differences settings with staggered adoption. *Journal of Econometrics*, 226(1):62–79.

Bodory, H., Huber, M., and Kueck, J. (2025). causalweight: Estimation Methods for Causal Inference Based on Inverse Probability Weighting and Doubly Robust Estimation.

Borusyak, K., Jaravel, X., and Spiess, J. (2024). Revisiting event-study designs: Robust and efficient estimation. *The Review of Economic Studies*, 91(6):3253–3285.

Callaway, B., Goodman-Bacon, A., and Sant'Anna, P. H. C. (2025). Difference-in-differences with a continuous treatment.

Callaway, B. and Sant'Anna, P. H. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225(2):200–230.

Cameron, A. C. and Trivedi, P. K. (2005). *Microeconometrics : methods and applications*. Cambridge University Press.

Chang, N.-C. (2020). Double/debiased machine learning for difference-in-differences models. *The Econometrics Journal*, 23(2):177–191.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Cid, J. A. and von Davier, A. A. (2015). Examining potential boundary bias effects in kernel smoothing on equating: An introduction for the adaptive and epanechnikov kernels. *Applied Psychological Measurement*, 39(3):208–222.

de Chaisemartin, C. and D'Haultfœuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, 110(9):2964–96.

de Chaisemartin, C., D'Haultfœuille, X., Pasquier, F., Sow, D., and Vazquez-Bare, G. (2025). Difference-in-differences estimators for treatments continuously distributed at every period.

D'Haultfœuille, X., Hoderlein, S., and Sasaki, Y. (2023). Nonparametric difference-in-differences in repeated cross-sections with continuous treatments. *Journal of Econometrics*, 234(2):664–690.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225(2):254–277.

Haddad, M. F. C., Huber, M., and Zhang, L. Z. (2024). Difference-in-differences with time-varying continuous treatments using double/debiased machine learning.

Huber, M., Hsu, Y.-C., Lee, Y.-Y., and Lettry, L. (2020). Direct and indirect effects of continuous treatments based on generalized propensity score weighting. *Journal of Applied Econometrics*, 35(7):814–840.

Müller, H. G. and Stadtmüller, U. (2002). Multivariate boundary kernels and a continuous least squares principle. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 61(2):439–458.

Roth, J., Sant'Anna, P. H., Bilinski, A., and Poe, J. (2023). What's trending in difference-in-differences? a synthesis of the recent econometrics literature. *Journal of Econometrics*, 235(2):2218–2244.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational psychology*, 66(5):688–701.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Chapman and Hall.

Sun, L. and Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 225(2):175–199.

Terrell, G. R. and Scott, D. W. (1992). Variable Kernel Density Estimation. *The Annals of Statistics*, 20(3):1236 – 1265.

Zhang, L. (2025). Continuous difference-in-differences with double/debiased machine learning.