

Inferring Treatment Effects in Large Panels by Uncovering Latent Similarities

Ben Deaner*

Chen-Wei Hsiang*

Andrei Zeleneev*

This version: January 31, 2025

Preliminary and incomplete; please do not cite and circulate

Abstract

The presence of unobserved confounders is one of the main challenges in identifying treatment effects. In this paper, we propose a new approach to causal inference for panel data based on uncovering latent similarities among units from long pre-treatment histories of outcomes. To this end, we assume that the untreated outcomes follow a nonparametric nonlinear factor model satisfying minimal smoothness requirements. We impute both missing counterfactual outcomes and propensity scores using kernel smoothing based on the constructed measure of similarity between units, and demonstrate that our estimator (nearly) achieves the optimal nonparametric rate of convergence. Finally, we construct a doubly robust estimator of the average treatment effect on the treated (ATT), and provide conditions, under which this estimator is $\sqrt{N}$ -consistent, and asymptotically normal and unbiased. Our simulation study demonstrates that our method provides accurate inference for a wide range of data generating processes, whereas conventional panel methods fail to do so.

*Department of Economics, University College London

1 Introduction

The potential presence of latent confounding variables is one of the main challenges for causal inference. With panel data opening up possibilities for controlling for time-invariant unobservables (fixed effects) and imputation of counterfactual outcomes, methods for causal inference in panels have become dominant in empirical work, and their development is an active and rapidly growing area of research (see, e.g., [Arkhangelsky and Imbens, 2024](#)).

To control for selection based on the unobserved fixed effects and/or to impute potential outcomes, existing methods for causal inference in panels impose different assumptions on the untreated outcomes. For example, difference-in-differences (DiD) and two-way fixed effects (TWFE) methods rely on various parallel trend assumptions, whereas synthetic control (SC) and (most of) matrix completion methods require the untreated potential outcomes to follow a linear factor model. While such assumptions are routinely imposed for simplicity of estimation and inference or motivated by limited data availability, they are known to be and often criticized as restrictive.

The goal of this paper is to develop a new method for estimation of and inference on causal effects of a binary treatment $w_{i,t}$ in large panels, which remains valid under weaker assumptions imposed on the untreated potential outcomes. Specifically, we assume that the untreated potential outcome of unit i in period t follows a nonlinear factor model

$$Y_{i,t}(0) = \mu^{(0)}(\alpha_i, \lambda_t) + u_{i,t}, \quad (1)$$

where α_i and λ_t are unobserved unit and time effects, $\mu^{(0)}(\cdot, \cdot)$ is an *unknown* function, and $u_{i,t}$ is an (exogenous) residual.¹ Importantly, we do not make any assumptions about the functional form and shape of $\mu^{(0)}(\cdot, \cdot)$ (other than imposing minimal smoothness requirements). Thus, the considered untreated potential outcomes model (1) is substantially more general than the commonly used formulations $\mu^{(0)}(\alpha_i, \lambda_t) = \alpha_i + \lambda_t$ and $\mu^{(0)}(\alpha_i, \lambda_t) = \alpha_i' \lambda_t$ imposed by DiD/TWFE and SC methods, respectively. Finally, α_i is also allowed to affect both the treated potential outcomes $Y_{i,t}(1)$ and selection into treatment, i.e., it plays the role of the latent confounder in our framework.

While the generality of formulation (1) is appealing, it obscures identification and estimation of treatment effects because α_i cannot be identified unless additional restrictions are imposed on the form of $\mu^{(0)}(\cdot, \cdot)$. Instead of trying to model $\mu^{(0)}(\cdot, \cdot)$ and estimate α_i directly, we

¹It is also straightforward to allow for a linear covariates adjustment in (1), but we deliberately abstract from this extension for clarity of exposition of the main ideas.

follow the approach of [Zhang et al. \(2017\)](#) and use a long history of pre-treatments outcomes to construct a pseudo-distance $\hat{d}_{i,j}$ informative about how close units i and j are in terms of their latent characteristics α_i and α_j . This allows us to estimate the *values* of the conditional mean $\mu_{i,t}^{(0)} := \mathbb{E}[Y_{i,t}(0)|\alpha_i]$ and of the propensity scores $p_{i,t} := \mathbb{E}[w_{i,t}|\alpha_i]$ for a given post-treatment period t for all units by kernel smoothing based on $\hat{d}_{i,j}$, even though α_i is not observed neither identified.

We construct estimators of $\mu_{i,t}^{(0)}$ and $p_{i,t}$ and establish their rates of convergence (uniformly over i) when both the number of units N and the number of pre-treatment periods T_0 go to infinity. Notably, if T_0 goes to infinity sufficiently fast, our estimators achieve the optimal nonparametric rate of convergence (up to a log term) as if we observed the latent α_i ([Stone, 1980](#)).

Once the values of the conditional means and the propensity scores are imputed, we construct a doubly robust estimate of the average treatment effect on the treated (ATT) for a given post-treatment period t (e.g., [Robins et al., 1994](#); [Chernozhukov et al., 2018a](#)). We propose a simple cross-fitting scheme, which, together with double robustness, ensures that our treatment effect estimator is $\sqrt{N}$ -consistent, and asymptotically normal and unbiased, even though $\mu_{i,t}^{(0)}$ and $p_{i,t}$ are estimated at slower nonparametric rates. Remarkably, under certain regularity conditions and if T_0 grows sufficiently fast, our estimator achieves the semiparametric efficiency bound as if α_i was observed.

Related Literature (*preliminary and incomplete*)

This paper contributes to the vast and rapidly growing literature on causal inference in panels. We refer the reader to [Arkhangelsky and Imbens \(2024\)](#) for an excellent recent overview of this field.

DiD and TWFE have been dominant and still growing in popularity causal inference methods in applied work due to their practicality and perceived transparency (e.g., [Goldsmith-Pinkham, 2024](#)); see [De Chaisemartin and d'Haultfoeuille \(2023\)](#) and [Roth et al. \(2023\)](#) for extensive reviews of the DiD and TWFE literature. While allowing for heterogeneous treatment effects has been one of the main focuses of this literature (e.g., [De Chaisemartin and d'Haultfoeuille, 2020](#); [Callaway and Sant'Anna, 2021](#); [Sun and Abraham, 2021](#); [Wooldridge, 2021](#); [Borusyak et al., 2024](#)), the validity of DiD and TWFE methods crucially relies on the parallel trends assumption. A number of recent papers re-evaluated the restrictiveness of this assumption and its implications for selection into treatment; the proposed approaches to relaxing it within the DiD/TWFE framework typically involve putting a priori restrictions

on the possible violations of the parallel trends assumption, and result in partially identified treatment effects (e.g., [Manski and Pepper, 2018](#); [Rambachan and Roth, 2023](#); [Ghanem et al., 2022](#)). In contrast to this strand of the literature, our framework substantially relaxes the parallel trends assumption while still allowing for rich heterogeneity of treatment effects and selection mechanisms, and maintaining point identification, at the cost of relying on a sufficiently long pre-treatment history.

Alternative popular approaches to causal inference in panels include SC and matrix completion methods; see [Abadie \(2021\)](#) for a recent review of SC methods.² Similarly to the interactive fixed effects panel literature (e.g., [Pesaran, 2006](#); [Bai, 2009](#)), SC and matrix completion methods generalize the TWFE framework by allowing the (untreated) outcomes to follow a factor model. Similarly to this paper, these methods impute the missing counterfactual outcomes leveraging the factor structure of a (long) history of pre-treatment outcomes. However, to establish statistical guarantees and validity of inference, most methods assume that $\mu^{(0)}(\alpha_i, \lambda_t) = \alpha_i' \lambda_t$ follows the exact linear factor model and that the factors are strong, whereas we consider the general nonparametric nonlinear factor model.³ While some matrix completion methods consider or allude to possible extensions to nonlinear factor models, these papers either do not provide inferential theory and/or impose strong smoothness requirements on $\mu^{(0)}(\cdot, \cdot)$, e.g., assume that $\mu^{(0)}(\cdot, \cdot)$ belongs to a Hölder class (e.g., [Agarwal et al., 2020](#); [Fernández-Val et al., 2021](#)). In contrast, our imputation method (nearly) achieves the optimal nonparametric rate of convergence and allows us to provide a semiparametrically efficient estimator of the ATTs under a substantially weaker Lipschitz continuity assumption.

The rest of the paper is organized as follows. Section 2 introduces the framework and describes our estimator. Section 3 provides formal statistical guarantees. Section 4 presents numerical evidence. All auxiliary lemmas and proofs are provided in Appendix A.

²Some recent developments in the rapidly growing SC and causal matrix completion literatures literature include, among other, [Arkhangelsky et al. \(2021\)](#); [Cattaneo et al. \(2021\)](#); [Chernozhukov et al. \(2021, 2018b\)](#) and [Athey et al. \(2021\)](#); [Bai and Ng \(2021\)](#); [Agarwal et al. \(2023\)](#); [Abadie et al. \(2024\)](#), respectively.

³The potential presence of weak factors is a known challenge to estimation and inference even when the treatment effect is constant. [Armstrong et al. \(2022\)](#) provide a more detailed discussion of this problem, and propose a new debiased estimator and construct bias-aware confidence intervals robust to the presence of weak factors. Allowing for weak factors in settings with heterogeneous treatment effects, to the best of our knowledge, remain an open problem.

2 Notation and Proposed Method

2.1 Model and Notation

Our sample consists of individuals indexed by $i = 1, \dots, n$ and time periods indexed by $t = 1, \dots, T$. For each individual i and period t , we observe an outcome $Y_{i,t}$ and a binary indicator $w_{i,t}$. The indicator $w_{i,t}$ is equal to one if individual i is treated at or prior to time t and zero otherwise. We let $Y_{i,s:t}$ denote the vector of outcomes for individual i from periods s to t inclusive and similarly for other variables.

Let $Y_{i,t}(0)$ denote the potential outcome of individual i at time t under a counterfactual in which the individual has not yet received treatment at period t . We assume that if $w_{i,t} = 0$ then $Y_{i,t} = Y_{i,t}(0)$. This precludes the possibility that individuals anticipate future treatment and that this impacts their outcomes (e.g., [Abbring and Van Den Berg, 2003](#); [Borusyak et al., 2022](#); [Sun and Abraham, 2021](#)). We state this formally in Assumption 1 below.

Assumption 1 (No Anticipation). If $w_{i,t} = 0$ then $Y_{i,t} = Y_{i,t}(0)$.

Central to our analysis is the assumption of a non-linear factor model for the untreated potential outcomes. Let α_i be some latent and time-invariant characteristics of individual i , and let λ_t be period-specific factors. α_i may capture say, unobserved demographic characteristics, individual preferences, or innate ability. λ_t may capture e.g., unobserved macro-economic conditions, government policy, or environmental factors. We model the potential outcomes and realized treatments as follows.

$$Y_{i,t}(0) = \mu^{(0)}(\alpha_i, \lambda_t) + u_{i,t}, \quad \mathbb{E}[u_{i,t} | \alpha_i, \lambda_t] = 0. \quad (2)$$

$$w_{i,t} = p_t(\alpha_i) + \epsilon_{i,t}, \quad \mathbb{E}[\epsilon_{i,t} | \alpha_i, \lambda_t] = 0. \quad (3)$$

The residuals $u_{i,t}$ and $\epsilon_{i,t}$ are unobserved. The functions $\mu^{(0)}$ and p_t are unknown and we do not assume that they have any particular functional form. Note that the model above implies that the mean of $Y_{i,t}(0)$ given the latent factors is time-invariant and does not depend on the individual i . However, the time-subscript on p_t allows the conditional mean of $w_{i,t}$ to vary over time, which reflects the fact that this variable is increasing and that it is therefore non-stationary.

Knowledge of the latent factors α_i and λ_t is insufficient to identify causal quantities of interest. This is because the residual in the treatment model (3) may be correlated with the residual in the outcome equation (2). In particular, there may be confounding factors that are not included in α_i and λ_t and which influence both the outcome and treatment status.

In order to achieve identification we make the key assumption that the latent factors α_i and λ_t together account for all confounding between the outcome and treatment. Formally, we assume that after controlling for these latent factors, there is no residual dependence between the untreated potential outcome and treatment.

Assumption 2 (Latent Unconfoundedness). $Y_{i,t}(0) \perp\!\!\!\perp w_{i,t} | \alpha_i, \lambda_t$.

The assumption of latent unconfoundedness is common in the literature (e.g. [Abadie et al., 2024](#); [Agarwal et al., 2021](#); [Arkhangelsky and Imbens, 2022](#); [Athey et al., 2021](#); [Fernández-Val et al., 2021](#)). Note that the assumption effectively requires that the latent factors are sufficiently rich.

Exposure to macroeconomic conditions and other shared aggregate factors may induce dependence between the potential outcomes of different individuals. We assume that the time-specific factors λ_t are sufficiently rich that after controlling for these factors and the individual-specific factors, the untreated potential outcomes of any two individuals are independent.

Assumption 3 (Latent Independence). For any individuals $i \neq j$, $Y_{i,t}(0) \perp\!\!\!\perp Y_{j,t}(0) | \{\alpha_k\}_{k=1}^N, \lambda_t$.

Under Assumption 2, key counterfactual quantities of interest can be written in terms of the latent factors and functions $\mu^{(0)}$ and p . For notational convenience, let $p_{i,t} := p_t(\alpha_i)$ and $\mu_{i,t}^{(0)} := \mu^{(0)}(\alpha_i, \lambda_t)$. Under Assumption 2, the average effect of treatment on the treated at time t , which is defined as $\mathbb{E}_t[Y_{i,t} - Y_{i,t}(0) | w_{i,t} = 1]$, can be written in doubly-robust form as follows.

$$ATT_t = \frac{1}{P(w_{i,t} = 1)} \mathbb{E}_t \left[Y_{i,t} w_{i,t} - \frac{(1 - w_{i,t}) Y_{i,t} p_{i,t} + (w_{i,t} - p_{i,t}) \mu_{i,t}^{(0)}}{1 - p_{i,t}} \right]$$

The time subscript on the expectation above indicates that it is taken with respect to the period t -specific distribution of the observables (i.e., conditional on λ_t). The doubly-robust form of ATT above is proposed and used in panel data settings by [Sant'Anna and Zhao \(2020\)](#). Additionally, the proposed method can also adopt other causal estimands with their respective doubly robust score functions (e.g., [Arkhangelsky and Imbens, 2022](#)).

In this paper we provide new methods for estimating $p_{i,t}$ and $\mu_{i,t}^{(0)}$. Given these estimates, one can construct a doubly-robust estimate of ATT_t and perform inference on this object. Let $\hat{p}_{i,t}$ and $\hat{\mu}_{i,t}^{(0)}$ be estimates of $p_{i,t}$ and $\mu_{i,t}^{(0)}$ respectively. Then a corresponding doubly-robust estimate of ATT_t is given below, where N is the sample size, and $N_{1,t}$ the number of individuals treated by time t .

$$\hat{ATT}_t = \frac{1}{N_{1,t}} \sum_{i=1}^N \left(Y_{i,t} w_{i,t} - \frac{(1 - w_{i,t}) Y_{i,t} \hat{p}_{i,t} + (w_{i,t} - \hat{p}_{i,t}) \hat{\mu}_{i,t}^{(0)}}{1 - \hat{p}_{i,t}} \right) \quad (4)$$

2.2 Proposed Method

Our proposed estimation method uses pre-treatment outcomes to find untreated individuals whose latent factors α_i are similar to those of treated individuals. To motivate our approach, let us first suppose α_i were observed. Under Assumptions 1 and 2, we have

$$\mu^{(0)}(\alpha, \lambda_t) = \mathbb{E}_t[Y_{i,t} | \alpha_{i,t} = \alpha, w_{i,t} = 0], \quad p_t(\alpha) = \mathbb{E}_t[w_{i,t} | \alpha_{i,t} = \alpha].$$

The objects on the right-hand sides of each equation above are regression functions. If $\alpha_{i,t}$ were observable, we could apply Nadaraya-Watson to non-parametrically estimate the functions $\mu^{(0)}(\cdot, \lambda_t)$ and $p_t(\cdot)$ and thus $p_{i,t}$ and $\mu_{i,t}^{(0)}$. To be precise, we could obtain the following estimates.

$$\hat{\mu}_{i,t}^{(0)} = \frac{\sum_{j: w_{j,t}=0} K(\|\alpha_i - \alpha_j\|/h) Y_{j,t}}{\sum_{j: w_{j,t}=0} K(\|\alpha_i - \alpha_j\|/h)}, \quad \hat{p}_{i,t} = \frac{\sum_j K(\|\alpha_i - \alpha_j\|/h) w_{j,t}}{\sum_j K(\|\alpha_i - \alpha_j\|/h)}$$

The estimates above are infeasible because in practice, we do not directly observe α_i for any individual i . In order to obtain feasible estimates, we replace the infeasible distance $\|\alpha_i - \alpha_j\|$ in the expressions above, with a feasible pseudo-distance. To define this pseudo-distance, suppose that no individuals in the population are treated prior to some period $T_0 + 1$ and let $\langle \cdot, \cdot \rangle$ be the Euclidean inner-product. We use the pseudo-distance defined below:

$$\hat{d}_{i,j} = \frac{1}{T_0} \max_{k \notin \{i,j\}} |\langle Y_{k,1:T_0}, Y_{i,1:T_0} - Y_{j,1:T_0} \rangle|$$

A pseudo-distance of the form above is employed in [Zhang et al. \(2017\)](#). To motivate the use of the pseudo-distance, suppose Assumption 3 holds. Then conditional on the individual latent factors, the inner-product in the pseudo-distance is an unbiased estimate of the object on the right-hand side below.

$$\mathbb{E}[\langle Y_{k,1:T_0}, Y_{i,1:T_0} - Y_{j,1:T_0} \rangle | \alpha_i, \alpha_j, \alpha_k] = \int \mu^{(0)}(\alpha, \lambda) (\mu^{(0)}(\alpha_i, \lambda) - \mu^{(0)}(\alpha_j, \lambda)) d\pi(\lambda).$$

In the above, π is the stationary distribution of λ_t . As such, we can understand the pseudo-distance as the sample analogue of the population pseudo-distance below, where $\mathcal{A}$ is the support of α_i .

$$d_{i,j} = \sup_{\alpha \in \mathcal{A}} \left| \int \mu^{(0)}(\alpha, \lambda) (\mu^{(0)}(\alpha_i, \lambda) - \mu^{(0)}(\alpha_j, \lambda)) d\pi(\lambda) \right| \quad (5)$$

$$\geq \frac{1}{2} \int (\mu^{(0)}(\alpha_i, \lambda) - \mu^{(0)}(\alpha_j, \lambda))^2 d\pi(\lambda) \quad (6)$$

The inequality relates the size of the pseudo-distance $d_{i,j}$ to a squared L_2 distance between the functions $\mu^{(0)}(\alpha_i, \cdot)$ and $\mu^{(0)}(\alpha_j, \cdot)$. Thus $d_{i,j}$ measures similarity of the latent factors to the

extent that they impact outcomes. We provide sufficient conditions for the consistency of the sample pseudo-metric to this quantity.

It is worth contrasting the sample pseudo-distance above with the Euclidean distance between the history of pre-treatment outcomes, which is defined as $\|Y_{i,1:T_0} - Y_{j,1:T_0}\|^2$. The mean of the squared Euclidean distance conditional on the individual latent factors is given below, where we again assume that untreated potential outcomes of different individuals are independent conditional on the latent factors.

$$\begin{aligned}\mathbb{E}[\|Y_{i,1:T_0} - Y_{j,1:T_0}\|^2 | \alpha_i, \alpha_j] &= \int (\mu^{(0)}(\alpha_i, \lambda) - \mu^{(0)}(\alpha_j, \lambda))^2 d\pi(\lambda) \\ &+ \int E[u_{i,t}^2 | \alpha_i, \lambda_t = \lambda] + E[u_{j,t}^2 | \alpha_j, \lambda_t = \lambda] d\pi(\lambda)\end{aligned}$$

Thus the Euclidean distance is increasing in the conditional residual variance $E[u_{j,t}^2 | \alpha_j, \lambda_t = \lambda]$. Thus in the presence of conditional heteroskedasticity, a small Euclidean distance may reflect that individual j 's outcomes have a low residual variance and not that individual j and i have similar latent factors. It is the need to be robust to conditional heteroskedasticity that motivates our use of the pseudo-metric.

Given the pseudo-distance, we may form feasible estimates of $p_{i,t}$ and $\mu_{i,t}^{(0)}$ as follows.

$$\hat{\mu}_{i,t}^{(0)} = \frac{\sum_{j:w_{j,t}=0} K(\hat{d}_{i,j}/h) Y_{j,t}}{\sum_{j:w_{j,t}=0} K(\hat{d}_{i,j}/h)}, \quad \hat{p}_{i,t} = \frac{\sum_j K(\hat{d}_{i,j}/h) w_{j,t}}{\sum_j K(\hat{d}_{i,j}/h)}$$

One could plug the estimates above into the the formula for the doubly-robust ATT_t estimate (4). However, we instead employ a cross-fitting scheme to further de-bias our estimates. This is in-line with the extensive literature on double machine learning, which demonstrates the utility of cross-fitting for reducing bias and obtaining valid inference (e.g., [Chernozhukov et al., 2018a](#); [Abadie et al., 2024](#)). The full algorithm with cross-fitting is detailed below along with a variance estimate and confidence interval.

Algorithm 1 Doubly-Robust Estimation and Inference with Cross-Fitting

Inputs: Number of folds J , bandwidth h , confidence level α . Returns: ATT_t estimate $\hat{ATT}_t$, variance estimate $\hat{V}_t$, and level $1-\alpha$ confidence interval.

- 1: Randomly partition $[N] = \{1, 2, \dots, N\}$ into J folds $\{\mathcal{I}_j\}_{j=1}^J$ of size $\approx N/J$. Let $\mathcal{I}_{-j} = [N] \setminus \mathcal{I}_j$.
- 2: **for all** $j \in [J]$ and $i \in \mathcal{I}_j$ **do**
- 3: Calculate $\hat{\mu}_{i,t}^{(0)}$ and $\hat{p}_{i,t}$ as follows.

$$\hat{\mu}_{i,t}^{(0)} \leftarrow \frac{\sum_{j \in \mathcal{I}_{-j}; w_{j,t}=0} K(\hat{d}_{i,j}/h) Y_{j,t}}{\sum_{j \in \mathcal{I}_{-j}; w_{j,t}=0} K(\hat{d}_{i,j}/h)}, \quad \hat{p}_{i,t} \leftarrow \frac{\sum_{j \in \mathcal{I}_{-j}} K(\hat{d}_{i,j}/h) w_{j,t}}{\sum_{j \in \mathcal{I}_{-j}} K(\hat{d}_{i,j}/h)}$$

4: **end for**

- 5: Construct $\hat{ATT}_t$ using the formula below.

$$\begin{aligned} \hat{ATT}_t &\leftarrow \frac{1}{N_{1,t}} \sum_{i=1}^N \left(Y_{i,t} w_{i,t} - \frac{(1 - w_{i,t}) Y_{i,t} \hat{p}_{i,t} + (w_{i,t} - \hat{p}_{i,t}) \hat{\mu}_{i,t}^{(0)}}{1 - \hat{p}_{i,t}} \right) \\ \hat{V}_t &\leftarrow \frac{N}{N_{1,t}^2} \sum_{i=1}^N \left(Y_{i,t} w_{i,t} - \frac{(1 - w_{i,t}) Y_{i,t} \hat{p}_{i,t} + (w_{i,t} - \hat{p}_{i,t}) \hat{\mu}_{i,t}^{(0)}}{1 - \hat{p}_{i,t}} - \frac{N_{1,t}}{N} \hat{ATT}_t \right)^2 \end{aligned}$$

- 6: Form $1 - \alpha$ level confidence interval $[\hat{ATT}_t \pm Z_{(1-\alpha)/2} \sqrt{\hat{V}_t/N}]$.
-

3 Large Sample Theory

The estimator $\hat{ATT}_t$ is doubly robust and employs cross-fitting. An extensive literature (e.g., [Robins et al. \(2008\)](#); [Chernozhukov et al. \(2018a\)](#); [Abadie et al. \(2024\)](#)) provides sufficient conditions for root- N consistency of such estimates and asymptotically correct coverage of the corresponding confidence intervals. A key condition is that the first stage nuisance-parameter estimates (in our case $\hat{\mu}_{i,t}^{(0)}$ and $\hat{p}_{i,t}$) converge sufficiently quickly. For this condition, the following rates suffice:

$$\max_{i \in [N]} |\hat{\mu}_{i,t}^{(0)} - \mu_{i,t}^{(0)}| = o_p(N^{-1/4}) \quad \text{and} \quad \max_{i \in [N]} |\hat{p}_{i,t} - p_{i,t}| = o_p(N^{-1/4})$$

We establish convergence rates for the first-stage estimates, which may be of interest per se. We show that under certain conditions, the estimates achieve the [Stone, 1980](#) optimal rate for non-parametric regression on α_i under Lipschitz continuity. That is, we can achieve the same optimal rate attainable for the infeasible estimates that take α_i as known. However, in order to achieve this rate, we require that the number of pre-treatment periods T_0

$$\|\hat{\mu}^{(0)} - \mu^{(0)}\| = O_p(R_\mu) \quad \text{and} \quad \|\hat{p} - p\| = O_p(R_p)$$

Stone, 1980. The centric difficult inference procedure of nonparametric estimators is due to failing to achieve the root- N rate of convergence, unlike other common parametric estimators.

The augmented inverse propensity weighted (AIPW) estimator offers a valid approach for inference-making with nonparametric estimators, primarily due to its rate double robustness property related to convergence rates. Although the convergence rates for both the outcome model and the propensity score model are slower than the root- N rate, the inference process remains accessible under certain regularity conditions if the product of their rates is sufficiently fast and attains the root- N rate. That is, suppose the rates of convergence for the outcome and the propensity score model are denoted as

$$\|\hat{\mu}^{(0)} - \mu^{(0)}\| = O_p(R_\mu) \quad \text{and} \quad \|\hat{p} - p\| = O_p(R_p)$$

Although both R_μ and R_p are slower than root- N when $\mu^{(w)}$ and p are being nonparametrically estimated, the doubly robust estimator for the target causal estimand can still achieve asymptotic normality

$$\sqrt{N}(\text{DR Estimator} - \text{Causal Estimand}) \rightarrow N(0, \text{AsyVar})$$

whenever $R_\mu = o(1)$, $R_p = o(1)$, and $R_\mu R_p = o(N^{-1/2})$. This concept, often referred to as rate double robustness in the literature, is discussed in works such as Smucler et al. (2019) and utilized in studies like Robins et al. (2008); Chernozhukov et al. (2018a); Abadie et al. (2024).

In Section 3.1, we will formally illustrate the assumptions and the rate of convergence for the proposed method, which has been shown to achieve double robustness. Then, in Section 3.2, we will utilize this rate of convergence to establish the asymptotic inference result, along with the corresponding doubly robust estimator.

3.1 Rate of Convergence for Outcome and Propensity Score

In order to derive convergence rates for our first-stage estimates of $\mu^{(0)}$ and p_t we impose the following additional assumptions.

Assumption 4 (Model and Latent Factors).

- (i) There is some $\underline{c} > 0$ so that $\mathbb{P}(w_{i,t} = 1 | \alpha_i) \geq \underline{c}$ almost surely for all i and $t > T_0$.

- (ii) $\mathbb{E}[u_{i,t}|\alpha, \lambda] = 0$, and for some $\delta > 0$, $\mathbb{E}[\exp(\eta u_{i,t})|\alpha, \lambda] \leq \exp(\delta \eta^2)$ for all $\eta \in \mathbb{R}$ almost surely and likewise for $\epsilon_{i,t}$. These error terms are jointly independent across time and across individuals conditional on the latent factors.
- (iii) $\text{supp}(\alpha) \subseteq \mathcal{A}$ where $\mathcal{A}$ is a compact subset of $\mathbb{R}^{d_\alpha}$. α_i and λ_t are jointly i.i.d. across i and t , respectively. There are constants $0 < \underline{c} < \bar{c} < \infty$ so that for any fixed $\alpha \in \text{supp}(\alpha)$ and any η , $\underline{c}\eta^{d_\alpha} \leq \mathbb{P}(\|\alpha - \alpha_i\| \leq \eta) \leq \bar{c}\eta^{d_\alpha}$.
- (iv) The functions $\mu^{(0)}, \mu^{(1)}, p : \mathcal{A} \times \mathcal{L} \rightarrow \mathbb{R}$ are uniformly bounded and satisfy, for some $L_0, L_p < \infty$,

$$|\mu^{(0)}(\alpha_1, \lambda) - \mu^{(0)}(\alpha_2, \lambda)| \leq L_0 \|\alpha_1 - \alpha_2\|$$

$$|p_t(\alpha_1) - p_t(\alpha_2)| \leq L_p \|\alpha_i - \alpha_j\|$$

for all $\alpha_i, \alpha_j \in \mathcal{A}$, $\lambda \in \mathcal{L}$, and $t \in [T]$.

- (v) $K : \mathbb{R}_+ \rightarrow \mathbb{R}$ is weakly positive and strictly positive at zero, supported on compact set and bounded by $\bar{K} < \infty$. K satisfies $|K(z) - K(z')| \leq \bar{K}'|z - z'|$ for all $z, z' \in \mathbb{R}_+$ for some $\bar{K}' > 0$.
- (vi) There exist constants $c_1, c_2 > 0$ so that for all $\alpha_1, \alpha_2 \in \mathcal{A}$

$$\int_{\lambda \in \text{supp}(\lambda)} (\mu^{(0)}(\alpha_1, \lambda) - \mu^{(0)}(\alpha_2, \lambda))^2 d\pi(\lambda) \geq c_1 \|\alpha_1 - \alpha_2\|^2 \quad (7)$$

and

$$\begin{aligned} & \sup_{\alpha \in \text{supp}(\alpha)} \int_{\lambda \in \text{supp}(\lambda)} \mu^{(0)}(\alpha, \lambda) (\mu^{(0)}(\alpha_1, \lambda) - \mu^{(0)}(\alpha_2, \lambda)) d\pi(\lambda) \\ & \geq c_2 \sqrt{\int_{\lambda \in \text{supp}(\lambda)} (\mu^{(0)}(\alpha_1, \lambda) - \mu^{(0)}(\alpha_2, \lambda))^2 d\pi(\lambda)} \end{aligned} \quad (8)$$

Assumption 4 (i) is a standard overlap condition and would be required for regular estimation of the ATT even if α_i were observed. Note that because we are interested in the average effect of treatment on the treated, we only require that the conditional probability of treatment is bounded below away from zero and not above away from one. Assumption 4 (ii) restricts that the tail behavior of the error terms, requiring them to be sub-Gaussian. The assumption also imposes that the residuals are independent over time given the time-specific factors. This condition allows us to apply particular concentration inequalities in order to obtain fast rates of convergence.

Assumption 4 (iii) imposes that the latent factors are independent and identically distributed and that the individual-specific factors have compact support. Compact support of the individual-specific latent factors helps to ensure that with high probability, for each treated individual i , there exist untreated individuals in the sample whose latent factors are close to those of i . Independence and identical distribution of the individual-specific latent factors follows if we understand individuals to be drawn identically and independently from the underlying population. Independence of λ_i over time may be plausible if time periods are sufficiently far apart. This restriction on λ_i allows us to apply concentration inequalities and ensure fast convergence of the pseudo-distance.

Assumption 4 (iv) imposes that the functions $\mu^{(0)}$ and p_t are bounded and vary smoothly with the individual-specific factors. This smoothness assumption ensures that individuals with similar latent factors also have similar conditional-mean potential outcomes and treatments. Assumption 4 (v) stipulates properties of the kernel K . Kernels that satisfy this condition include are common in the literature, with the Epanechnikov kernel a particularly prevalent choices.

Assumption 4 (vi) relates the pseudo-distance to the distance between individual latent factors. The assumption imposes first, that the mean-squared distance between $\mu^{(0)}(\alpha_1, \cdot)$ and $\mu^{(0)}(\alpha_2, \cdot)$ is at least proportional to the Euclidean distance between α_1 and α_2 . In addition, the assumption states that the population pseudo-distance is at least proportional to this mean-squared distance. Consider the special case in which $\mu^{(0)}(\alpha, \lambda) = \alpha' M \lambda$ for some fixed matrix M . In this case, the condition (7) holds so long as the matrix $M \int_{\lambda \in \text{supp}(\lambda)} \lambda \lambda' d\pi(\lambda) M'$ is strictly positive definite. If this is the case, then the second condition (8) holds if there is a $c > 0$ so that for any $\alpha_1, \alpha_2 \in \mathcal{A}$ we have $\frac{c}{\|\alpha_1 - \alpha_2\|}(\alpha_1 - \alpha_2) \in \mathcal{A}$. This is true, for example, if $\mathcal{A}$ is a Euclidean ball centered at zero.

Remark 1. Bi-Lipschitz-type assumptions that play a similar role to Assumption 4 (vi) are common in the literature. Note that Assumptions 4 (vi) allows for multi-dimensional latent factors α . For example, Zhang et al. (2017) and Issartel (2020) assume the standard bi-Lipschitz condition on the graphon function in within a network context. Feng (2024a,b) directly assume that the actual distance between latent factors can be inferred from the noise-free oracle pseudo distance as both N and T_0 grow to infinity, refer to Assumption 4.1 and Theorem A.2 in Feng (2024b). Auerbach (2022) requires a specific type of continuous function. Similar arguments can be found in the network graphon literature, such as Theorem 13.27 in Lovász (2012); for further discussion, see Auerbach (2022).

Theorem 1 (Rate of Convergence for the Outcome Model). *Suppose Assumptions 1 to 4 hold and $\frac{1}{h} \sqrt{\frac{\log(N)}{T_0}} \rightarrow 0$ and $\frac{\log(N)}{h^{2d_\alpha} N} \rightarrow 0$, then*

$$\begin{aligned} \max_{i \in [N]} |\hat{\mu}_{i,t}^{(0)} - \mu_{i,t}^{(0)}| &= O_p \left(h + \sqrt{\frac{\log(N)}{h^{d_\alpha} N}} \right) \\ \max_{i \in [N]} |\hat{p}_{i,t}^{(0)} - p_{i,t}^{(0)}| &= O_p \left(h + \sqrt{\frac{\log(N)}{h^{d_\alpha} N}} \right) \end{aligned}$$

The convergence rate in Theorem 1 is identical to that of the Nadaraya-Watson estimator with observed covariates up to a log term. Setting $h \propto N^{-1/(d_\alpha+2)}$ we get convergence rate $N^{-1/(d_\alpha+2)} \sqrt{\log(N)}$ which, up to a log term, is the optimal rate in Stone (1980) under Lipschitz continuity. However, the rate at which h can converge to zero is restricted by the condition that $\frac{1}{h} \sqrt{\frac{\log(N)}{T_0}} \rightarrow 0$. Thus the theorem above allows us to achieve the rate in Stone (1980), up to log terms, if and only if T_0 grows sufficiently quickly that $N^{2/(d_\alpha+2)}/T_0 \rightarrow 0$. In order to obtain root- N consistency and root- N centered asymptotic normality of the doubly-robust ATT estimator, we require $\max_{i \in [N]} |\hat{\mu}_{i,t}^{(0)} - \mu_{i,t}^{(0)}| = o_p(N^{-1/4})$ and similarly for $\hat{p}_{i,t}$. In the case of a one-dimensional individual latent factor ($d_\alpha = 1$), Theorem 1 ensures that this holds for an appropriate choice of h if and only if T_0 grows quickly enough that $\frac{N^{1/2} \log(N)}{T_0} \rightarrow 0$.

3.2 Inference Framework with Rate Double Robustness

To achieve rate double robustness, different sets of regularity conditions or methodological modifications are required. To avoid over-fitting bias while still attaining the root- N rate of the doubly robust estimator, two approaches can be considered. The first approach involves assuming strong Donsker conditions, i.e., strong smoothness conditions, for both the nuisance functions of the outcome model and the propensity score (e.g., Robins et al., 2008). The second approach utilizes sample splitting techniques to retrieve independence and eliminate over-fitting bias (e.g., Chernozhukov et al., 2018a; Abadie et al., 2024). Given that one advantage of the proposed method is to not rely on structural assumptions about the underlying data-generating process and to impose minimal smoothness conditions, the first approach of imposing Donsker conditions defeats this purpose. Therefore, we adopt the sample splitting approach. Additionally, employing the sample splitting method with a cross-fitting algorithm as suggested in the aforementioned literature and in Algorithm 1 can help us recover any efficiency loss in the estimation procedure within each fold.

However, achieving rate double robustness is particularly challenging in the presence of latent confounders. Abadie et al. (2024) provides an inference procedure for doubly robust

estimators specifically designed for such situations. We will illustrate the assumptions and theorems from [Abadie et al. \(2024\)](#) adapted to our notation. Also, it should be noted that these assumptions are consistent with the assumptions of our proposed method.

Assumption AADS.1 (Positivity; Assumption 1 in [Abadie et al., 2024](#)). The unknown propensity score matrix P is such that

$$\delta \leq p_{i,t} \leq 1 - \delta,$$

for all $i \in \{1, 2, \dots, N\}$ and $t \in \{1, 2, \dots, T\}$, where $0 < \delta \leq 1/2$ is a constant.

Assumption AADS.2 (Zero-mean, independent, and sub-Gaussian noise; Assumption 2 in [Abadie et al., 2024](#)). Fix any $t \in \{1, 2, \dots, T\}$. Then,

- (i) $\{(u_{i,t}, v_{i,t}, \epsilon_{i,t}) : i \in \{1, 2, \dots, N\}\}$ are mean zero and independent (across i),
- (ii) $(u_{i,t}, v_{i,t}) \perp \epsilon_{i,t}$ for every $i \in \{1, 2, \dots, N\}$, and
- (iii) $u_{i,t}$ and $v_{i,t}$ have sub-Gaussian norm bounded by a constant $\bar{\sigma}$ for every $i \in \{1, 2, \dots, N\}$.

Assumption AADS.3 (Assumption 3 in [Abadie et al., 2024](#)). The estimated propensity score matrix $\hat{P}$ is such that

$$\bar{\lambda} \leq \hat{p}_{i,t} \leq 1 - \bar{\delta},$$

for all $i \in \{1, 2, \dots, N\}$ and $t \in \{1, 2, \dots, T\}$, where $0 < \bar{\delta} \leq \delta$ is a constant.

Assumption AADS.4 (Assumption 4 in [Abadie et al., 2024](#)). Fix any $t \in \{1, 2, \dots, T\}$. There exists a non-empty partition $(\mathcal{I}_1, \mathcal{I}_2)$ of the units $\{1, 2, \dots, N\}$ such that

$$\left\{ \left(\hat{p}_{i,t}, \hat{\mu}_{i,t}^{(w)} \right) \right\}_{i \in \mathcal{I}_j} \perp \{ \epsilon_{i,t} \}_{i \in \mathcal{I}_j}$$

and

$$\{ \hat{p}_{i,t} \}_{i \in \mathcal{I}_j} \perp \left\{ \left(\epsilon_{i,t}, \hat{\mu}_{i,t}^{(w)} \right) \right\}_{i \in \mathcal{I}_j}$$

for every $w \in \{0, 1\}$ and $j \in \{1, 2\}$

Assumptions [AADS.1](#) and [AADS.3](#) define the overlapping conditions between the population and its sample estimates for propensity scores. Assumption [AADS.2](#) ensures the independence of error terms across individuals in both the outcome and propensity score models. Additionally, it restricts the tail behavior of the error term distribution, which is

crucial for determining the bounds on the rate. Assumption [AADS.4](#) specifies the allowable sample splitting algorithm that retrieves independence for eliminating over-fitting bias. This assumption is satisfied by the proposed K -fold sample splitting with a cross-fitting algorithm. With these assumptions, the asymptotic normality for the doubly robust estimator is guaranteed by the following theorem.

Theorem AADS.1 (Asymptotic Normality for Doubly Robust Estimator; Theorem 2 in [Abadie et al., 2024](#)). Denote

$$\mathcal{E}(\hat{P}) := \frac{\max_{t \in \{1, \dots, T\}} \left\{ \sum_{i \in \{1, \dots, N\}} (\hat{p}_{i,t} - p_{i,t})^2 \right\}^{1/2}}{\sqrt{N}}$$

and

$$\mathcal{E}(\hat{\mu}) := \sum_{w \in \{0,1\}} \mathcal{E}(\hat{\mu}^{(w)}) := \sum_{w \in \{0,1\}} \frac{\max_{t \in \{1, \dots, T\}} \left\{ \sum_{i \in \{1, \dots, N\}} (\hat{\mu}_{i,t}^{(w)} - \mu_{i,t}^{(w)})^2 \right\}^{1/2}}{\sqrt{N}}$$

Suppose Assumptions [AADS.1-AADS.4](#) and the following conditions hold,

- (i) $\mathcal{E}(\hat{P}) = o_p(1)$ and $\mathcal{E}(\hat{\mu}) = o_p(1)$
- (ii) $\mathcal{E}(\hat{P})\mathcal{E}(\hat{\mu}) = o_p(N^{-1/2})$
- (iii) For every $i \in \{1, \dots, N\}$ and $t \in \{1, \dots, T\}$, let $\sigma_{i,t}^{(0)}$ and $\sigma_{i,t}^{(1)}$ be the standard deviation of $u_{i,t}$ and $v_{i,t}$, respectively. The sequence

$$\bar{\sigma}_t^2 := \frac{1}{N} \sum_{i \in \{1, \dots, N\}} \frac{(\sigma_{i,t}^{(1)})^2}{p_{i,t}} + \frac{1}{N} \sum_{i \in \{1, \dots, N\}} \frac{(\sigma_{i,t}^{(0)})^2}{1 - p_{i,t}},$$

is bounded away from zero as N increases.

Then, for all $t \in \{1, \dots, T\}$,

$$\sqrt{N} \left(\hat{\text{ATE}}_{\cdot,t}^{\text{DR}} - \text{ATE}_{\cdot,t} \right) / \bar{\sigma}_t \xrightarrow{d} \mathcal{N}(0, 1),$$

as $N \rightarrow \infty$.

While the theorem is established with the average treatment effect being the target estimand, as mentioned in [Abadie et al. \(2024\)](#), other causal estimands, such as our main parameter of interest ATT_t , can also achieve asymptotic normality for inference. This is attainable when having proper doubly robust score functions and the appropriate form of variance estimator, as the result is immediate.

Employing Theorem [AADS.1](#) along with the rate of convergence result in Theorem 1, we establish the asymptotic inference for the proposed estimator:

Theorem 2 (Asymptotic Inference for the Proposed Estimator). *Assume Assumptions 1-4 hold, $\frac{N^{1/2} \log(N)}{T_0} \rightarrow 0$ and h is chosen so that $\frac{1}{h} \sqrt{\frac{\log(N)}{T_0}} \rightarrow 0$, it follows that the product rate of the outcome model and the propensity score model becomes faster than $N^{-1/4}$. With the results from Theorem AADS.1, the aforementioned doubly robust estimator for ATT_t , outlined in Algorithm 1, achieves asymptotic normality.*

4 Simulation Study

In this section, we conduct a Monte Carlo simulation to illustrate the finite sample coverage of the proposed inference method that addresses the issue of latent confounders. We consider a large panel data setting with individuals labeled $i = 1, \dots, N$ over time periods $t = 1, \dots, T$, where $N = 500$ and $T = 500$ in this numerical study. The treatment assignment mechanism depends on unobserved latent characteristics for each i and time t . Specifically, the propensity score $p_{i,t}$ follows the model

$$p_{i,t} = \frac{\exp(\alpha_i + \lambda_t)}{1 + \exp(\alpha_i + \lambda_t)}$$

where the individual latent factor $\alpha_i \sim \text{Uniform}(-1, 1)$ and time latent factor $\lambda_t \sim \text{Uniform}(-1, 1)$. With the configuration of both factors located on compact supports, it results in roughly $p_{i,t} \in [0.1192, 0.8808]$, which satisfies the overlapping condition required for the proposed method.

For the outcome model, we consider various data generating processes, as outlined in Table 1, to investigate how the proposed method performs under common linear and nonlinear models. Additionally, we examine how different nonlinear models with varying degrees of smoothness affect the finite sample coverage.

Table 1: Outcome models

Outcome model	Data generating process
1. Additive model	$y_{i,t} = \alpha_i + \lambda_t + \tau w_{i,t} + u_{i,t}$
2. Interactive fixed effects model	$y_{i,t} = \alpha_i \lambda_t + \tau w_{i,t} + u_{i,t}$
3. Gaussian kernel (non-smooth)	$y_{i,t} = \frac{1}{\sqrt{2\pi\theta}} \exp\left\{-\frac{(\alpha_i - \lambda_t)^2}{\theta}\right\} + \tau w_{i,t} + u_{i,t},$ where $\theta = 1/16$
4. Gaussian kernel (moderately smooth)	$y_{i,t} = \frac{1}{\sqrt{2\pi\theta}} \exp\left\{-\frac{(\alpha_i - \lambda_t)^2}{\theta}\right\} + \tau w_{i,t} + u_{i,t},$ where $\theta = 1/2$
5. Gaussian kernel (smooth)	$y_{i,t} = \frac{1}{\sqrt{2\pi\theta}} \exp\left\{-\frac{(\alpha_i - \lambda_t)^2}{\theta}\right\} + \tau w_{i,t} + u_{i,t},$ where $\theta = 1$
6. Different functions for counterfactuals	$y_{i,t} = \begin{cases} \frac{3}{4} (1 - \alpha_i - \lambda_t ^2) + u_{i,t} & , \text{ for } w_{i,t} = 1 \\ \alpha_i \lambda_t + u_{i,t} & , \text{ for } w_{i,t} = 0 \end{cases}$
7. Binary outcome with logit	$\Pr(y_{i,t} = 1) = \frac{\exp(\alpha_i \lambda_t)}{1 + \exp(\alpha_i \lambda_t)} + \tau w_{i,t}$
8. Poisson arrival model	$\Pr(y_{i,t} = k) = \begin{cases} \frac{(5 \cdot \alpha_i \lambda_t + \tau w_{i,t})^k \exp[-(5 \cdot \alpha_i \lambda_t + \tau w_{i,t})]}{k!} & , \text{ for } k = 0, 1, 2, \dots \\ 0 & , \text{ o.w.} \end{cases}$

Notes: Error terms are set to be $u_{i,t} \sim N(0, 0.5^2)$ for all i and t in model 1 to 6. For the treatment effect, τ is set to be 0.5 for models 1 to 5 and 8, and τ is set to be 0.1 for model 7. For model 6 with a more complex data generating process, it has the treatment effect at time t being $\frac{1}{2} - \frac{3}{4} \cdot \lambda_t^2$.

Models 1 and 2 are the workhorse additive and interactive fixed effects models that are commonly used in the panel data literature for tackling unknown confounders, which can be estimated by methods such as [Moon and Weidner \(2015\)](#). Models 3, 4, and 5 are the non-linear Gaussian kernel, which is also used in the simulation design of [Fernández-Val et al. \(2021\)](#). As discussed in [Griebel and Harbrecht \(2014\)](#), the corresponding smoothness of such a nonlinear function can be adjusted by alternating the parameter θ in the Gaussian kernel. Specifically, a lower value of θ results in a panel, or matrix, generated by the Gaussian kernel that exhibits a slower speed of singular value decay. This makes these non-smooth models more challenging to estimate using methods that depend on low-rank approximations, a property also noted in [Zhang et al. \(2017\)](#). Model 6 illustrates a scenario where counterfactual outcomes in different treatment statuses arise from distinct data-generating processes. As the proposed method calculates the distance of latent similarities only based

on pre-treatment periods that share the same treatment status, it allows the following imputation step to be more flexible and can accommodate different designs. Models 7 and 8 are discrete outcome models that are difficult to estimate when latent confounders are present. The proposed method is expected to address these challenges, as the pseudo distance is originally used to estimate the graphon with adjacency matrix in a network context (e.g., [Zhang et al., 2017](#); [Zeleneev, 2020](#); [Wang, 2022](#)).

We simulate the aforementioned models with the number of replications being 500. In these simulations, we utilize the Epanechnikov kernel, that satisfies the boundedness assumption and is defined as $K(x) = \frac{3}{4}(1 - x^2) \cdot \mathbf{1}\{|x| \leq 1\}$, for both outcome and propensity score imputations in the proposed method. Using this kernel, we compared the performance of different methods: (i) Knowing the true latent factor α_i for similarity in ℓ_2 distance calculation and the true propensity score for doubly robust estimation. (ii) Knowing the true latent factor α_i for similarity in ℓ_2 distance calculation, but the true propensity score is unknown and needs to be estimated with the proposed method. (iii) Estimate the pseudo distance and impute both outcome and propensity score without cross-fitting. (iv) Estimate the pseudo distance and impute both outcome and propensity score with 2-fold cross-fitting. Methods (i) and (ii) serve as benchmark methods for comparison, allowing us to assess how well the pseudo distance reflects the actual similarities in the unobserved latent characteristics.

In the numerical experiment, the bandwidth is selected between 0.05 and 3. For methods (i) to (iii), we choose the smallest bandwidth that still allowed the estimation of the outcome and the propensity score, in line with the assumption that the bandwidth shrinks toward zero. For method (iv), the bandwidth is determined using cross-validation through the 2-fold cross-fitting procedure, which generally yields a value close to the smallest bandwidth that keeps the model estimable in our numerical experiments. In [Table 2](#), we report the 95% confidence interval coverage and the average bandwidth selection across 500 replications in our simulation for each outcome model and each method.

Table 2: Simulation results

Outcome model	Model	95% C.I. coverage	Average bandwidth
1. Additive model	With true $p_{i,t}$	93.80%	0.0541
	With true α_i	93.40%	0.0541
	Pseudo distance	90.00%	0.1156
	Pseudo distance w/ CF	95.60%	0.1575
2. Interactive fixed effects model	With true $p_{i,t}$	94.00%	0.0541
	With true α_i	93.40%	0.0541
	Pseudo distance	94.00%	0.1000
	Pseudo distance w/ CF	94.20%	0.1008
3. Gaussian kernel (non-smooth)	With true $p_{i,t}$	92.40%	0.0541
	With true α_i	93.80%	0.0541
	Pseudo distance	90.40%	0.6798
	Pseudo distance w/ CF	94.20%	1.1600
4. Gaussian kernel (moderately smooth)	With true $p_{i,t}$	94.00%	0.0541
	With true α_i	93.20%	0.0541
	Pseudo distance	94.60%	0.1000
	Pseudo distance w/ CF	94.40%	0.1001
5. Gaussian kernel (smooth)	With true $p_{i,t}$	94.00%	0.0541
	With true α_i	93.20%	0.0541
	Pseudo distance	88.60%	0.1000
	Pseudo distance w/ CF	89.00%	0.1000
6. Different functions for counterfactuals	With true $p_{i,t}$	92.00%	0.0541
	With true α_i	91.20%	0.0541
	Pseudo distance	88.60%	0.0998
	Pseudo distance w/ CF	88.20%	0.1000
7. Binary outcome with logit	With true $p_{i,t}$	90.00%	0.0541
	With true α_i	90.80%	0.0541
	Pseudo distance	90.00%	0.7142
	Pseudo distance w/ CF	94.40%	0.8476
8. Poisson arrival model	With true $p_{i,t}$	93.40%	0.0541
	With true α_i	92.40%	0.0541
	Pseudo distance	94.40%	0.1000
	Pseudo distance w/ CF	94.40%	0.1007

For models 1 and 2 with the additive and interactive fixed effect model, the proposed method demonstrates performance comparable to the benchmark methods (i) and (ii) regarding the 95% confidence interval coverage. This suggests that the pseudo distance measurement effectively captures the unobserved latent similarities in low-rank panels with smooth data generating processes.

In models 3 to 5, which involve Gaussian kernel data generating processes with varying degrees of smoothness, the proposed method continues to perform well, particularly in the case of Model 3, which represents an extremely non-smooth data generating process. This finding aligns with the work of [Zhang et al. \(2017\)](#): their neighborhood smoothing approach—estimating the graphon using adjacency matrices within a network context—performs effectively well even when the synthetic graphons generated are full rank. Moreover, the proposed kernel method, which assigns different weights to individuals based on their latent similarities as measured by the pseudo distance, advantages significantly from non-smooth data generating processes, as individuals with similar latent characteristics can be more accurately separated.

In model 6, where counterfactual outcomes are generated by different functions, the proposed method shows slightly lower confidence interval coverage. However, it still performs comparably to the benchmark methods (i) and (ii), which rely on the scenario that latent characteristics need to be observed.

For models 7 and 8, which focus on discrete outcomes, the proposed method, when combined with cross-fitting, achieves good confidence interval coverage.

These simulation results indicate that across all outcome models—whether they are linear or nonlinear, smooth or non-smooth, and continuous or discrete—the proposed doubly robust estimation procedure utilizing pseudo distance measurements for latent similarities performs well, even in the presence of confounding treatment assignment mechanisms.

References

- Abadie, A. (2021). Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of economic literature*, 59(2):391–425.
- Abadie, A., Agarwal, A., Dwivedi, R., and Shah, A. (2024). Doubly robust inference in causal latent factor models. *arXiv preprint arXiv:2402.11652*.
- Abbring, J. H. and Van Den Berg, G. J. (2003). The nonparametric identification of treatment effects in duration models. *Econometrica*, 71:1491–1517.
- Agarwal, A., Dahleh, M., Shah, D., and Shen, D. (2021). Causal matrix completion. *arXiv preprint arXiv:2109.15154*.
- Agarwal, A., Dahleh, M., Shah, D., and Shen, D. (2023). Causal matrix completion. In *The thirty sixth annual conference on learning theory*, pages 3821–3826. PMLR.
- Agarwal, A., Shah, D., and Shen, D. (2020). Synthetic interventions. *arXiv preprint arXiv:2006.07691*.
- Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W., and Wager, S. (2021). Synthetic difference-in-differences. *American Economic Review*, 111(12):4088–4118.
- Arkhangelsky, D. and Imbens, G. (2024). Causal models for longitudinal and panel data: A survey. *The Econometrics Journal*, 27(3):C1–C61.
- Arkhangelsky, D. and Imbens, G. W. (2022). Doubly robust identification for causal panel data models. *The Econometrics Journal*, 25(3):649–674.
- Armstrong, T. B., Weidner, M., and Zeleneev, A. (2022). Robust estimation and inference in panels with interactive fixed effects. *arXiv preprint arXiv:2210.06639*.
- Athey, S., Bayati, M., Doudchenko, N., Imbens, G., and Khosravi, K. (2021). Matrix completion methods for causal panel data models. *Journal of the American Statistical Association*, 116(536):1716–1730.
- Auerbach, E. (2022). Identification and estimation of a partially linear regression model using network data. *Econometrica*, 90(1):347–365.
- Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica*, 77(4):1229–1279.

- Bai, J. and Ng, S. (2021). Matrix completion, counterfactuals, and factor analysis of missing data. *Journal of the American Statistical Association*, 116(536):1746–1763.
- Bennett, G. (1962). Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45.
- Borusyak, K., Jaravel, X., and Spiess, J. (2022). Revisiting event study designs: Robust and efficient estimation. *arXiv preprint arXiv:2108.12419*.
- Borusyak, K., Jaravel, X., and Spiess, J. (2024). Revisiting event-study designs: robust and efficient estimation. *Review of Economic Studies*, page rdae007.
- Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.
- Callaway, B. and Sant’Anna, P. H. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225(2):200–230.
- Cattaneo, M. D., Feng, Y., and Titiunik, R. (2021). Prediction intervals for synthetic control methods. *Journal of the American Statistical Association*, 116(536):1865–1880.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018a). Double/debiased machine learning for treatment and structural parameters: Double/debiased machine learning. *The Econometrics Journal*, 21(1).
- Chernozhukov, V., Wuthrich, K., and Zhu, Y. (2018b). A t -test for synthetic controls. *arXiv preprint arXiv:1812.10820*.
- Chernozhukov, V., Wüthrich, K., and Zhu, Y. (2021). An exact and robust conformal inference method for counterfactual and synthetic controls. *Journal of the American Statistical Association*, 116(536):1849–1864.
- De Chaisemartin, C. and d’Haultfoeuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American economic review*, 110(9):2964–2996.
- De Chaisemartin, C. and d’Haultfoeuille, X. (2023). Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey. *The Econometrics Journal*, 26(3):C1–C30.
- Feng, Y. (2024a). Causal inference in possibly nonlinear factor models. *Working paper*.

- Feng, Y. (2024b). Optimal estimation of large-dimensional nonlinear factor models. *Working paper*.
- Fernández-Val, I., Freeman, H., and Weidner, M. (2021). Low-rank approximations of non-separable panel models. *The Econometrics Journal*, 24(2):C40–C77.
- Ghanem, D., Sant’Anna, P. H., and Wüthrich, K. (2022). Selection and parallel trends. *arXiv preprint arXiv:2203.09001*.
- Goldsmith-Pinkham, P. (2024). Tracking the credibility revolution across fields. *arXiv preprint arXiv:2405.20604*.
- Griebel, M. and Harbrecht, H. (2014). Approximation of bi-variate functions: singular value decomposition versus sparse grids. *IMA journal of numerical analysis*, 34(1):28–54.
- Issartel, Y. (2020). *Inference on random networks*. PhD thesis, Faculté des sciences d’Orsay, Université Paris-Saclay.
- Lovász, L. (2012). *Large networks and graph limits*, volume 60. American Mathematical Soc.
- Manski, C. F. and Pepper, J. V. (2018). How do right-to-carry laws affect crime rates? coping with ambiguity using bounded-variation assumptions. *Review of Economics and Statistics*, 100(2):232–244.
- Moon, H. R. and Weidner, M. (2015). Linear regression for panel with unknown number of factors as interactive fixed effects. *Econometrica*, 83(4):1543–1579.
- Pesaran, M. H. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica*, 74(4):967–1012.
- Rambachan, A. and Roth, J. (2023). A more credible approach to parallel trends. *Review of Economic Studies*, 90(5):2555–2591.
- Robins, J., Li, L., Tchetgen, E., van der Vaart, A., et al. (2008). Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and statistics: essays in honor of David A. Freedman*, volume 2, pages 335–422. Institute of Mathematical Statistics.
- Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866.

- 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.
- Sant’Anna, P. H. and Zhao, J. (2020). Doubly robust difference-in-differences estimators. *Journal of econometrics*, 219(1):101–122.
- Smucler, E., Rotnitzky, A., and Robins, J. M. (2019). A unifying approach for doubly-robust ℓ_1 regularized estimation of causal contrasts. *arXiv preprint arXiv:1904.03737*.
- Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *The annals of Statistics*, pages 1348–1360.
- Sun, L. and Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 225(2):175–199.
- Wang, Z. (2022). The linking effect: Causal identification and estimation of the effect of peer relationship. In *4th EUI Alumni Conference in Economics*.
- Wooldridge, J. M. (2021). Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators. *Available at SSRN 3906345*.
- Zeleneev, A. (2020). Identification and estimation of network models with nonparametric unobserved heterogeneity.
- Zhang, Y., Levina, E., and Zhu, J. (2017). Estimating network edge probabilities by neighbourhood smoothing. *Biometrika*, 104(4):771–783.

Appendix

A Proofs

A.1 Supporting Lemmas

Lemma A.1 (Rate of Convergence for the Pseudo Distance Estimator). *Suppose Assumption 4 holds and $\sqrt{\frac{\log(N)}{T_0}} \rightarrow 0$. Then for all $i, j \in [N]$*

$$\max_{i,j} |\hat{d}_{i,j} - d_{i,j}| = O_p \left(\sqrt{\frac{\log(N)}{T_0}} + \frac{\log(N)}{N} \right).$$

Proof of Lemma A.1. By Assumption 4 (iii) we can write $d_{i,j}$ as follows:

$$d_{i,j} = \sup_{\alpha_k \in \text{supp}(\alpha)} |\mathbb{E}[\mu^{(0)}(\alpha_k, \lambda_t)(\mu^{(0)}(\alpha_i, \lambda_t) - \mu^{(0)}(\alpha_j, \lambda_t)) | \alpha_i, \alpha_j, \alpha_k]|$$

We further denote the oracle distance $\tilde{d}_{i,j}$, which will be helpful throughout the proof, as follows:

$$\tilde{d}_{i,j} := \max_{k \notin \{i,j\}} \left| \frac{1}{T_0} \sum_{t=1}^{T_0} \mu_{k,t}^{(0)} (\mu_{i,t}^{(0)} - \mu_{j,t}^{(0)}) \right|$$

The proof is then separated into two parts: (i) the rate of convergence from $\hat{d}_{i,j}$ to $\tilde{d}_{i,j}$ and (ii) the rate of convergence from $\tilde{d}_{i,j}$ to $d_{i,j}$.

From $\hat{d}_{i,j}$ to $\tilde{d}_{i,j}$ By applying the reverse triangle inequality and multiplying out, we derive the max difference between estimated distance and oracle distance as

$$\begin{aligned} & \max_{i,j} |\hat{d}_{i,j} - \tilde{d}_{i,j}| \\ &= \left| \max_{i,j,k: k \notin \{i,j\}} \left| \frac{1}{T_0} \sum_{t=1}^{T_0} (\mu_{k,t}^{(0)} + u_{k,t})(\mu_{i,t}^{(0)} - \mu_{j,t}^{(0)} + u_{j,t} - u_{j,t}) \right| - \max_{k \notin \{i,j\}} \left| \frac{1}{T_0} \sum_{t=1}^{T_0} \mu_{k,t}^{(0)} (\mu_{i,t}^{(0)} - \mu_{j,t}^{(0)}) \right| \right| \\ &\leq \max_{i,j,k: k \notin \{i,j\}} \left| \frac{1}{T_0} \sum_{t=1}^{T_0} [u_{k,t} (\mu_{i,t}^{(0)} - \mu_{j,t}^{(0)}) + \mu_{k,t}^{(0)} (u_{j,t} - u_{j,t}) + u_{k,t} (u_{j,t} - u_{j,t})] \right| \end{aligned} \quad (9)$$

By Assumptions 4 (ii), for each $r, q \in \{i, j, k\}$ with $r \neq q$, it follows that $\mathbb{E}[u_{k,t} | \alpha_i, \alpha_j, \alpha_k, \lambda_t] = 0$ and $\mathbb{E}[u_{k,t}(u_{j,t} - u_{j,t}) | \alpha_i, \alpha_j, \alpha_k, \lambda_t] = 0$. Moreover, by Assumption 4 (iv), $\mu_{i,t}^{(0)}$, $\mu_{j,t}^{(0)}$, and $\mu_{k,t}^{(0)}$ are bounded, and by Assumption 4 (ii), $\{(u_{i,t}, u_{j,t}, u_{k,t})\}_{t=1}^{T_0}$ are jointly independent over time and sub-Gaussian conditional on $\alpha_i, \alpha_j, \alpha_k, \lambda_t$. Thus, from Bernstein's inequality in Lemma A.7, we get that for constants C, a , and b , for any $\eta > 0$, it follows that

$$\mathbb{P} \left(\left| \frac{1}{T_0} \sum_{t=1}^{T_0} [u_{k,t} (\mu_{i,t}^{(0)} - \mu_{j,t}^{(0)}) + \mu_{k,t}^{(0)} (u_{j,t} - u_{j,t}) + u_{k,t} (u_{j,t} - u_{j,t})] \right| > \eta \mid \alpha_i, \alpha_j, \alpha_k, \lambda_t \right) \leq C \exp \left(-\frac{T_0 \eta^2}{a + b \eta} \right)$$

And so, applying the law of iterated expectations gives

$$\mathbb{P}\left(\left|\frac{1}{T_0} \sum_{t=1}^{T_0} [(u_{k,t}(\mu_{i,t}^{(0)} - \mu_{j,t}^{(0)}) + \mu_{k,t}^{(0)}(u_{j,t} - u_{j,t}) + u_{k,t}(u_{j,t} - u_{j,t}))]\right| > \eta\right) \leq C \exp\left(-\frac{T_0 \eta^2}{a + b\eta}\right)$$

By taking the union bound and (9), we then get

$$\begin{aligned} \mathbb{P}\left(\max_{i,j} |\hat{d}_{i,j} - \tilde{d}_{i,j}| > \eta\right) &\leq \binom{N}{3} C \exp\left(-\frac{T_0 \eta^2}{a + b\eta}\right) \\ &\leq N^3 C \exp\left(-\frac{T_0 \eta^2}{a + b\eta}\right) \end{aligned}$$

By setting $\eta = \sqrt{\frac{3a \log(N)}{T_0}}$ in the above inequality, it follows that

$$\mathbb{P}\left(\max_{i,j} |\hat{d}_{i,j} - \tilde{d}_{i,j}| > \sqrt{\frac{a \log(N)}{T_0}}\right) \leq C \exp\left(3 \log(N) - \frac{3a \log(N)}{a + b \sqrt{\frac{a \log(N)}{T_0}}}\right)$$

As the RHS goes to C , hence, we conclude the rate for convergence from $\hat{d}_{i,j}$ to $\tilde{d}_{i,j}$ being

$$\max_{i,j} |\hat{d}_{i,j} - \tilde{d}_{i,j}| = \mathcal{O}_p\left(\sqrt{\frac{\log(N)}{T_0}}\right)$$

From $\tilde{d}_{i,j}$ to $d_{i,j}$ By applying the triangle inequality and reverse triangle inequality, we derive the max difference between oracle distance and actual distance as

$$\begin{aligned} &\max_{i,j,k: k \notin \{i,j\}} |d_{i,j} - \tilde{d}_{i,j}| \\ &\leq \max_{i,j,k: k \notin \{i,j\}} \left| \frac{1}{T_0} \sum_{t=1}^{T_0} \mu_{k,t}^{(0)} (\mu_{i,t}^{(0)} - \mu_{j,t}^{(0)}) - \mathbb{E}[\mu_{k,t}^{(0)} (\mu_{i,t}^{(0)} - \mu_{j,t}^{(0)}) \mid \alpha_i, \alpha_j, \alpha_k] \right| \end{aligned} \quad (10)$$

$$+ \max_{i,j} \sup_{\alpha \in \text{supp}(\alpha)} \left| \mathbb{E}[\mu_{k,t}^{(0)} (\mu_{i,t}^{(0)} - \mu_{j,t}^{(0)}) \mid \alpha_i, \alpha_j] - \max_{i,j,k: k \notin \{i,j\}} \left| \mathbb{E}[\mu_{k,t}^{(0)} (\mu_{i,t}^{(0)} - \mu_{j,t}^{(0)}) \mid \alpha_i, \alpha_j, \alpha_k] \right| \right| \quad (11)$$

We first derive the rate for (10). By Assumptions 4 (iii) and (iv), it follows that $\mu_{i,t}^{(0)}$, $\mu_{j,t}^{(0)}$, and $\mu_{k,t}^{(0)}$ are bounded and independent over time conditional on $\alpha_i, \alpha_j, \alpha_k$. Then by Hoeffding's inequality, there are constants C , a , and b such that for any $\eta > 0$,

$$\mathbb{P}\left(\left|\frac{1}{T_0} \sum_{t=1}^{T_0} \mu_{k,t}^{(0)} (\mu_{i,t}^{(0)} - \mu_{j,t}^{(0)}) - \mathbb{E}[\mu_{k,t}^{(0)} (\mu_{i,t}^{(0)} - \mu_{j,t}^{(0)}) \mid \alpha_i, \alpha_j, \alpha_k]\right| \mid \alpha_i, \alpha_j, \alpha_k\right) \leq C \exp\left(-\frac{T_0 \eta^2}{a + b\eta}\right)$$

which further leads to

$$\mathbb{P}\left(\left|\frac{1}{T_0} \sum_{t=1}^{T_0} \mu_{k,t}^{(0)} (\mu_{i,t}^{(0)} - \mu_{j,t}^{(0)}) - \mathbb{E}[\mu_{k,t}^{(0)} (\mu_{i,t}^{(0)} - \mu_{j,t}^{(0)}) \mid \alpha_i, \alpha_j, \alpha_k]\right| \leq C \exp\left(-\frac{T_0 \eta^2}{a + b\eta}\right)\right)$$

Then applying union bound gives

$$\mathbb{P}\left(\max_{i,j,k: k \notin \{i,j\}} \left|\frac{1}{T_0} \sum_{t=1}^{T_0} \mu_{k,t}^{(0)} (\mu_{i,t}^{(0)} - \mu_{j,t}^{(0)}) - \mathbb{E}[\mu_{k,t}^{(0)} (\mu_{i,t}^{(0)} - \mu_{j,t}^{(0)}) \mid \alpha_i, \alpha_j, \alpha_k]\right| \leq N^3 C \exp\left(-\frac{T_0 \eta^2}{a + b\eta}\right)\right)$$

By setting $\eta = \sqrt{\frac{3a \log(N)}{T_0}}$ in the above inequality, it follows that

$$\max_{i,j,k: k \notin \{i,j\}} \left| \frac{1}{T_0} \sum_{t=1}^{T_0} \mu_{k,t}^{(0)} (\mu_{i,t}^{(0)} - \mu_{j,t}^{(0)}) - \mathbb{E}[\mu_{k,t}^{(0)} (\mu_{i,t}^{(0)} - \mu_{j,t}^{(0)}) \mid \alpha_i, \alpha_j, \alpha_k] \right| = O_p \left(\sqrt{\frac{\log(N)}{T_0}} \right)$$

Then we derive the rate of convergence for (11). By the reverse triangle inequality, the boundedness of $\mu_{i,t}^{(0)}$, and the Lipschitz condition in Assumption 4 (iv), we have that, for some constant C ,

$$\begin{aligned} & \max_{i,j} \sup_{\alpha \in \text{supp}(\alpha)} \left| \mathbb{E}[\mu^{(0)}(\alpha, \lambda_t) (\mu_{i,t}^{(0)} - \mu_{j,t}^{(0)}) \mid \alpha_i, \alpha_j] \right| - \max_{i,j,k: k \notin \{i,j\}} \left| \mathbb{E}[\mu_{k,t}^{(0)} (\mu_{i,t}^{(0)} - \mu_{j,t}^{(0)}) \mid \alpha_i, \alpha_j, \alpha_k] \right| \\ & \leq \max_{i,j} \sup_{\alpha \in \text{supp}(\alpha)} \min_{k \notin \{i,j\}} \left| \mathbb{E}[(\mu^{(0)}(\alpha, \lambda_t) - \mu^{(0)}(\alpha_k, \lambda_t)) (\mu_{i,t}^{(0)} - \mu_{j,t}^{(0)}) \mid \alpha_i, \alpha_j, \alpha_k] \right| \\ & \leq C \sup_{\alpha \in \text{supp}(\alpha)} \min_k \|\alpha - \alpha_k\| \end{aligned}$$

By the compactness of the support of α in Assumption 4 (iii), with a constant C , there is an η -covering of the support of α_i that consists of $J \leq \frac{C}{\eta^{d_\alpha}}$ open balls whose centers we will denote by $\{\tilde{\alpha}_j\}_{j=1}^J$. Then by definition of the covering, it follows that

$$\sup_{\alpha \in \text{supp}(\alpha)} \min_k \|\alpha - \alpha_k\| \leq \max_{j \in [J]} \min_{k \in [N]} \|\tilde{\alpha}_j - \alpha_k\| + L_0 \eta$$

Then applying the union bound gives

$$\mathbb{P} \left(\max_{j \in [J]} \min_{k \in [N]} \|\tilde{\alpha}_j - \alpha_k\| > \epsilon \right) \leq \sum_{j=1}^J \mathbb{P} \left(\min_{k \in [N]} \|\tilde{\alpha}_j - \alpha_k\| > \epsilon \right)$$

By Assumption 4 (iii), there is a $0 < \delta$ so that for any fixed $\alpha \in \text{supp}(\alpha)$ and any η , $\mathbb{P}(\|\alpha - \alpha_k\| > \eta) \leq 1 - \delta\eta$. With the independence of $\{\alpha_k\}_{k=1}^N$, we have

$$\mathbb{P} \left(\min_{k \in [N]} \|\tilde{\alpha}_j - \alpha_k\| > \epsilon \right) = \left\{ \mathbb{P}(\|\tilde{\alpha}_j - \alpha_k\| > \epsilon) \right\}^N \leq (1 - \delta\epsilon)^N$$

which further leads to

$$\mathbb{P} \left(\max_{j \in [J]} \min_{k \in [N]} \|\tilde{\alpha}_j - \alpha_k\| > \epsilon \right) \leq J \delta^N \epsilon^N \leq \frac{C}{\eta^{d_\alpha}} (1 - \delta\epsilon)^N$$

By setting $\epsilon = \frac{d_\alpha \log(N)}{\delta N}$ and $\eta = \frac{1}{N}$, it follows that

$$\mathbb{P} \left(\max_{j \in [J]} \min_{k \in [N]} \|\tilde{\alpha}_j - \alpha_k\| > \frac{\log(N)}{\delta N} \right) \leq J \delta^N \epsilon^N \leq \frac{C}{N^{d_\alpha}} \left(1 - \frac{d_\alpha \log(N)}{N} \right)^N \rightarrow C$$

Hence, we have

$$\sup_{\alpha \in \text{supp}(\alpha)} \min_k \|\alpha - \alpha_k\| = O_p \left(\frac{\log(N)}{N} \right)$$

Combining with previous derivation, therefore, we conclude the rate for convergence from $\tilde{d}_{i,j}$ to $d_{i,j}$ being

$$\max_{i,j} |\hat{d}_{i,j} - d_{i,j}| = O_p \left(\sqrt{\frac{\log(N)}{T_0}} + \frac{\log(N)}{N} \right)$$

□

Lemma A.2. Under Assumptions 4 (iv) and (vi), there exists a $c > 0$ so that for all $\alpha_1, \alpha_2 \in \mathcal{A}$:

$$\sup_{\alpha \in \text{supp}(\alpha)} \int_{\lambda \in \text{supp}(\lambda)} \mu^{(0)}(\alpha_1, \lambda) (\mu^{(0)}(\alpha_1, \lambda) - \mu^{(0)}(\alpha_2, \lambda)) d\pi(\lambda) \geq c \|\alpha_1 - \alpha_2\|$$

Proof of Lemma A.2. By Assumption 4 (vi), for some $C > 0$ we have:

$$\begin{aligned} & C \sqrt{\int_{\lambda \in \text{supp}(\lambda)} (\mu^{(0)}(\alpha_1, \lambda) - \mu^{(0)}(\alpha_2, \lambda))^2 d\pi(\lambda)} \\ & \leq \sup_{\alpha \in \text{supp}(\alpha)} \int_{\lambda \in \text{supp}(\lambda)} \mu^{(0)}(\alpha, \lambda) (\mu^{(0)}(\alpha_1, \lambda) - \mu^{(0)}(\alpha_2, \lambda)) d\pi(\lambda) \end{aligned}$$

and the object on the LHS is bounded below by $C\|\alpha_1 - \alpha_2\|$. Thus letting $c = C^2 > 0$ gives the result. □

Lemma A.3. There exists a $\delta > 0$ so that with probability approaching 1, for all $i, j \in [N]$:

$$K\left(\frac{\hat{d}_{i,j}}{h}\right) = \mathbf{1}\{\|\alpha_i - \alpha_j\| \leq \delta h\} K\left(\frac{\hat{d}_{i,j}}{h}\right)$$

Proof of Lemma A.3. By Lemma A.2, there exists a constant C , such that for any finite $\delta > 0$, we have

$$\|\alpha_i - \alpha_j\| > \delta h \implies d_{i,j} > 2C\delta h$$

Now, note that for all i, j ,

$$\frac{\hat{d}_{i,j}}{h} \geq \frac{d_{i,j}}{h} - \frac{\max_{i,j} |\hat{d}_{i,j} - d_{i,j}|}{h}.$$

Thus, it follows that

$$\|\alpha_i - \alpha_j\| > \delta h \implies \frac{\hat{d}_{i,j}}{h} > \delta C - \frac{\max_{i,j} |\hat{d}_{i,j} - d_{i,j}|}{h}$$

By Lemma A.1, we have

$$\frac{\max_{i,j} |\hat{d}_{i,j} - d_{i,j}|}{h} = O_p \left(\frac{1}{h} \sqrt{\frac{\log(N)}{T_0}} + \frac{\log(N)}{hN} \right)$$

By supposition, $\frac{1}{h} \sqrt{\frac{\log(N)}{T_0}}$ and $\frac{\log(N)}{hN}$ go to zero, and so with probability approaching one, we have $\frac{\max_{i,j} |\hat{d}_{i,j} - d_{i,j}|}{h} \leq \delta C$. Hence, with probability approaching one, for all i, j , it follows that

$$\|\alpha_i - \alpha_j\| > \delta h \implies \frac{\hat{d}_{i,j}}{h} > \delta C$$

As Assumption 4 (v) states that the kernel is zero everywhere other than some compact (and therefore bounded) sets. Hence, there is a constant c such that

$$\frac{\hat{d}_{i,j}}{h} > c \implies K\left(\frac{\hat{d}_{i,j}}{h}\right) = 0$$

Finally, setting $\delta > \frac{c}{\bar{c}}$ gives the desired result. $\square$

Lemma A.4. Suppose Assumption 4 holds, and $\frac{\log N}{Nh^{2d_\alpha}} \rightarrow 0$. Then for any fixed δ , there are constants $0 < \underline{c} < \bar{c} < \infty$ and $0 < \tilde{c}$, such that with probability approaching one and for all $i \in [N]$, it follows that

$$\underline{c}h^{d_\alpha} \leq \frac{1}{N} \sum_{j:w_{j,t}=0} \mathbf{1}\{\|\alpha_i - \alpha_j\| \leq \delta h\} \leq \bar{c}h^{d_\alpha}$$

and

$$\tilde{c} \leq \frac{1}{Nh^{d_\alpha}} \sum_{j:w_{j,t}=0} K(d_{i,j}/h)$$

Proof of Lemma A.4. For the first statement in Lemma A.4, by Assumption 4 (i), for some $c > 0$, it follows that $\mathbb{P}(1 - w_{j,t} | \alpha_j) \geq c$ almost surely. And with the fact that $w_{j,t}$ is independent of α_i and that $\{\alpha_j\}_{j=1}^N$ are i.i.d. (by Assumption 4 (iii)), we have

$$\begin{aligned} \mathbb{E} \left[\frac{1}{N} \sum_j (1 - w_{j,t}) \mathbf{1}\{\|\alpha_i - \alpha_j\| \leq \delta h\} \mid \alpha_i \right] &= \mathbb{E} \left[\frac{1}{N} \sum_j P(1 - w_{j,t} | \alpha_j) \mathbf{1}\{\|\alpha_i - \alpha_j\| \leq \delta h\} \mid \alpha_i \right] \\ &\geq c \cdot \mathbb{E} \left[\frac{1}{N} \sum_j \mathbf{1}\{\|\alpha_i - \alpha_j\| \leq \delta h\} \mid \alpha_i \right] \\ &= c \cdot \mathbb{P}(\|\alpha_i - \alpha_j\| \leq \delta h \mid \alpha_i) \end{aligned}$$

With Assumption 4 (iii), for any fixed δ , there are constants $0 < \underline{c} < \bar{c} < \infty$ such that

$$2\underline{c}h^{d_\alpha} \leq \mathbb{P}(\|\alpha_i - \alpha_j\| \leq \delta h \mid \alpha_i) \leq \frac{1}{2}\bar{c}h^{d_\alpha}$$

By applying the union bound and Hoeffding's inequality for Bernoulli random variables, it follows that, for any C ,

$$\begin{aligned} &\mathbb{P} \left(\max_{i \in [N]} \left| \frac{1}{N} \sum_j (1 - w_{j,t}) \cdot \mathbf{1}\{\|\alpha_i - \alpha_j\| \leq \delta h\} - \mathbb{E} \left[\frac{1}{N} \sum_j (1 - w_{j,t}) \cdot \mathbf{1}\{\|\alpha_i - \alpha_j\| \leq \delta h\} \mid \alpha_i \right] \right| > Ch^{d_\alpha} \right) \\ &\leq 2N \cdot \exp(-2NC^2h^2) \\ &= 2 \cdot \exp(\log(N) - 2NC^2h^{2d_\alpha}) \end{aligned}$$

The RHS goes to zero if $\frac{\log N}{Nh^{2d_\alpha}} \rightarrow 0$, which holds by supposition. By setting $C = \underline{c}$ and $C = \frac{1}{2}\bar{c}$, we see that with probability approaching one, for all $i \in [N]$, it follows that

$$\underline{c}h^{d_\alpha} \leq \frac{1}{N} \sum_{j: w_{j,t}=0} (1 - w_{j,t}) \mathbf{1}\{\|\alpha_i - \alpha_j\| \leq \delta h\} \leq \bar{c}h^{d_\alpha}$$

Now for the second statement in Lemma A.4, by Assumption 4 (iv), it follows that $\mu(\cdot, \lambda)$ is Lipschitz continuous and bounded (uniformly over λ in both cases). Thus, for some constant $C < \infty$, we have

$$\|\alpha_i - \alpha_j\| \leq x \implies d_{i,j} \leq Cx$$

By Assumption 4 (v) of the kernel being positive and strictly positive at zero, and alongside with Lipschitz continuous, thus, there must be some $a, b > 0$ such that for all $x \leq a$, $K(x) \geq b$. Combining these results gives

$$\|\alpha_i - \alpha_j\| \leq \frac{a}{C}h \implies d_{i,j} \leq ah \implies K(d_{i,j}/h) \geq b$$

Then with the positivity of the kernel, it follows that

$$b \cdot \frac{1}{N} \sum_j (1 - w_{j,t}) \mathbf{1}\{\|\alpha_i - \alpha_j\| \leq \frac{a}{C} \cdot h\} \leq \frac{1}{N} \sum_j (1 - w_{j,t}) \cdot K(d_{i,j}/h)$$

But since we already established that for any fixed δ , there is a constant $\underline{c} > 0$ such that $\underline{c}h^{d_\alpha} \leq \frac{1}{N} \sum_j (1 - w_{j,t}) \mathbf{1}\{\|\alpha_i - \alpha_j\| \leq \delta h\}$ for all $i \in [N]$ with probability approaching one, therefore, there is a constant $\tilde{c} > 0$ such that with probability approaching one, for all i ,

$$\tilde{c} \leq \frac{1}{Nh^{d_\alpha}} \sum_j (1 - w_{j,t}) \cdot K(d_{i,j}/h)$$

□

A.2 Proof of Theorem 1

Proof of Theorem 1. We prove the result for $\hat{\mu}_{i,t}^{(0)}$. The result for $\hat{p}_{i,t}$ follows by almost identical steps. By Lemma A.3, there exists a $\delta < \infty$ such that with probability approaching one, for all $i, j \in [N]$, we have

$$K(\hat{d}_{i,j}/h) = K(\hat{d}_{i,j}/h) \cdot \mathbf{1}\{\|\alpha_i - \alpha_j\| \leq \delta h\}$$

And with the above equality holds, we can then decompose $\hat{\mu}_{i,t}^{(0)} - \mu_{i,t}^{(0)}$ as follows:

$$\begin{aligned} \hat{\mu}_{i,t}^{(0)} - \mu_{i,t}^{(0)} &= \frac{\sum_{j:w_{j,t}=0} \mathbf{1}\{\|\alpha_i - \alpha_j\| \leq \delta h\} K(\hat{d}_{i,j}/h) \cdot \mu_{j,t}^{(0)}}{\sum_{j:w_{j,t}=0} \mathbf{1}\{\|\alpha_i - \alpha_j\| \leq \delta h\} K(\hat{d}_{i,j}/h)} - \mu_{i,t}^{(0)} \\ &\quad + \frac{\frac{1}{h^{d_\alpha} N} \sum_{j:w_{j,t}=0} \mathbf{1}\{\|\alpha_i - \alpha_j\| \leq \delta h\} K(\hat{d}_{i,j}/h) \cdot u_{j,t}}{\frac{1}{h^{d_\alpha} N} \sum_{j:w_{j,t}=0} K(\hat{d}_{i,j}/h)} \end{aligned} \quad (12)$$

The quantity on the first line of the RHS above is a weighted average of $\mu_{j,t}^{(0)}$ over individuals j for whom $\|\alpha_i - \alpha_j\| \leq \delta h$. Using the Lipschitz condition in Assumption 4 (iv), we have

$$\begin{aligned} \left| \frac{\sum_{j:w_{j,t}=0} \mathbf{1}\{\|\alpha_i - \alpha_j\| \leq \delta h\} K(\hat{d}_{i,j}/h) \cdot \mu_{j,t}^{(0)}}{\sum_{j:w_{j,t}=0} \mathbf{1}\{\|\alpha_i - \alpha_j\| \leq \delta h\} K(\hat{d}_{i,j}/h)} - \mu_{i,t}^{(0)} \right| &\leq \sup_{\|\alpha - \alpha_i\| \leq \delta h} |\mu(\alpha, \lambda_t) - \mu(\alpha_i, \lambda_t)| \\ &\leq L_0 \delta h \end{aligned}$$

We now consider the term in the second line (12). First consider the numerator. By applying union bound, we have

$$\begin{aligned} &\mathbb{P}_t \left(\max_{i \in [N]} \left| \frac{1}{Nh} \sum_{j:w_{j,t}=0} \mathbf{1}\{\|\alpha_i - \alpha_j\| \leq \delta h\} K(\hat{d}_{i,j}/h) \cdot u_{j,t} \right| > \eta \right) \\ &\leq N \cdot \mathbb{P}_t \left(\left| \frac{1}{Nh} \sum_{j:w_{j,t}=0} \mathbf{1}\{\|\alpha_i - \alpha_j\| \leq \delta h\} K(\hat{d}_{i,j}/h) \cdot u_{j,t} \right| > \eta \right) \end{aligned}$$

By Lemma A.4, there exist constants $0 < \underline{c} < \bar{c} < \infty$ such that, with probability approaching one, for all i $\underline{c}h^{d_\alpha} \leq \frac{1}{N} \sum_{j:w_{j,t}=0} \mathbf{1}\{\|\alpha_i - \alpha_j\| \leq \delta h\} \leq \bar{c}h^{d_\alpha}$, it follows that In this case:

$$\begin{aligned} &\mathbb{P}_t \left(\left| \frac{1}{Nh^{d_\alpha}} \sum_{j:w_{j,t}=0} \mathbf{1}\{\|\alpha_i - \alpha_j\| \leq \delta h\} K(\hat{d}_{i,j}/h) \cdot u_{j,t} \right| > \eta \mid \{\alpha_j, \hat{d}_{i,j}, w_{j,t}\}_{j=1}^N, \lambda_t \right) \\ &\leq \mathbb{P}_t \left(\left| \frac{1}{\sum_{j:w_{j,t}=0} \mathbf{1}\{\|\alpha_i - \alpha_j\| \leq \delta h\}} \sum_{j:w_{j,t}=0, \|\alpha_i - \alpha_j\| \leq \delta h} K(\hat{d}_{i,j}/h) \cdot u_{j,t} \right| > \frac{1}{\bar{c}} \cdot \eta \mid \{\alpha_j, \hat{d}_{i,j}, w_{j,t}\}_{j=1}^N, \lambda_t \right) \end{aligned}$$

First, note that, given $\{\alpha_j, \hat{d}_{i,j}, w_{j,t}\}_{j=1}^N$ and λ_t , the terms $\{K(\hat{d}_{i,j}/h)\}_{j=1}^N$ are constants, and these are bounded above by Assumption 4 (v). By Assumptions 1 and 4 (ii), $\{u_{j,t}\}_{j=1}^N$ are sub-Gaussian, zero-mean, and independent random variables conditional on $\{\alpha_j, \hat{d}_{i,j}, w_{j,t}\}_{j=1}^N$ and λ_t . Thus, from the Bernstein inequality in Lemma A.7, it follows that for constants C, a , and b , for any

$\eta > 0$,

$$\begin{aligned}
& \mathbb{P}_t \left(\left| \frac{1}{\sum_{j; w_{j,t}=0} \mathbf{1} \{ \|\alpha_i - \alpha_j\| \leq \delta h \}} \sum_{j; w_{j,t}=0, \|\alpha_i - \alpha_j\| \leq \delta h} K(\hat{d}_{i,j}/h) \cdot u_{j,t} \right| > \frac{1}{\bar{c}} \cdot \eta \mid \left\{ \alpha_j, \hat{d}_{i,j}, w_{j,t} \right\}_{j=1}^N, \lambda_t \right) \\
& \leq C \cdot \exp \left(- \frac{\sum_j \mathbf{1} \{ \|\alpha_i - \alpha_j\| \leq \delta h \} \cdot \bar{c}^{-2} \eta^2}{a + b\eta} \right) \\
& \leq C \cdot \exp \left(\log(N) - \frac{c N h^{d_\alpha} \bar{c}^{-2} \eta^2}{a + b\eta} \right)
\end{aligned}$$

where the last line again uses that $\underline{c} h^{d_\alpha} \leq \frac{1}{N} \sum_{j; w_{j,t}=0} \mathbf{1} \{ \|\alpha_i - \alpha_j\| \leq \delta h \} \leq \bar{c} h^{d_\alpha}$. Then by the law of iterated expectations, we have

$$N \cdot \mathbb{P}_t \left(\left| \frac{1}{N h^{d_\alpha}} \sum_{j; w_{j,t}=0} \mathbf{1} \{ \|\alpha_i - \alpha_j\| \leq \delta h \} K(\hat{d}_{i,j}/h) \cdot u_{j,t} \right| > \eta \right) \leq C \cdot \exp \left(\log(N) - \frac{c N h^{d_\alpha} \bar{c}^{-2} \eta^2}{a + b\eta} \right)$$

By setting $\eta = \bar{c} \sqrt{\frac{a \log(N)}{c h^{d_\alpha} N}}$ in the above inequality, it gives that

$$\begin{aligned}
& N \cdot \mathbb{P}_t \left(\left| \frac{1}{N h^{d_\alpha}} \sum_{j; w_{j,t}=0} \mathbf{1} \{ \|\alpha_i - \alpha_j\| \leq \delta h \} K(\hat{d}_{i,j}/h) \cdot u_{j,t} \right| > \bar{c} \sqrt{\frac{a \log(N)}{c h^{d_\alpha} N}} \right) \\
& \leq C \cdot \exp \left(\log(N) - \frac{a \log(N)}{a + b\eta} \right) \rightarrow C
\end{aligned}$$

and leads to

$$\begin{aligned}
& \mathbb{P}_t \left(\max_{i \in [N]} \left| \frac{1}{N h^{d_\alpha}} \sum_{j; w_{j,t}=0} \mathbf{1} \{ \|\alpha_i - \alpha_j\| \leq \delta h \} K(\hat{d}_{i,j}/h) \cdot u_{j,t} \right| > \bar{c} \sqrt{\frac{a \log(N)}{c h^{d_\alpha} N}} \right) \\
& \leq C + o(1)
\end{aligned}$$

Finally, we consider the denominator term $\frac{1}{N h^{d_\alpha}} \sum_{j; w_{j,t}=0} K(\hat{d}_{i,j}/h)$. By adding and subtracting terms, we have

$$\begin{aligned}
& \frac{1}{N h^{d_\alpha}} \sum_{j; w_{j,t}=0} \mathbf{1} \{ \|\alpha_i - \alpha_j\| \leq \delta h \} K(\hat{d}_{i,j}/h) \\
& = \frac{1}{N h^{d_\alpha}} \sum_{j; w_{j,t}=0} \mathbf{1} \{ \|\alpha_i - \alpha_j\| \leq \delta h \} K(d_{i,j}/h) \\
& + \frac{1}{N h^{d_\alpha}} \sum_{j; w_{j,t}=0} \mathbf{1} \{ \|\alpha_i - \alpha_j\| \leq \delta h \} [K(\hat{d}_{i,j}/h) - K(d_{i,j}/h)]
\end{aligned}$$

Using that the kernel is Lipschitz by Assumption 4 (v), we have

$$\begin{aligned}
& \max_{i \in [N]} \left| \frac{1}{N h^{d_\alpha}} \sum_{j; w_{j,t}=0} \mathbf{1} \{ \|\alpha_i - \alpha_j\| \leq \delta h \} [K(\hat{d}_{i,j}/h) - K(d_{i,j}/h)] \right| \\
& \leq \bar{K}' \frac{\max_{j \neq i} |\hat{d}_{i,j} - d_{i,j}|}{h} \max_{i \in [N]} \frac{1}{N h^{d_\alpha}} \sum_{j; w_{j,t}=0} \mathbf{1} \{ \|\alpha_i - \alpha_j\| \leq \delta h \}
\end{aligned}$$

By Lemma A.1, $\frac{\max_{j \neq i} |\hat{d}_{i,j} - d_{i,j}|}{h} = o_p(1)$ and by Lemma A.4 $\max_{i \in [N]} \frac{1}{Nh^{d_\alpha}} \sum_{j; w_{j,i}=0} \mathbf{1} \{ \|\alpha_i - \alpha_j\| \leq \delta h \} = O_p(1)$, it follows that

$$\begin{aligned} & \min_{i \in [N]} \frac{1}{Nh^{d_\alpha}} \sum_{j; w_{j,i}=0} \mathbf{1} \{ \|\alpha_i - \alpha_j\| \leq \delta h \} K(\hat{d}_{i,j}/h) \\ &= \min_{i \in [N]} \frac{1}{Nh^{d_\alpha}} \sum_{j; w_{j,i}=0} \mathbf{1} \{ \|\alpha_i - \alpha_j\| \leq \delta h \} K(d_{i,j}/h) + o_p(1) \end{aligned}$$

Finally, by Lemma A.4, there exist a $\tilde{c} > 0$ such that, with probability approaching one,

$$\min_{i \in [N]} \frac{1}{Nh^{d_\alpha}} \sum_{j; w_{j,i}=0} \mathbf{1} \{ \|\alpha_i - \alpha_j\| \leq \delta h \} K(d_{i,j}/h) = \min_{i \in [N]} \frac{1}{Nh^{d_\alpha}} \sum_{j; w_{j,i}=0} K(d_{i,j}/h) > \tilde{c}$$

Therefore, we conclude that

$$\max_{i \in [N]} |\hat{\mu}_{i,t}^{(0)} - \mu_{i,t}^{(0)}| = O_p \left(h + \sqrt{\frac{\log(N)}{h^{d_\alpha} N}} \right)$$

and finish the proof. $\square$

A.3 Bernstein Inequalities

In this section, we specify the Bernstein inequalities in Lemma A.5 to A.7 that are extensively used throughout the proofs.

Lemma A.5 (Bernstein inequality; see [Bennett \(1962\)](#)). *Let $Z_1, \dots, Z_n$ be mean zero independent random variables. Assume there exists a positive constant M such that $|Z_i| \leq M$ with probability one for each i . Also let $\sigma^2 := \frac{1}{n} \sum_{i=1}^n \mathbb{E}[Z_i^2]$. Then for all $\epsilon > 0$*

$$\mathbb{P} \left(\left| \frac{1}{n} \sum_{i=1}^n Z_i \right| \geq \epsilon \right) \leq 2 \exp \left(- \frac{n\epsilon^2}{2(\sigma^2 + \frac{1}{3}M\epsilon)} \right).$$

Lemma A.6 (Bernstein inequality for unbounded random variables; see [Boucheron et al. \(2013\)](#)). *Let $Z_1, \dots, Z_n$ be independent random variables. Assume that there exist some positive constants ν and c such that $\frac{1}{n} \sum_{i=1}^n \mathbb{E}[Z_i^2] \leq \nu$ such that for all integers $q \geq 3$,*

$$\frac{1}{n} \sum_{i=1}^n \mathbb{E}[|Z_i|^q] \leq \frac{q!c^{q-2}}{2} \nu.$$

Then for all $\epsilon > 0$,

$$\mathbb{P} \left(\left| \frac{1}{n} \sum_{i=1}^n (Z_i - \mathbb{E}[Z_i]) \right| \geq \epsilon \right) \leq 2 \exp \left(- \frac{n\epsilon^2}{2(\nu + c\epsilon)} \right).$$

Proof of Lemma A.6. See Theorem 2.1 in [Boucheron et al. \(2013\)](#) for detailed discussion. $\square$

Lemma A.7 (Bernstein inequality for sub-exponential random variables; see [Zeleneev \(2020\)](#)).
Let $Z_1, \dots, Z_n$ be mean zero independent random variables. Assume that there exist some $\nu > 0$ such that $\mathbb{E}[\exp(\lambda Z_i)] \leq \exp(\nu \lambda^2)$ for all $\lambda \in \mathbb{R}$. Then, there exist some positive constants C , a , and b such that for all constants $\alpha_1, \dots, \alpha_n$ satisfying $\max_i |\alpha_i| < \bar{\alpha}$ and for all $\epsilon > 0$,

$$\mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n \alpha_i Z_i\right| \geq \epsilon\right) \leq C \exp\left(-\frac{n\epsilon^2}{a + b\epsilon}\right)$$

and

$$\mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n \alpha_i (Z_i^2 - \mathbb{E}[Z_i^2])\right| \geq \epsilon\right) \leq C \exp\left(-\frac{n\epsilon^2}{a + b\epsilon}\right).$$

Proof of Lemma A.7. See Corollary A.1 in [Zeleneev \(2020\)](#) for detailed discussion. $\square$