

Simple Specification-Free Causal Inference in Large Panels*

Ben Deaner[†]

Chen-Wei Hsiang[†]

Andrei Zeleneev[†]

This version: September 23, 2026

First Circulated: January 31, 2025

Abstract

We propose a simple and transparent approach to causal inference using panel data with a large number of units and periods. Instead of imposing a particular parametric specification on the untreated potential outcomes, we assume they follow a nonparametric nonseparable factor model. We use long pre-treatment outcome histories to construct a measure of similarity between the latent factors of different units. Using this measure, we impute missing counterfactual means and propensity scores through kernel smoothing. Under weak smoothness conditions, these estimates can attain the optimal nonparametric convergence rate up to logarithmic factors, provided that the pre-treatment history grows sufficiently quickly with the sample size. We construct a doubly robust estimator of the period-specific average treatment effect on the treated (ATT) and provide conditions under which it is $\sqrt{N}$ -consistent and asymptotically normal, centered at the true ATT. Simulations demonstrate accurate inference for a wide range of linear and nonlinear data-generating processes.

Keywords: treatment effects, panel data, nonlinear factor models, doubly robust estimation

*This paper was previously circulated under the title “Inferring Treatment Effects in Large Panels by Uncovering Latent Similarities”. The first version of this paper was prepared for the 2025 ESWC submission on January 31, 2025, and is also available [here](#). We thank Alberto Abadie, Timothy B. Armstrong, Andrew Chesher, Timothy Christensen, Guido Imbens, Aureo de Paula, Kirill Ponomarev, Liyang Sun, and participants in numerous seminars and conferences for their valuable comments and suggestions.

[†]Department of Economics, University College London, and CeMMAP.

1 Introduction

The presence of unobserved confounding factors presents a major challenge for causal inference. The availability of rich panel data can help address this problem. Panel datasets contain multiple observations for each individual and can thus provide a means of controlling for time-invariant confounders (fixed effects). In addition, panel data can enable the imputation of individual counterfactual outcomes. Methods for causal inference with panel data are widely used in empirical work, and their development is an active and rapidly growing area of research (see, e.g., [Arkhangelsky and Imbens, 2024](#)).

Many popular methods for causal inference using panel data assume that untreated potential outcomes have a linear or affine structure. For example, difference-in-differences (DiD) and two-way fixed effects (TWFE) methods rely on parallel-trends assumptions, whereas synthetic control (SC) and most matrix-completion methods require untreated potential outcomes to follow a linear factor model. While such restrictions can simplify estimation and inference, and may be necessary given limited data availability, it is natural to ask whether they are fundamental for identifying treatment effects in large panels and, most importantly, whether they can be relaxed while retaining a theoretically disciplined and practical approach that performs well in empirically relevant sample sizes.

This paper addresses this question by proposing a new method for estimating and conducting inference on the causal effects of a binary treatment $w_{i,t}$ in large panels. We assume that the untreated potential outcome of unit i in period t follows the possibly nonlinear and nonseparable 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 We assume that, conditional on the latent factors, untreated potential outcomes are independent of treatment assignment. Importantly, we do not assume that $\mu^{(0)}(\cdot, \cdot)$ has any particular parametric form. Thus, model (1) is substantially more general than the commonly used linear formulations $\mu^{(0)}(\alpha_i, \lambda_t) = \alpha_i + \lambda_t$ and $\mu^{(0)}(\alpha_i, \lambda_t) = \alpha'_i \lambda_t$ imposed by TWFE and SC methods, respectively.

While the generality of formulation (1) is appealing, it complicates the identification and estimation of treatment effects because α_i cannot be identified without strong additional restrictions imposed on the form of $\mu^{(0)}(\cdot, \cdot)$. Rather than attempting to model $\mu^{(0)}(\cdot, \cdot)$ and es-

¹It is also conceptually straightforward to allow for a linear adjustment for observed covariates in (1). We abstract from this extension for clarity of exposition.

timate α_i directly, we adapt an approach first developed by [Zhang et al. \(2017\)](#) for graphon estimation in network data. Specifically, we use a long history of pre-treatment outcomes to construct a pseudo-distance $\hat{d}_{i,j}$ that is informative about the similarity between the latent characteristics α_i and α_j of units i and j . Using this pseudo-distance, we impute the *values* of the conditional means of untreated potential outcomes and the propensity scores (both conditional on the latent factors) in a given post-treatment period t for *all units* through kernel smoothing. Notably, we can do so even though α_i is *neither observed nor identified* in our setting.

We construct estimators of the untreated conditional means $\mu_{i,t}^{(0)}$ and the propensity scores $p_{i,t}$, and establish their rates of convergence (uniformly over i) as both the number of units N and the number of pre-treatment periods T_0 go to infinity. If T_0 grows sufficiently quickly relative to N , these estimators attain the optimal nonparametric rate of convergence established by [Stone \(1980\)](#), up to a logarithmic factor, as if the latent characteristics α_i were observed.

Even at the optimal rate, however, these nonparametric estimators generally converge too slowly for conventional plug-in inference on the average treatment effect on the treated (ATT). We therefore combine them in a doubly robust estimator of the ATT for each post-treatment period t (e.g., [Robins et al., 1994](#); [Chernozhukov et al., 2018](#)). We propose a cross-fitting scheme that, together with double robustness, helps ensure that the resulting estimator is $\sqrt{N}$ -consistent, asymptotically normal, and centered at the true ATT. Remarkably, under certain regularity conditions, and if T_0 grows sufficiently fast, our estimator attains the semiparametric efficiency bound that would apply if the latent characteristics α_i were observed.

Compared to some competing approaches, an advantage of our method is its simplicity and transparency. We avoid the need to decompose an incomplete matrix of potential outcomes or to regularize the components of that decomposition. Instead, we impute untreated potential outcomes for treated units by averaging the outcomes of untreated units, placing greater weight on those with similar pre-treatment outcome histories. Researchers can readily inspect the pseudo-distances and corresponding kernel weights to see which comparison units drive each imputation.

We provide extensive simulation evidence that our procedure performs well across a broad range of linear and nonlinear data-generating processes. When the additive TWFE model is correctly specified, our estimator delivers confidence-interval coverage comparable to that of TWFE. Under interactive and nonlinear factor structures, it continues to provide accurate coverage, whereas TWFE can perform poorly. Our method is also competitive

with—and often outperforms—the local PCA approach of [Feng \(2026\)](#) in small and moderate panels, particularly when the pre-treatment history is relatively short. Finally, in an empirical illustration based on the Walmart sales panel, we conduct empirically calibrated Monte Carlo exercises in which our estimator performs nearly as well as an infeasible benchmark that uses the true latent characteristics. Taken together, our theoretical and numerical results show that relaxing linear factor restrictions in large panels need not come at the expense of reliable inference in empirically relevant sample sizes.

To assist empirical practitioners who wish to apply our method, we provide an R implementation at <https://github.com/chenweihsiang/PanelLatSim>.

Related literature 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 are widely used and continue to grow in popularity in applied work, perhaps because of their practicality and perceived transparency ([Goldsmith-Pinkham, 2024](#)); see [de Chaisemartin and d’Haultfoeuille \(2023\)](#) and [Roth et al. \(2023\)](#) for extensive reviews of the DiD and TWFE literature. A major focus of recent econometric work on these methods has been accommodating heterogeneous treatment effects (e.g., [de Chaisemartin and d’Haultfoeuille, 2020](#); [Callaway and Sant’Anna, 2021](#); [Sun and Abraham, 2021](#); [Wooldridge, 2025](#); [Borusyak et al., 2024](#)). However, their validity still relies crucially on parallel-trends assumptions. Several recent papers have re-examined the restrictiveness of these assumptions and, in some cases, proposed relaxations at the cost of giving up point identification of the causal parameters of interest (e.g., [Manski and Pepper, 2018](#); [Rambachan and Roth, 2023](#); [Ghanem et al., 2022](#)). By contrast, our framework allows for rich heterogeneity in treatment effects and selection mechanisms while retaining point identification without imposing parallel trends. To achieve this, however, we require a sufficiently long pre-treatment history.

Alternative popular approaches to causal inference in panels include synthetic control (SC) and matrix-completion methods; see [Abadie \(2021\)](#) for a recent review of SC methods. Recent contributions to the SC and causal matrix-completion literatures include, among others, [Arkhangelsky et al. \(2021\)](#); [Cattaneo et al. \(2021\)](#); [Chernozhukov et al. \(2021, 2026\)](#) and [Athey et al. \(2021\)](#); [Bai and Ng \(2021\)](#); [Agarwal et al. \(2023\)](#); [Xiong and Pelger \(2023\)](#); [Choi and Yuan \(2024\)](#); [Abadie et al. \(2024\)](#), respectively. 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 potential outcomes to follow a factor

model. As in this paper, these methods use a long history of pre-treatment outcomes to impute missing counterfactual means. Most of these methods, however, rely on a linear factor model, $\mu^{(0)}(\alpha_i, \lambda_t) = \alpha_i' \lambda_t$, to establish statistical guarantees and justify inference, whereas our framework allows for a general nonparametric and nonlinear factor structure.

Some recent matrix-completion methods also accommodate nonlinear factor structures of the type considered in this paper by approximating the corresponding matrix of counterfactual means with a low-rank matrix. The resulting approximation error is typically controlled by imposing relatively strong smoothness restrictions on $\mu^{(0)}(\cdot, \cdot)$, such as requiring it to belong to a sufficiently smooth Hölder class (e.g., [Fernández-Val et al., 2021](#); [Choi et al., 2024](#); [Abadie et al., 2025](#)).² By contrast, our imputation method does not rely on a low-rank approximation of $\mu^{(0)}(\cdot, \cdot)$. Under the substantially weaker smoothness assumption requiring Lipschitz continuity only in the first argument of $\mu^{(0)}(\cdot, \cdot)$, our imputation method achieves a nearly optimal nonparametric rate of convergence (given the imposed smoothness conditions) and allows us to construct a semiparametrically efficient estimator of the ATT.³

Our imputation procedure builds on the pseudo-distance proposed by [Zhang et al. \(2017\)](#), which allows us to find units with similar latent characteristics even though those characteristics are neither observed nor identified. Variants of this pseudo-distance have been employed for nonparametric graphon estimation ([Zhang et al., 2017](#); [Zeleneev, 2026](#); [Nowakowicz, 2024](#); [Sun and Zhang, 2026](#)), controlling for unobservables using network data ([Auerbach, 2022](#)), and estimation of nonlinear factor models ([Feng, 2024](#)). Related distance-based methods have also been applied to treatment-effect estimation in network and panel settings by [Hoshino and Yanagi \(2024\)](#) and [Athey and Imbens \(2025\)](#). Unlike this paper, however, these studies neither establish asymptotic normality of their treatment-effect estimators nor provide asymptotically valid confidence intervals.⁴

Imputation of the propensity scores alongside the counterfactual means is an important and distinctive ingredient of our approach that allows us to estimate the ATT using doubly robust moments (e.g., [Robins et al., 1994](#); [Robins and Rotnitzky, 1995](#); [Hahn, 1998](#); [Scharfstein](#)

²Among these papers, [Choi et al. \(2024\)](#) develops an inference procedure for average treatment effects, whereas [Fernández-Val et al. \(2021\)](#) and [Abadie et al. \(2025\)](#) establish estimation results but do not provide corresponding inference procedures.

³[Dwivedi et al. \(2025\)](#) provide entrywise non-asymptotic error bounds for a nonlinear factor model under a similar Lipschitz-continuity assumption. Their imputation method, however, relies crucially on homoskedastic idiosyncratic errors. Moreover, they do not investigate whether their estimator achieves the optimal nonparametric rate of convergence and do not provide semiparametrically efficient estimators of average treatment effects.

⁴In particular, [Athey and Imbens \(2025\)](#) focus on identification of the counterfactual means and do not provide asymptotic theory for their treatment effect estimator. [Hoshino and Yanagi \(2024\)](#) provide error bounds for various average dyad treatment effects but do not establish their asymptotic distribution. None of these papers consider treatment effect estimation in a doubly robust form.

et al., 1999; Chernozhukov et al., 2018, 2022). We demonstrate that, under suitable conditions, the resulting doubly robust estimator is $\sqrt{N}$ -consistent, asymptotically normal, and centered at the true ATT, whereas conventional imputation estimators in the matrix-completion literature generally inherit the slower nonparametric convergence rate of the estimated counterfactual means.

Although propensity scores are routinely estimated and used in doubly robust moments to adjust for *observed* covariates in panel settings (e.g., Sant’Anna and Zhao, 2020; Arkhangelsky and Imbens, 2022; Chen et al., 2025), this strategy remains underexplored in latent factor models, where standard matrix-completion methods focus almost exclusively on imputing counterfactual means. A notable recent exception is Abadie et al. (2024), who estimate propensity scores by applying matrix-completion methods to the observed treatment-assignment matrix. However, our propensity-score imputation approach is fundamentally different from theirs: we use pre-treatment outcome histories rather than rely on a low-rank approximation of the propensity-score matrix.⁵

Finally, Feng (2026) studies causal inference with unobserved confounders when a large number of auxiliary “proxies” for those confounders are available; in panel settings, pre-treatment outcomes can serve as such proxies. As in this paper, Feng (2026) estimates counterfactual means and propensity scores and combines them in doubly robust estimators of causal effects. Their imputation procedure, however, is quite different from ours. Feng (2026) first constructs K -nearest-neighbor neighborhoods using the pseudo-distance of Zhang et al. (2017) or another distance, applies local principal component analysis (PCA) to extract factor loadings within each neighborhood, and then uses the estimated loadings as regressors in local quasi-maximum-likelihood estimation (QMLE). Feng (2026) establishes asymptotic results under higher-order smoothness conditions. Our method instead directly estimates the counterfactual means and propensity scores using kernel smoothing based on the pseudo-distance, avoiding local PCA, the selection of the number of components, and the subsequent QMLE step. A benefit of the resulting procedure is that it is simple and transparent, with counterfactual means and propensity scores expressed directly as smoothly weighted averages of observed outcomes and treatments. The simplicity of the procedure and our focus on panel data also carry over to the theoretical analysis. Under primitive conditions on the nonlinear factor structure and conditions on the serial dependence structure of the panel data, we establish nearly optimal uniform convergence rates for the estimated con-

⁵Outside of the panel causal inference literature, recent applications of doubly robust estimation in linear panel regression models with unobserved heterogeneity having a nonlinear factor form also include Beyhum and Mugnier (2024) and Freeman and Kristensen (2026).

ditional means and propensity scores under Lipschitz continuity of the model, as opposed to higher-order smoothness.⁶ We show that under our primitive conditions, the resulting ATT estimator is $\sqrt{N}$ -consistent and asymptotically normal. Our simulation results suggest that the simplicity of our method does not come at the expense of finite-sample performance: our estimator is competitive with local PCA across the designs considered and often delivers coverage closer to nominal in both small and moderate panels, particularly when the pre-treatment history is relatively short.⁷

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 simulation evidence, and Section 5 provides an empirical illustration based on Monte Carlo exercises calibrated to the Walmart sales panel. All auxiliary lemmas and proofs are provided in Appendix A.

2 Setup and Proposed Method

2.1 Setup

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.⁸ We assume throughout that no individual in the population is treated prior to a period $T_0 + 1$ with $T_0 < T$. 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 no individual is ever treated. We assume that if $w_{i,t} = 0$ then $Y_{i,t} = Y_{i,t}(0)$. This precludes both anticipation and spillover effects from treated to untreated units (see, e.g., Abbring and van den Berg, 2003; Borusyak et al., 2024; Sun and Abraham, 2021). It does

⁶While our imputation method achieves the nearly optimal rate of convergence under Lipschitz continuity, it does not take advantage of the potential higher-order smoothness, whereas the approach of Feng (2026) can achieve faster rates of convergence in the latter case.

⁷We note that Feng (2026) also discusses a simpler estimator that extracts a single principal component, which is described as analogous to local-constant smoothing. However, the one-component procedure still uses local PCA and QMLE and therefore does not generally reduce to conventional nearest-neighbor averaging.

⁸For concreteness, we assume that treatment is an absorbing state, which is typical for staggered treatment adoption applications. Our approach also accommodates treatment switching by restricting the period- t comparison to currently treated and never-treated units, provided that the maintained assumptions hold for this comparison sample. In this extension, the equality in Assumption 1 is required for all units during the first T_0 periods and for never-treated controls thereafter.

not rule out spillover effects between treated units. We state this formally in Assumption 1 below.

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

Central to our analysis is the assumption of a nonlinear 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. These time-specific factors may be *non-stationary* (distributed differently over time) and in particular, their distribution in the post-treatment period may *differ* entirely from that in the pre-treatment period. We model the untreated 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 $\mu^{(0)}(\cdot, \cdot)$ and $p_t(\cdot)$ are unknown, and, importantly, we do not assume that they have any particular functional form. Note that the model above implies that the conditional mean function $\mu^{(0)}(\cdot, \cdot)$ is the same for every individual and every period. However, the conditional mean $\mu^{(0)}(\alpha_i, \lambda_t)$ may vary with both i and t through the latent factors. Notice that $p_t(\cdot)$, the conditional mean function of $w_{i,t}$, is allowed to vary over time, which reflects the fact that this variable is increasing and therefore subject to secular trends that need not be captured in λ_t .

Remark 1 (Nonseparable models). The additively separable models (2) and (3) may be derived from non-separable models under primitive conditions. For example, suppose that $Y_{i,t}(0) = q^{(0)}(\alpha_i, \lambda_t, v_{i,t})$ for some function $q^{(0)}(\cdot)$ and random variable $v_{i,t}$ that is independent of α_i and λ_t . If $v_{i,t}$ is exchangeable (separately in both i and t), then (2) holds with

$$\mu^{(0)}(\alpha_i, \lambda_t) := \mathbb{E}[q^{(0)}(\alpha_i, \lambda_t, v_{i,t})|\alpha_i, \lambda_t],$$

and the residual $u_{i,t}$ is mean-independent of α_i and λ_t by construction.

Without additional assumptions, causal quantities of interest are not identified from the model above. This would remain true even if the latent factors α_i and λ_t were observed. 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 factors that influence both the outcome

and treatment status (that is, confounding factors) and which are included in neither α_i nor λ_t . 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., 2023](#); [Arkhangelsky and Imbens, 2022](#); [Athey et al., 2021](#); [Fernández-Val et al., 2021](#)). Note that, since, in the spirit of the DiD and TWFE methods, we focus on estimating average treatment effects for the treated units, Assumption 2 does not place restrictions on the treated potential outcomes. If we additionally require the treated potential outcomes to be conditionally independent of treatment, our method straightforwardly extends to cover other counterfactuals including average treatment effects (ATEs).

Under Assumption 2, key counterfactual quantities of interest can be written in terms of the latent factors and functions $\mu^{(0)}(\cdot, \cdot)$ and $p_t(\cdot)$. 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 and a support overlap condition, the average effect of treatment on the treated at time $t \geq T_0 + 1$, which is defined as $ATT_t := \mathbb{E}_t[Y_{i,t} - Y_{i,t}(0) | w_{i,t} = 1]$, can be expressed in the following doubly robust form:

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

The time subscript indicates that each expectation is taken with respect to the period t -specific distribution of the observables (i.e., conditional on λ_t). Related doubly robust approaches to estimating ATT-type parameters in panel and DiD settings are studied by, among others, [Sant’Anna and Zhao \(2020\)](#); [Arkhangelsky and Imbens \(2022\)](#); [Chen et al. \(2025\)](#).

Remark 2. For clarity and concreteness, we focus on the period-specific average treatment effect on the treated ATT_t . Under staggered treatment adoption, this parameter coincides with the calendar-time average treatment effect considered by [Callaway and Sant’Anna \(2021\)](#). In particular, letting G_i denote the period in which unit i is first treated and defining

$$ATT(g, t) := \mathbb{E}_t [Y_{i,t} - Y_{i,t}(0) | G_i = g],$$

we have

$$ATT_t = \sum_{g \leq t} \mathbb{P}_t(G_i = g | G_i \leq t) ATT(g, t).$$

Our analysis can accommodate other standard ATT-type estimands analogously: one can estimate the cohort-time effects $\text{ATT}(g, t)$ using the corresponding doubly robust moments and aggregate them with appropriate weights to obtain event-time, cohort-specific, or overall treatment effects. However, in order to achieve identification in this case, Assumption 2 must be strengthened slightly to $Y_{i,t}(0) \perp\!\!\!\perp G_i | \alpha_i, \lambda_t$ and one must impose a suitable support overlap condition.

In this paper, we provide new methods for estimating the *values* of $p_{i,t}$ 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 takes the form

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

where $N_{1,t}$ is the number of individuals treated by time t .

The arguments we use to consistently estimate $p_{i,t}$ and $\mu_{i,t}^{(0)}$ require an additional condition. Consider that exposure to aggregate/macroeconomic conditions may induce dependence between the potential outcomes of different individuals. We assume, in effect, that the time-specific factors λ_t capture the influence of all such factors on untreated potential outcomes. The condition also imposes that at least in the absence of treatment, the outcomes and latent factors of an individual j cannot impact the untreated outcomes of any distinct individual i .

Assumption 3 (Conditional cross-sectional independence). For any two individuals $i \neq j$, we have $Y_{i,t}(0) \perp\!\!\!\perp Y_{j,t}(0) | \alpha_i, \alpha_j, \lambda_t$ and $Y_{i,t}(0) \perp\!\!\!\perp \alpha_j | \alpha_i, \lambda_t$.

Assumption 3 is crucial to our arguments. The condition ensures that for any two distinct individuals i and j ,

$$\mathbb{E}[\langle Y_{i,1:T_0}, Y_{j,1:T_0} \rangle | \alpha_i, \alpha_j] = \mathbb{E}[\langle \mu_{i,1:T_0}^{(0)}, \mu_{j,1:T_0}^{(0)} \rangle | \alpha_i, \alpha_j],$$

where $\langle \cdot, \cdot \rangle$ is the Euclidean inner product (i.e., $\langle a, b \rangle = a'b$ for two column vectors a and b of the same length). Thus, this assumption allows us to relate observable quantities, i.e., the history of pre-treatment outcomes, to the history of $\mu_{i,t}^{(0)}$, and thus to the individual-specific latent factors.

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

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

The objects on the right-hand sides of each equation above are regression functions. If α_i were observable, we could nonparametrically estimate $\mu^{(0)}(\cdot, \lambda_t)$ and $p_t(\cdot)$ (as functions of α), and thus also $p_{i,t}$ and $\mu_{i,t}^{(0)}$. For example, we could obtain the following Nadaraya-Watson estimates:

$$\hat{\mu}_{i,t}^{(0)} = \frac{\sum_{j:w_{j,t}=0} K(\|\alpha_i - \alpha_j\|/h) Y_{j,t}}{\sum_{j:w_{j,t}=0} K(\|\alpha_i - \alpha_j\|/h)}, \quad \hat{p}_{i,t} = \frac{\sum_j K(\|\alpha_i - \alpha_j\|/h) w_{j,t}}{\sum_j K(\|\alpha_i - \alpha_j\|/h)}, \quad (5)$$

where $K(\cdot)$ and h are some kernel and bandwidth.

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 the following feasible pseudo-distance:

$$\hat{d}_{i,j} = \frac{1}{T_0} \max_{k \notin \{i,j\}} |\langle Y_{k,1:T_0}, Y_{i,1:T_0} \rangle - \langle Y_{k,1:T_0}, Y_{j,1:T_0} \rangle|. \quad (6)$$

A pseudo-distance of the form above was first employed in [Zhang et al. \(2017\)](#) for nonparametric graphon estimation.

To motivate the use of this pseudo-distance, suppose Assumption 3 holds and that the aggregate factors λ_t in each period $t \leq T_0$ are jointly independent of the individual-specific factors. Then the difference in conditional expected values of the inner-products in the pseudo-distance can be expressed as

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

where π_t denotes the (unconditional) distribution of λ_t , which may vary with t , i.e., λ_t does *not* need to be stationary. As such, we can understand the pseudo-distance as the sample analogue of the population pseudo-distance

$$\begin{aligned} d_{i,j} &= \sup_{\alpha \in \mathcal{A}} \left| \frac{1}{T_0} \sum_{t=1}^{T_0} \int \mu^{(0)}(\alpha, \lambda) (\mu^{(0)}(\alpha_i, \lambda) - \mu^{(0)}(\alpha_j, \lambda)) d\pi_t(\lambda) \right| \\ &= \sup_{\alpha \in \mathcal{A}} \left| \int \mu^{(0)}(\alpha, \lambda) (\mu^{(0)}(\alpha_i, \lambda) - \mu^{(0)}(\alpha_j, \lambda)) d\bar{\pi}_{T_0}(\lambda) \right| \\ &\geq \frac{1}{2} \int (\mu^{(0)}(\alpha_i, \lambda) - \mu^{(0)}(\alpha_j, \lambda))^2 d\bar{\pi}_{T_0}(\lambda), \end{aligned}$$

where $\mathcal{A}$ is the support of α_i , and $\bar{\pi}_{T_0} := \frac{1}{T_0} \sum_{t=1}^{T_0} \pi_t$ is the average of the probability laws of λ_t over the pre-treatment period; notice that we suppress the dependence of $d_{i,j}$ on T_0 for

notational convenience.⁹ The inequality above relates the size of the population pseudo-distance $d_{i,j}$ to a squared L^2 distance (with respect to the measure $\bar{\pi}_{T_0}$) between the functions $\mu^{(0)}(\alpha_i, \cdot)$ and $\mu^{(0)}(\alpha_j, \cdot)$.¹⁰ Thus $d_{i,j}$ measures similarity of the latent factors to the extent that they impact untreated outcomes. Finally, while the population pseudo-distance $d_{i,j}$ is not directly observed, it can be consistently estimated by the sample pseudo-distance $\hat{d}_{i,j}$. This means that $\hat{d}_{i,j}$ also effectively measures the similarity between the functions $\mu^{(0)}(\alpha_i, \cdot)$ and $\mu^{(0)}(\alpha_j, \cdot)$, which, under an additional injectivity condition, also translates into the similarity between α_i and α_j .

Remark 3. It is worth contrasting the sample pseudo-distance above with the (normalized) Euclidean distance between the histories of pre-treatment outcomes, i.e., $\frac{1}{T_0} \|Y_{i,1:T_0} - Y_{j,1:T_0}\|^2$. Under the previously employed conditional independence assumption, its population analogue can be expressed as

$$\begin{aligned} \mathbb{E} \left[\frac{1}{T_0} \|Y_{i,1:T_0} - Y_{j,1:T_0}\|^2 \mid \alpha_i, \alpha_j \right] &= \int \left(\mu^{(0)}(\alpha_i, \lambda) - \mu^{(0)}(\alpha_j, \lambda) \right)^2 d\bar{\pi}_{T_0}(\lambda) \\ &+ \frac{1}{T_0} \sum_{s=1}^{T_0} \int \left(\mathbb{E}[u_{i,s}^2 \mid \alpha_i, \lambda_s = \lambda] + \mathbb{E}[u_{j,s}^2 \mid \alpha_j, \lambda_s = \lambda] \right) d\pi_s(\lambda). \end{aligned}$$

The above demonstrates that the (population) Euclidean distance is increasing in the conditional residual variance $\mathbb{E}[u_{j,t}^2 \mid \alpha_j, \lambda_t = \lambda]$. Therefore, in the presence of conditional heteroskedasticity, a small Euclidean distance between a fixed unit i and unit j might not necessarily indicate that α_j is similar to α_i : instead, it may simply reflect that unit j 's outcomes have a low residual variance. It is the need to be robust to conditional heteroskedasticity that motivates our use of the pseudo-distance (6).

Given the constructed pseudo-distance $\hat{d}_{ij}$, we obtain feasible estimates of $p_{i,t}$ and $\mu_{i,t}^{(0)}$:

$$\hat{\mu}_{i,t}^{(0)} = \frac{\sum_{j:w_{j,t}=0} K(\hat{d}_{i,j}/h) Y_{j,t}}{\sum_{j:w_{j,t}=0} K(\hat{d}_{i,j}/h)}, \quad \hat{p}_{i,t} = \frac{\sum_j K(\hat{d}_{i,j}/h) w_{j,t}}{\sum_j K(\hat{d}_{i,j}/h)}. \quad (7)$$

In principle, one could directly plug the estimates above into the formula for the doubly robust ATT_t estimate (4). However, to further de-bias our estimate, we propose additionally combining (4) with a cross-fitting scheme. 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., 2018; Abadie et al., 2024). The

⁹If λ_t is stationary during the pre-treatment period, we trivially have $\pi_t = \pi^*$ and $\bar{\pi}_{T_0} = \pi^*$, where π^* denotes the stationary distribution of λ_t for $t = 1, \dots, T_0$.

¹⁰The inequality follows from taking $\alpha = \alpha_i$ and $\alpha = \alpha_j$ in the supremum and applying the triangle inequality.

full algorithm with cross-fitting is detailed below along with a variance estimate and confidence interval. Our asymptotic results assume a fixed number of folds for cross-fitting and a deterministic bandwidth sequence. However, in practice, we select the bandwidth h by cross-validation.¹¹

Algorithm 1 Doubly Robust Estimation and Inference with Cross-Fitting

Inputs: Number of folds $\mathcal{K}$, bandwidth h , confidence level $1 - \zeta$

Returns: estimate $\hat{\text{ATT}}_t$, asymptotic variance estimate $\hat{V}_t$, and $1 - \zeta$ level confidence interval

- 1: Randomly partition $[N] = \{1, 2, \dots, N\}$ into $\mathcal{K}$ folds $\{\mathcal{I}_k\}_{k=1}^{\mathcal{K}}$ of size $\approx N/\mathcal{K}$. Let $\mathcal{I}_{-k} = [N] \setminus \mathcal{I}_k$
- 2: **for all** $k \in [\mathcal{K}]$ **do**
- 3: **for all** $i \in \mathcal{I}_k$ and $j \in \mathcal{I}_{-k}$ **do**
- 4: Calculate $\hat{d}_{i,j}$, the pseudo-distance between $i \in \mathcal{I}_k$ and $j \in \mathcal{I}_{-k}$, as follows:

$$\hat{d}_{i,j} \leftarrow \frac{1}{T_0} \max_{\ell \notin \{i,j\}; \ell \in \mathcal{I}_{-k}} \left| \langle Y_{\ell,1:T_0}, Y_{i,1:T_0} - Y_{j,1:T_0} \rangle \right|$$

- 5: **end for**
- 6: **end for**
- 7: **for all** $k \in [\mathcal{K}]$ and $i \in \mathcal{I}_k$ **do**
- 8: Calculate $\hat{\mu}_{i,t}^{(0)}$ and $\hat{p}_{i,t}$ as follows:

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

- 9: **end for**
- 10: Construct $\hat{\text{ATT}}_t$ and $\hat{V}_t$ using the formulas below

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

- 11: Construct $1 - \zeta$ level confidence interval $[\hat{\text{ATT}}_t \pm z_{1-\zeta/2} \sqrt{\hat{V}_t/N}]$, where $z_{1-\zeta/2}$ is the $(1 - \zeta/2)$ -th quantile of the standard normal distribution.
-

¹¹Implementation of the proposed method in R is also available at <https://github.com/chenweihsiang/PanellatSim>.

3 Large Sample Theory

The estimator $\hat{\text{ATT}}$, in Algorithm 1 is doubly robust and employs cross-fitting. An extensive literature provides sufficient conditions for $\sqrt{N}$ -consistency of such estimators and asymptotically correct coverage of the corresponding confidence intervals (e.g., [Robins et al., 2008](#); [Chernozhukov et al., 2018](#); [Abadie et al., 2024](#)). A key condition is that the first stage nuisance-parameter estimates (in our case $\hat{\mu}_{i,t}^{(0)}$ and $\hat{p}_{i,t}$) have sufficiently fast rates of convergence. In particular, this requirement is satisfied if

$$\max_{i \in [N]} |\hat{\mu}_{i,t}^{(0)} - \mu_{i,t}^{(0)}| = o_p(N^{-1/4}) \quad \text{and} \quad \max_{i \in [N]} |\hat{p}_{i,t} - p_{i,t}| = o_p(N^{-1/4}). \quad (8)$$

We establish convergence rates for the first-stage estimates. These rates may be of interest per se because $\mu_{i,t}^{(0)}$ is the optimal prediction of individual i 's untreated potential outcome given the latent factors. In particular, we show that, under certain conditions, $\hat{\mu}_{i,t}^{(0)}$ and $\hat{p}_{i,t}$ achieve, up to logarithmic factors, the optimal convergence rate of [Stone \(1980\)](#) for nonparametric regression on α_i under Lipschitz continuity. That is, provided that the number of pre-treatment periods T_0 grows sufficiently quickly with N , our feasible estimators attain, up to logarithmic factors, the optimal pointwise rate available when α_i is observed.

3.1 First-stage rates of convergence

In order to derive convergence rates for our first-stage estimates $\hat{\mu}_{i,t}^{(0)}$ and $\hat{p}_{i,t}$, we impose the following additional assumptions.

Assumption 4 (Model and latent factors). The models (2) and (3) hold and satisfy the following conditions.

- (i) The factors $\alpha_1, \alpha_2, \dots, \alpha_N, \lambda_1, \lambda_2, \dots, \lambda_T$ are jointly independent, and $\alpha_1, \alpha_2, \dots, \alpha_N$ are identically distributed. For each i , α_i has support $\mathcal{A} \subset \mathbb{R}^{d_\alpha}$ and there are constants $0 < \underline{c} < \bar{c} < \infty$ and $\bar{\eta} > 0$ so that for any $0 < \eta \leq \bar{\eta}$ and $\alpha \in \mathcal{A}$, $\underline{c}\eta^{d_\alpha} \leq \mathbb{P}(\|\alpha - \alpha_i\| \leq \eta) \leq \bar{c}\eta^{d_\alpha}$.
- (ii) There is some $\underline{c} > 0$ so that $\underline{c} \leq p_{i,t} \leq 1 - \underline{c}$ almost surely for all i and $t > T_0$.
- (iii) For some $\delta > 0$, $\mathbb{E}[\exp(\eta u_{i,t}) | \{\alpha_j\}_{j=1}^N, \{\lambda_s\}_{s=1}^T, \{w_{j,t}\}_{j=1}^N] \leq \exp(\delta \eta^2)$ for all $\eta \in \mathbb{R}$ almost surely for all i and t . For each i and t , we have $u_{i,t}, \epsilon_{i,t} \perp\!\!\!\perp \{\{u_{j,s}, \epsilon_{j,s}\}_{s=1}^t, \alpha_j\}_{j \neq i}, \{u_{i,s}\}_{s=1}^{t-1}, \{\lambda_s\}_{s \neq t} \mid \alpha_i, \lambda_t$.
- (iv) The functions $\mu^{(0)}(\cdot, \cdot)$ and $p_t(\cdot)$ are uniformly bounded (on the relevant supports) and

satisfy, for some $L_0, L_p < \infty$,

$$\begin{aligned} |\mu^{(0)}(\alpha_1, \lambda) - \mu^{(0)}(\alpha_2, \lambda)| &\leq L_0 \|\alpha_1 - \alpha_2\|, \\ |p_t(\alpha_1) - p_t(\alpha_2)| &\leq L_p \|\alpha_1 - \alpha_2\|, \end{aligned}$$

for all $\alpha_1, \alpha_2 \in \mathcal{A}$, $\lambda \in \mathcal{L}$, and t .¹²

(v) $K : \mathbb{R}_+ \rightarrow \mathbb{R}$ is weakly positive and strictly positive at zero, supported on a compact set and bounded by $\bar{K} < \infty$. Moreover, $|K(z) - K(z')| \leq \bar{K}'|z - z'|$ for all $z, z' \in \mathbb{R}_+$ for some $\bar{K}' > 0$.

(vi) There exist constants $c_1, c_2 > 0$ so that if T_0 is sufficiently large, then for all $\alpha_1, \alpha_2 \in \mathcal{A}$

$$\int_{\lambda \in \text{supp}(\lambda)} (\mu^{(0)}(\alpha_1, \lambda) - \mu^{(0)}(\alpha_2, \lambda))^2 d\bar{\pi}_{T_0}(\lambda) \geq c_1 \|\alpha_1 - \alpha_2\|^2 \quad (9)$$

and

$$\begin{aligned} &\sup_{\alpha \in \mathcal{A}} \left| \int_{\lambda \in \text{supp}(\lambda)} \mu^{(0)}(\alpha, \lambda) (\mu^{(0)}(\alpha_1, \lambda) - \mu^{(0)}(\alpha_2, \lambda)) d\bar{\pi}_{T_0}(\lambda) \right| \\ &\geq c_2 \sqrt{\int_{\lambda \in \text{supp}(\lambda)} (\mu^{(0)}(\alpha_1, \lambda) - \mu^{(0)}(\alpha_2, \lambda))^2 d\bar{\pi}_{T_0}(\lambda)}. \end{aligned} \quad (10)$$

Assumption 4(i) requires the individual-specific latent factors to be independent and identically distributed as if units are (independently) drawn from the underlying population. At the same time, while Assumption 4(i) requires λ_t to be serially independent (and independent of the individual factors), it does *not* need to be stationary. Serial independence of λ_t simplifies the exposition and proofs, but may be stronger than needed. For example, it is straightforward to allow for weakly dependent λ_t ; see Remark 4. Moreover, with a simple modification, our method can also explicitly incorporate additive time effects allowed to follow an arbitrary (potentially strongly dependent) process; see Remark 5 for details. The final part of Assumption 4(i) requires 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 , which implies that the support of α_i is bounded.

Assumption 4(ii) is a standard overlap condition and would be required for regular estimation of the ATE even if α_i were observed. The bound away from one permits us to impute untreated potential outcomes for treated units using the outcomes of nearby untreated units, and it keeps the inverse weight $1/(1 - p_{i,t})$ in the doubly robust score bounded. The bound

¹²Here $\mathcal{L} := \bigcup_t \text{supp}(\lambda_t)$, and we use $\text{supp}(\lambda)$ as shorthand for $\mathcal{L}$ throughout.

away from zero ensures that the treated share $N_{1,t}/N$ is bounded away from zero, so that ATT_t can be estimated at the $\sqrt{N}$ rate.

Assumption 4(iii) enables us to apply particular concentration inequalities in order to obtain fast rates of convergence. The condition restricts the tail behavior of the error term $u_{i,t}$, requiring it to be uniformly sub-Gaussian conditional on some other variables. The assumption also imposes that, conditional on the latent factors α_i and λ_t , the residuals $u_{i,t}$ and $\epsilon_{i,t}$ are jointly independent of (a) the histories of these variables for all individuals other than i , (b) the individual-specific latent factors of individuals other than i , (c) lags of $u_{i,t}$, and (d) the aggregate factors λ_s for $s \neq t$. In short, variables must be cross-sectionally independent after conditioning on the aggregate factors λ_t , and the current outcome and treatment residuals must be serially independent of the lagged outcome residuals after conditioning on individual-specific factors α_i . One could relax this requirement and allow for serially correlated outcome residuals; see Remark 4 for details. The conditional independence restrictions in Assumption 4(iii) strengthen Assumption 3 to aid in the derivation of uniform convergence rates. While Assumption 3 is then redundant, we retain it for exposition because it plays a crucial role in identification.

Assumption 4(iv) imposes that the functions $\mu^{(0)}(\cdot, \cdot)$ and $p_t(\cdot)$ are bounded (on the relevant supports) and vary smoothly with the individual-specific factors.¹³ 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 $K(\cdot)$. Kernels that satisfy this condition are common in the literature, with the Epanechnikov kernel being a particularly prevalent choice.

Assumption 4(vi) relates the pseudo-distance to the distance between individual latent factors.¹⁴ 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 squared Euclidean distance between α_1 and α_2 . In addition, the assumption states that the population pseudo-distance is at least proportional to the square root of 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 (9) holds so long as the matrix $M \int_{\lambda \in \text{supp}(\lambda)} \lambda \lambda' d\pi_{T_0}(\lambda) M'$ is strictly positive definite, with its smallest eigenvalue

¹³Note that $p_t(\cdot)$ is bounded by construction. Boundedness of $\mu^{(0)}(\cdot, \cdot)$ is imposed for simplicity and can be relaxed. In particular, it is straightforward to allow for an unbounded common trend in the outcome model (1); see Remark 5 for details.

¹⁴See Sun and Zhang (2026) for a detailed discussion of a network version of this assumption and primitive conditions sufficient for it to hold.

bounded away from zero uniformly over all sufficiently large T_0 . If this is the case, then the second condition (10) holds if there is a $c > 0$ so that for any $\alpha_1, \alpha_2 \in \mathcal{A}$ with $\alpha_1 \neq \alpha_2$, we have $\frac{c}{\|\alpha_1 - \alpha_2\|}(\alpha_1 - \alpha_2) \in \mathcal{A}$. This is true, for example, if $\mathcal{A}$ contains a Euclidean ball centered at zero.

Theorem 1 (Rates of convergence). *Suppose Assumptions 1 to 4 hold and that $\frac{1}{h} \sqrt{\frac{\log(N)}{T_0}} \rightarrow 0$, $h \rightarrow 0$, and $\frac{\log(N)}{h^{d_\alpha} N} \rightarrow 0$. Then for $t > T_0$,*¹⁵

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

Theorem 1 establishes the uniform (over i) rate of convergence for the pseudo-distance based estimators (7). The derived rate is identical to the (pointwise) convergence rate of the infeasible Nadaraya-Watson estimators (5) that use the unobserved factors as regressors, up to an extra log term needed to ensure uniformity over i . Setting $h \propto N^{-1/(d_\alpha+2)}$ (which automatically satisfies the restriction $\frac{\log(N)}{h^{d_\alpha} N} \rightarrow 0$), we get convergence rate $N^{-1/(d_\alpha+2)} \sqrt{\log(N)}$ which, up to a log term, is the optimal (pointwise) nonparametric learning rate under Lipschitz continuity of the regression function (Stone, 1980). However, the rate at which h can converge to zero is also 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 optimal rate, up to log terms, provided that T_0 grows sufficiently quickly that $N^{2/(d_\alpha+2)} \log(N)/T_0 \rightarrow 0$.

Recall that, in order to obtain $\sqrt{N}$ -consistency and centered asymptotic normality of the doubly robust ATT estimator, it would be sufficient to demonstrate that (8) holds. In particular, 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 provided that T_0 grows quickly enough that $\frac{N^{1/2} \log(N)}{T_0} \rightarrow 0$. For $d_\alpha > 1$, Theorem 1 does not provide sufficiently fast rates for $\sqrt{N}$ -consistency of the ATT_t estimator.

Remark 4 (Relaxing serial independence). Assumption 4 imposes serial independence of the time-specific factors λ_t and residuals $u_{i,t}$. These conditions simplify the proof of Theorem 1 but may be stronger than needed.

In particular, these restrictions are used primarily to ensure that the sample pseudo-distance $\hat{d}_{i,j}$ is uniformly consistent for its population counterpart $d_{i,j}$, and to derive its rate of convergence. More precisely, the restrictions enable us to apply Bernstein-type inequalities

¹⁵Because T_0 is increasing with N , t is likewise understood to be a sequence increasing with N so that $t > T_0$.

(for independent random variables) to certain averages such as $\frac{1}{T_0} \sum_{t=1}^{T_0} u_{k,t} u_{i,t}$ and $\frac{1}{T_0} \sum_{t=1}^{T_0} \mu_{k,t}^{(0)} \mu_{i,t}^{(0)}$ to characterize their uniform (over i and k) rates of convergence to their population means. Thus, one may invoke appropriate concentration inequalities for weakly dependent (rather than serially independent) variables instead.

Remark 5 (Including additive time effects). Our assumptions do not rule out the possibility that some of the time-factors enter the model additively. However, the pseudo-distance in (6) is not invariant to these additive factors, and our stated formal results require that they satisfy the same conditions we place on non-additive factors. However, by adjusting the pseudo-distance we can eliminate additive time-effects entirely and thus avoid the need to place any conditions on these objects. In particular, γ_t can follow an arbitrary (potentially strongly dependent) process and be unbounded. To see this, suppose we include additive time-effects γ_t in model (2). That is, suppose that

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

The time-effects can be eliminated (in the pre-treatment period) in the usual way, by simply subtracting cross-sectional average outcomes. The corresponding sample and population pseudo-distances that no longer depend on the additive effects γ_t can be constructed as

$$\begin{aligned} \hat{d}_{i,j} &= \frac{1}{T_0} \max_{k \notin \{i,j\}} |\langle Y_{k,1:T_0} - \bar{Y}_{1:T_0}, Y_{i,1:T_0} - Y_{j,1:T_0} \rangle|, \\ d_{i,j} &= \sup_{\alpha \in \mathcal{A}} \left| \int (\mu^{(0)}(\alpha, \lambda) - \bar{\mu}^{(0)}(\lambda)) (\mu^{(0)}(\alpha_i, \lambda) - \mu^{(0)}(\alpha_j, \lambda)) d\bar{\pi}_{T_0}(\lambda) \right|, \end{aligned}$$

where $\bar{Y}_{1:T_0} := (\frac{1}{N} \sum_{i=1}^N Y_{i,1}, \frac{1}{N} \sum_{i=1}^N Y_{i,2}, \dots, \frac{1}{N} \sum_{i=1}^N Y_{i,T_0})'$ and $\bar{\mu}^{(0)}(\lambda) = \int \mu^{(0)}(\alpha, \lambda) d\pi_\alpha$ for π_α the distribution of α_k . One could adapt Theorem 1 (and the required assumptions) to apply with the alternative sample and population pseudo-distances above.

3.2 Asymptotic normality

In this section, we apply well-established ideas from the literature on double machine learning to the studied setting and demonstrate that, under certain conditions, our ATT estimator (with cross-fitting as specified in Algorithm 1) is $\sqrt{N}$ -consistent and has a centered asymptotically normal distribution.

Theorem 2 (Asymptotic normality). *Suppose that $\max_{i \in [N]} |\hat{p}_{i,t} - p_{i,t}| = o_p(1)$, $\max_{i \in [N]} |\hat{\mu}_{i,t}^{(0)} - \mu_{i,t}^{(0)}| = o_p(1)$, and that $\max_{i \in [N]} |(p_{i,t} - \hat{p}_{i,t})(\mu_{i,t}^{(0)} - \hat{\mu}_{i,t}^{(0)})| = o_p(N^{-1/2})$. In addition, suppose that the number of*

folds $\mathcal{K} \geq 2$ is fixed and, for each k , the size of fold $\mathcal{I}_k$ grows at the same rate as the sample size, that is, $\frac{|\mathcal{I}_k|}{N} \rightarrow c_k$ for some finite $c_k > 0$. Then, under Assumptions 1-4, we have that for $t > T_0$,

$$\sqrt{N}\hat{\text{ATT}}_t = \frac{\sqrt{N}}{N_{1,t}} \sum_{i=1}^N \left(Y_{i,t} w_{i,t} - \frac{(1 - w_{i,t})Y_{i,t}p_{i,t} + (w_{i,t} - p_{i,t})\mu_{i,t}^{(0)}}{1 - p_{i,t}} \right) + o_p(1),$$

in which case, if the usual conditions on the doubly robust ATT score $\psi_{i,t}$ apply conditional on λ_t ,¹⁶ we obtain $\sqrt{N}(\hat{\text{ATT}}_t - \text{ATT}_t) / \sqrt{V_t} \xrightarrow{d} N(0, 1)$, where V_t is the usual asymptotic variance of the doubly robust ATT estimator, i.e.,

$$V_t = \mathbb{E}_t[\psi_{i,t}^2], \quad \psi_{i,t} := \frac{1}{\mathbb{E}_t[w_{i,t}]} \left(Y_{i,t} w_{i,t} - \frac{(1 - w_{i,t})Y_{i,t}p_{i,t} + (w_{i,t} - p_{i,t})\mu_{i,t}^{(0)}}{1 - p_{i,t}} - w_{i,t}\text{ATT}_t \right).$$

The first result in Theorem 2 relates the estimate $\hat{\text{ATT}}_t$ to an infeasible estimate in which the first stage estimates of $\mu_{i,t}^{(0)}$ and $p_{i,t}$ are replaced by their true values. In particular, the theorem states that under certain conditions, the difference between $\hat{\text{ATT}}_t$ and the infeasible estimate disappears strictly faster than $N^{-1/2}$. It then follows from standard results that $\hat{\text{ATT}}_t$ is $\sqrt{N}$ -asymptotically normal and centered at ATT_t .¹⁷

The proof of Theorem 2 proceeds by similar steps to the results in Chernozhukov et al. (2018) for DML2 estimators. A complicating factor is that, unlike in standard DML2 estimation of the ATT, in our setting the estimates $\hat{\mu}_{i,t}^{(0)}$ and $\hat{p}_{i,t}$ are constructed using the pseudo-distance which includes outcome data for individual i in the pre-treatment period. That is, unlike in standard DML2 ATT estimation, the first-stage estimates $\hat{\mu}_{i,t}^{(0)}$ and $\hat{p}_{i,t}$ depend on some of the outcome data from the fold that contains individual i , which requires a more careful analysis.¹⁸

Theorem 2 requires not only that the first-stage estimates are uniformly consistent over the sample, but also that the product of their estimation errors is uniformly $o_p(N^{-1/2})$. A sufficient condition, mentioned earlier in this section, is that each of these estimates converges strictly faster than $N^{-1/4}$. The theorem further requires that each fold grows at the same rate as the sample size. This is satisfied if the number of folds is fixed and all of the folds are of equal size. In recent work, Velez (2026) provides conditions under which DML2 estimates are centered root- N asymptotically normal even if leave-one-out cross-fitting is applied, and so it may be possible to weaken this condition on the fold size in Theorem 2.

¹⁶Namely, that for each t , $\psi_{i,t}$ is iid across i conditional on λ_t , with $\mathbb{E}_t[\psi_{i,t}^2] \geq c$ and $\mathbb{E}[|\psi_{i,t}|^{2+\rho}] \leq c^{-1}$ for a fixed $\rho > 0$ and $0 < c < \infty$.

¹⁷Under the conditions of Theorem 2, $\hat{V}_t/V_t \xrightarrow{p} 1$, so the confidence interval in Algorithm 1 has asymptotic coverage $1 - \zeta$.

¹⁸In particular, it is straightforward to accommodate this dependence under the conditional serial independence of the errors imposed in Assumption 4. The provided argument also naturally extends to the case when the errors are weakly dependent.

Remark 6 (Semiparametric efficiency). The score $\psi_{i,t}$ in Theorem 2 is the well-known efficient influence function for the ATT in cross-sectional settings with *observed* unit characteristics α_i (Hahn, 1998). Therefore, under the conditions of the theorem, our estimator achieves the semiparametric efficiency bound for estimation of the ATT in the cross-sectional case with observed α_i , despite the fact that α_i is latent in our framework.

4 Simulation Study

In this section, we illustrate the finite-sample properties of the proposed method in two sets of numerical experiments. The first compares the proposed method with the workhorse two-way fixed effects (TWFE) approach, both when the additive TWFE specification is correct and when it is misspecified. The second compares the proposed method with the local PCA method of Feng (2026) under the nonlinear data generating processes.

4.1 Simulation with Linear DGPs

The TWFE approach is widely used in applied research and is designed to adjust for unobserved additive individual and time fixed effects. In this section, we demonstrate that the proposed method performs comparably well in settings in which the TWFE approach is valid. We also show that the proposed method maintains good finite sample properties in a more complicated setting in which the TWFE approach fails.

We consider a panel with individuals $i = 1, \dots, N$ and time periods $t = 1, \dots, T$, where $N \in \{50, 250\}$, the number of pre-treatment periods is $T_0 \in \{50, 250\}$, and $T = T_0 + 1$, so that there is exactly one post-treatment period. For the outcome we consider the following data generating processes:

$$\begin{cases} \text{Model 1 (Additive fixed effects):} & Y_{i,t} = \alpha_i + \lambda_t + 0.5 \cdot w_{i,t} + u_{i,t}, \\ \text{Model 2 (Interactive fixed effects):} & Y_{i,t} = \alpha_i \cdot \lambda_t + 0.5 \cdot w_{i,t} + u_{i,t}, \end{cases}$$

where $\alpha_i \sim \text{Uniform}(-1, 1)$, $\lambda_t \sim \text{Uniform}(-1, 1)$, and $u_{i,t} \sim N(0, 0.5^2)$ are drawn independently of one another and i.i.d. across i and t . The target estimand is the average treatment effect on the treated in the post-treatment period, $\text{ATT}_T = 0.5$. Models 1 and 2 are standard additive and interactive fixed effects specifications.

The treatment assignment mechanism depends on unobserved latent characteristics for each i , and the treatment occurs only in the final period. $w_{i,t} = 0$ for all $t \leq T_0$, and the

propensity score for the last period $p_{i,T}$ follows the model

$$w_{i,T} \mid \alpha_i \sim \text{Bernoulli}(p_{i,T}), \quad p_{i,T} = \frac{\exp(\alpha_i)}{1 + \exp(\alpha_i)}.$$

Because $\alpha_i \in [-1, 1]$, we have approximately $p_{i,T} \in [0.2689, 0.7311]$, which satisfies the overlap condition required for the proposed method.

We simulate 1000 replications from these models. In these simulations, we utilize the Epanechnikov kernel, which satisfies the requirements of Theorem 1 and is defined as $K(x) = \frac{3}{4}(1 - x^2) \cdot \mathbf{1}\{|x| \leq 1\}$. We use this kernel for both the outcome and propensity score imputations in our proposed method. Using this kernel, we compare the performance of different methods: (i) TWFE; (ii) the local PCA method proposed by Feng (2026) with 5-fold cross-validation for the choice of nearest neighbors;^{19,20} (iii) an infeasible doubly robust estimator in which the pseudo-distance is replaced by the oracle ℓ_2 distance between the true α 's; (iv) our proposed method.

Two sample-splitting procedures are used for different purposes. *Cross-fitting* refers to the sample splitting used when imputing the nuisance quantities $\hat{\mu}_{i,T}^{(0)}$ and $\hat{p}_{i,T}$ to control overfitting in doubly robust estimation. *Cross-validation* refers to the choice of the bandwidth h , which is further implemented with grid search and evaluated via post-treatment period squared prediction error of the imputed outcomes. Folds are drawn once per replication, and since N can be quite small in some settings, redrawing is sometimes needed if any training set contains fewer than two treated or two control units.

In the numerical experiments, the bandwidth is selected between 0.05 and 5. In Tables 1 and 2, we report the 95% confidence interval coverage and other statistics across 1000 replications in our simulation for each outcome model and each method. Additionally, estimated propensity scores close to one can generate large inverse-odds weights $\hat{p}_{i,T}/(1 - \hat{p}_{i,T})$ on untreated observations and hence very large estimates. Such replications are rare but heavy enough to distort mean-based statistics, and the effect is more pronounced when N is small. For this reason we report the median absolute error, the median of $|\text{estimate} - \text{true value}|$, and the median confidence interval length rather than their means.

¹⁹The implementation of the method proposed by Feng (2026) follows the replication files available on the author's website: https://github.com/yingjieum/replication-Feng_2026.

²⁰There are some minor numerical issues in the implementation of Feng (2026). During the simulation, LocalPCA occasionally fails to produce estimates during certain replications. This issue occurred in fewer than 10 replications for each different simulation design across various N, T_0 settings. Since the overall results are not significantly affected by these minor errors and we prefer to maintain the original implementation, we have decided to drop the NAs while averaging for the summary statistics. This also explains why some of the coverage results are not multiples of 0.001. Also note that the proposed method's implementation produced no errors in any of the simulations' replications.

Table 1: Simulation results for Model 1 (additive fixed effects)

		TWFE	Local PCA (Feng, 2026)	Proposed Method		
				W/ true α_i	Pseudo-Distance	
Cross-Fitting				LOO	None	5-Fold
Bandwidth Selection				LOO	LOO	5-Fold
$N = 50$ $T_0 = 50$	Median Abs. Err.	0.0951	0.1359	0.1233	0.1160	0.1342
	95% CI Coverage	0.9500	0.8862	0.9580	0.9030	0.9480
	Median CI Length	0.5642	0.6065	0.6897	0.5490	0.7121
$N = 50$ $T_0 = 250$	Median Abs. Err.	0.0935	0.1340	0.1191	0.1067	0.1271
	95% CI Coverage	0.9450	0.8740	0.9540	0.8980	0.9560
	Median CI Length	0.5588	0.5919	0.7027	0.5568	0.7203
$N = 250$ $T_0 = 50$	Median Abs. Err.	0.0442	0.0508	0.0490	0.0475	0.0483
	95% CI Coverage	0.9440	0.9320	0.9570	0.9250	0.9450
	Median CI Length	0.2507	0.2736	0.2789	0.2608	0.2827
$N = 250$ $T_0 = 250$	Median Abs. Err.	0.0442	0.0539	0.0514	0.0498	0.0504
	95% CI Coverage	0.9590	0.9410	0.9580	0.9400	0.9520
	Median CI Length	0.2488	0.2713	0.2801	0.2617	0.2828

Notes: 1,000 replications were conducted in each setting, where the estimand is $ATT_T = 0.5$. LOO and 5-fold denote leave-one-out and five-fold procedures, respectively, for cross-fitting or bandwidth cross-validation as indicated by the corresponding row; the bandwidth is chosen from the grid by minimizing post-period squared prediction error. If the selected bandwidth yields an empty kernel neighborhood, the next larger grid value is used. "None" indicates no cross-fitting. Local PCA follows the replication package of Feng (2026) with K chosen by 5-fold cross-validation.

For Model 1, the additive fixed effects model, the proposed method shows performance similar to the two-way fixed effects method in terms of 95% confidence interval coverage, which is under an ideal data generating process for TWFE. Additionally, the proposed method, which incorporates the pseudo-distance, effectively captures the distance between unobserved latent characteristics, which is evinced by the performance of our method relative to the oracle estimator which uses the true distance in the α 's. However, there is some loss of efficiency when the sample size N is small. This efficiency loss could perhaps be mitigated by increasing the number of folds or by using the leave-one-out procedure for bandwidth selection.

Table 2: Simulation results for Model 2 (interactive fixed effects)

		TWFE	Local PCA (Feng, 2026)	Proposed Method		
				W/ true α_i	Pseudo-Distance	
Cross-Fitting				LOO	None	5-Fold
Bandwidth Selection				LOO	LOO	5-Fold
$N = 50$ $T_0 = 50$	Median Abs. Err.	0.1725	0.1263	0.1189	0.1041	0.1228
	95% CI Coverage	0.8490	0.8999	0.9450	0.9110	0.9450
	Median CI Length	0.6780	0.5885	0.6290	0.5410	0.6415
$N = 50$ $T_0 = 250$	Median Abs. Err.	0.1724	0.1215	0.1108	0.1036	0.1163
	95% CI Coverage	0.8120	0.8899	0.9370	0.8920	0.9320
	Median CI Length	0.6728	0.5850	0.6337	0.5468	0.6432
$N = 250$ $T_0 = 50$	Median Abs. Err.	0.1644	0.0519	0.0485	0.0507	0.0512
	95% CI Coverage	0.4500	0.9410	0.9510	0.9250	0.9380
	Median CI Length	0.3014	0.2693	0.2687	0.2544	0.2704
$N = 250$ $T_0 = 250$	Median Abs. Err.	0.1501	0.0506	0.0498	0.0490	0.0504
	95% CI Coverage	0.5000	0.9390	0.9480	0.9300	0.9360
	Median CI Length	0.2991	0.2685	0.2690	0.2555	0.2692

Notes: Details are provided in the notes to Table 1.

For Model 2, the interactive fixed effects model, the TWFE estimator is misspecified. Its median absolute error remains between roughly 0.15 and 0.17 regardless of the sample size, while its confidence intervals shrink as N grows. As a result, the TWFE method’s coverage deteriorates sharply with N . In contrast, the proposed method maintains coverage close to the nominal level for all considered settings of N and T_0 .

4.2 Simulation with More Complex Nonlinear DGPs

We compare the performance of our approach with the method proposed by Feng (2026), which, to the best of our knowledge, is one of the few available inferential methods for average causal effects in nonlinear factor models. In this subsection, we use the same data generating processes employed in Feng (2026)²¹ but we consider different values of N and T .

²¹The replication package of Feng (2026) was originally released in 2024 on the GitHub page. The author later updated the package in 2026, making changes to both the simulation DGPs and the implementation of the LocalPCA method. Since the 2024 version includes a wider variety of nonlinear DGPs, we retain the simulation

The implementation is based on the replication code provided by the author. These numerical exercises are designed to investigate how the methods that we consider perform under relatively small sample sizes, which may be more typical of microeconomic applications.

Again, consider a large panel data setting with individuals labeled $i = 1, \dots, N$ over time periods $t = 1, \dots, T$, where $N \in \{50, 250, 1000\}$ and $T_0 \in \{50, 250, 1000\}$ with only one post-treatment time period. Our target estimand is the counterfactual mean $\mathbb{E}[Y_{i,T}(0)|w_{i,T} = 1]$ which is the object of interest in the simulations of [Feng \(2026\)](#). Estimation and inference for this counterfactual mean can be carried out in a manner analogous to those for the ATT_t . As with the ATT_t , one can plug our estimates of $\mu_{i,t}^{(0)}$ and $p_{i,t}$ into the standard doubly robust formulae.²² We consider the following DGPs for the untreated potential outcomes:

$$\begin{cases} \text{Model 3 (Nonlinear factor model):} & Y_{i,t}(0) = (\alpha_i - \lambda_t)^2 + u_{i,t}, \\ \text{Model 4 (Gaussian kernel):} & Y_{i,t}(0) = \frac{1}{0.1 \cdot \sqrt{2\pi}} \exp[-100(\alpha_i - \lambda_t)^2] + u_{i,t}, \\ \text{Model 5 (Exponential kernel):} & Y_{i,t}(0) = \exp(-10|\alpha_i - \lambda_t|) + u_{i,t}, \end{cases}$$

for $t \in \{1, \dots, T_0\}$. In the post-treatment time period, all three models follow

$$Y_{i,T}(0) = \alpha_i + \alpha_i^2 + \varepsilon_{i,T}^{(0)}, \quad Y_{i,T}(1) = 2\alpha_i + \alpha_i^2 + 1 + \varepsilon_{i,T}^{(1)},$$

where the latent factors and outcome residuals are drawn independently by $\alpha_i \sim \text{Uniform}(0, 1)$, $\lambda_t \sim \text{Uniform}(0, 1)$, $u_{i,t} \sim N(0, 0.5^2)$, and $\varepsilon_{i,t}^{(w)} \sim N(0, 1)$. Treatment is drawn independently across individuals by $w_{i,T}|\alpha_i \sim \text{Bernoulli}(p_{i,T})$, with the propensity score given by

$$p_{i,T} = \frac{\exp[(\alpha_i - 0.5) + (\alpha_i - 0.5)^2]}{1 + \exp[(\alpha_i - 0.5) + (\alpha_i - 0.5)^2]}.$$

The design of the DGP ensures that approximately $p_{i,T} \in [0.4378, 0.6792]$, which satisfies the required overlap condition.

The three designs differ only in the function that generates the pre-treatment history. The post-treatment outcome equations, the latent variables and the assignment mechanism are the same. The comparison therefore isolates how the quality of the pseudo-distance depends on the shape of the pre-treatment structure. Model 3 is a smooth polynomial factor model.

designs from that version. However, the LocalPCA implementation follows the exact same code as the most up-to-date 2026 paper and package. We also set the standard deviation of the pre-treatment disturbance to $\sigma = 0.5$, so that the pre-treatment signal-to-noise ratio is more favorable to all methods; results with $\sigma = 1$ are qualitatively similar.

²²To be precise, in this case the relevant estimator is $\hat{\theta}_t = \frac{1}{N_{i,t}} \sum_{i=1}^N Y_{i,t} w_{i,t} - \hat{ATT}_t$ and the variance estimate is $\frac{N}{N_{i,t}^2} \sum_{i=1}^N \left(\frac{(1-w_{i,t})Y_{i,t} \hat{p}_{i,t} + (w_{i,t} - \hat{p}_{i,t}) \hat{\mu}_{i,t}^{(0)}}{1 - \hat{p}_{i,t}} - w_{i,t} \hat{\theta}_t \right)^2$. For TWFE, here and in Section 5, we estimate the counterfactual mean as the mean post-treatment outcome among treated units minus the estimated treatment coefficient, with standard errors based on the influence function of this difference.

Models 4 and 5 are Gaussian- and exponential-kernel designs, which are harder to approximate. A similar simulation design is also used in [Fernández-Val et al. \(2021\)](#).

The results are provided in Tables 3, 4, and 5 below.²³ We consider the same methods and report the same statistics as for the previous numerical experiments. The number of replications is 1000.

In general, the proposed method provides good coverage across different sample sizes. Specifically, the configuration using five-fold cross-fitting (and five-fold cross-validation) generally delivers confidence interval coverage closer to the nominal level than the configuration without cross-fitting (and with leave-one-out cross-validation).

In Models 3, 4, and 5, the proposed method with five-fold cross-fitting achieves lower median absolute error than local PCA in 19 of the 27 designs, while delivering substantial improvements in confidence interval coverage in several settings. The coverage advantage is particularly pronounced when N is small or the pre-treatment history is short relative to N . For $N = 50$, local PCA has coverage below 90% in every design, whereas the proposed method attains coverage between 91.8% and 94.4%, accompanied by longer confidence intervals. In Model 5 with $N = 1000$ and $T_0 = 50$, the proposed method reduces median absolute error by approximately 30% relative to local PCA, from 0.0589 to 0.0413, while increasing coverage from 79.5% to 92.3%. Also note that the proposed method delivers these gains while estimating propensity scores nonparametrically, whereas local PCA uses local logistic regression with the same link function as the treatment assignment mechanism in the DGP. Across all reported designs, the proposed method's coverage is within three percentage points of the nominal 95% level in all but two cases, both of which occur in Model 3.

In summary, these simulations demonstrate that the proposed method performs comparably to TWFE when the additive TWFE specification is correct and continues to provide reliable inference in nonlinear settings where TWFE exhibits substantial undercoverage. Across the nonlinear designs, the proposed method often improves on local PCA in median absolute error and delivers substantial coverage gains in several settings, particularly when the cross-sectional sample size is small or the pre-treatment history is short relative to it. These results demonstrate that the proposed method combines simplicity with frequent gains in estimation accuracy and substantial improvements in coverage relative to the competing methods.

²³Note that the oracle benchmark uses the true individual factors α_i and post-treatment data. Within each model and for each fixed N , we use the same random seeds in corresponding replications to generate the individual factors, treatment assignments, and post-treatment potential outcomes across all T_0 settings. Since the benchmark does not use the pre-treatment history, its reported results remain identical as T_0 varies within each N block of a given table.

Table 3: Simulation results for Model 3 (nonlinear factor model)

		TWFE	Local PCA (Feng, 2026)	Proposed Method		
				W/ true α_i	Pseudo-Distance	
				LOO LOO	None LOO	5-Fold 5-Fold
Cross-Fitting Bandwidth Selection						
$N = 50$ $T_0 = 50$	Median Abs. Err.	0.1992	0.2147	0.1832	0.1751	0.2004
	95% CI Coverage	0.8740	0.8699	0.9330	0.8240	0.9180
	Median CI Length	0.8896	0.9498	1.0547	0.7292	1.0744
$N = 50$ $T_0 = 250$	Median Abs. Err.	0.2013	0.1916	0.1832	0.1666	0.2010
	95% CI Coverage	0.8720	0.8760	0.9330	0.8550	0.9410
	Median CI Length	0.8854	0.9552	1.0547	0.7823	1.0964
$N = 50$ $T_0 = 1000$	Median Abs. Err.	0.2017	0.1945	0.1832	0.1744	0.1938
	95% CI Coverage	0.8750	0.8899	0.9330	0.8730	0.9420
	Median CI Length	0.8857	0.9487	1.0547	0.8148	1.0959
$N = 250$ $T_0 = 50$	Median Abs. Err.	0.1716	0.0884	0.0754	0.0840	0.0841
	95% CI Coverage	0.6430	0.9160	0.9580	0.8950	0.9380
	Median CI Length	0.4119	0.4264	0.4353	0.3729	0.4289
$N = 250$ $T_0 = 250$	Median Abs. Err.	0.1706	0.0752	0.0754	0.0761	0.0838
	95% CI Coverage	0.6420	0.9400	0.9580	0.9150	0.9540
	Median CI Length	0.4103	0.4299	0.4353	0.3845	0.4485
$N = 250$ $T_0 = 1000$	Median Abs. Err.	0.1707	0.0745	0.0754	0.0744	0.0746
	95% CI Coverage	0.6420	0.9390	0.9580	0.9260	0.9540
	Median CI Length	0.4103	0.4317	0.4353	0.3969	0.4382
$N = 1000$ $T_0 = 50$	Median Abs. Err.	0.1633	0.0462	0.0378	0.0423	0.0412
	95% CI Coverage	0.1490	0.8910	0.9660	0.8790	0.9090
	Median CI Length	0.2060	0.2109	0.2130	0.1939	0.2094
$N = 1000$ $T_0 = 250$	Median Abs. Err.	0.1635	0.0383	0.0378	0.0371	0.0367
	95% CI Coverage	0.1360	0.9540	0.9660	0.9410	0.9590
	Median CI Length	0.2055	0.2132	0.2130	0.2009	0.2160
$N = 1000$ $T_0 = 1000$	Median Abs. Err.	0.1630	0.0382	0.0378	0.0375	0.0369
	95% CI Coverage	0.1420	0.9630	0.9660	0.9520	0.9620
	Median CI Length	0.2054	0.2139	0.2130	0.2020	0.2109

Notes: 1,000 replications were conducted in each setting, where the estimand is $\mathbb{E}[Y_{i,T}(0) | w_{i,T} = 1]$ with a true value around 0.9145. LOO and 5-fold denote leave-one-out and five-fold procedures, respectively, for cross-fitting or bandwidth cross-validation as indicated by the corresponding row; the bandwidth is chosen from the grid by minimizing post-period squared prediction error. If the selected bandwidth yields an empty kernel neighborhood, the next larger grid value is used. "None" indicates no cross-fitting. Local PCA follows the replication package of Feng (2026) with K chosen by 5-fold cross-validation.

Table 4: Simulation results for Model 4 (Gaussian kernel)

Cross-Fitting Bandwidth Selection		TWFE	Local PCA (Feng, 2026)	Proposed Method		
				W/ true α_i	Pseudo-Distance	
				LOO LOO	None LOO	5-Fold 5-Fold
$N = 50$ $T_0 = 50$	Median Abs. Err.	0.2104	0.2405	0.1832	0.1812	0.2410
	95% CI Coverage	0.8680	0.8679	0.9330	0.8330	0.9310
	Median CI Length	0.9113	1.0565	1.0547	0.8092	1.2549
$N = 50$ $T_0 = 250$	Median Abs. Err.	0.2052	0.2865	0.1832	0.1822	0.2427
	95% CI Coverage	0.8570	0.8529	0.9330	0.8420	0.9440
	Median CI Length	0.8875	1.0731	1.0547	0.8131	1.2789
$N = 50$ $T_0 = 1000$	Median Abs. Err.	0.2063	0.2567	0.1832	0.1738	0.2388
	95% CI Coverage	0.8550	0.8627	0.9330	0.8470	0.9420
	Median CI Length	0.8846	1.0406	1.0547	0.8094	1.2360
$N = 250$ $T_0 = 50$	Median Abs. Err.	0.1803	0.0829	0.0754	0.0753	0.0841
	95% CI Coverage	0.6040	0.9350	0.9580	0.9310	0.9660
	Median CI Length	0.4186	0.4404	0.4353	0.4116	0.4813
$N = 250$ $T_0 = 250$	Median Abs. Err.	0.1810	0.0820	0.0754	0.0762	0.0814
	95% CI Coverage	0.6010	0.9440	0.9580	0.9470	0.9660
	Median CI Length	0.4118	0.4405	0.4353	0.4140	0.4689
$N = 250$ $T_0 = 1000$	Median Abs. Err.	0.1824	0.0773	0.0754	0.0751	0.0761
	95% CI Coverage	0.5930	0.9430	0.9580	0.9470	0.9610
	Median CI Length	0.4096	0.4445	0.4353	0.4140	0.4630
$N = 1000$ $T_0 = 50$	Median Abs. Err.	0.1704	0.0384	0.0378	0.0382	0.0380
	95% CI Coverage	0.1230	0.9610	0.9660	0.9560	0.9660
	Median CI Length	0.2096	0.2150	0.2130	0.2105	0.2203
$N = 1000$ $T_0 = 250$	Median Abs. Err.	0.1745	0.0393	0.0378	0.0378	0.0385
	95% CI Coverage	0.1060	0.9580	0.9660	0.9620	0.9620
	Median CI Length	0.2064	0.2160	0.2130	0.2104	0.2178
$N = 1000$ $T_0 = 1000$	Median Abs. Err.	0.1730	0.0394	0.0378	0.0371	0.0367
	95% CI Coverage	0.0960	0.9560	0.9660	0.9620	0.9640
	Median CI Length	0.2055	0.2168	0.2130	0.2101	0.2167

Notes: Details are provided in the notes to Table 3.

Table 5: Simulation results for Model 5 (exponential kernel)

Cross-Fitting Bandwidth Selection		TWFE	Local PCA (Feng, 2026)	Proposed Method		
				W/ true α_i	Pseudo-Distance	
				LOO LOO	None LOO	5-Fold 5-Fold
$N = 50$ $T_0 = 50$	Median Abs. Err.	0.2005	0.1988	0.1832	0.1868	0.1979
	95% CI Coverage	0.8620	0.8730	0.9330	0.8040	0.9240
	Median CI Length	0.8850	0.9312	1.0547	0.7206	1.0791
$N = 50$ $T_0 = 250$	Median Abs. Err.	0.2003	0.2001	0.1832	0.1883	0.2128
	95% CI Coverage	0.8620	0.8769	0.9330	0.8210	0.9380
	Median CI Length	0.8818	0.9294	1.0547	0.7457	1.2056
$N = 50$ $T_0 = 1000$	Median Abs. Err.	0.1997	0.1881	0.1832	0.1727	0.2100
	95% CI Coverage	0.8610	0.8879	0.9330	0.8640	0.9390
	Median CI Length	0.8816	0.9205	1.0547	0.8073	1.1289
$N = 250$ $T_0 = 50$	Median Abs. Err.	0.1788	0.0942	0.0754	0.0835	0.0869
	95% CI Coverage	0.6080	0.8800	0.9580	0.8490	0.9270
	Median CI Length	0.4100	0.4189	0.4353	0.3565	0.4505
$N = 250$ $T_0 = 250$	Median Abs. Err.	0.1781	0.0804	0.0754	0.0797	0.0857
	95% CI Coverage	0.6140	0.9340	0.9580	0.9040	0.9550
	Median CI Length	0.4083	0.4198	0.4353	0.3729	0.4812
$N = 250$ $T_0 = 1000$	Median Abs. Err.	0.1783	0.0781	0.0754	0.0733	0.0751
	95% CI Coverage	0.6040	0.9230	0.9580	0.9360	0.9630
	Median CI Length	0.4079	0.4132	0.4353	0.4033	0.4400
$N = 1000$ $T_0 = 50$	Median Abs. Err.	0.1695	0.0589	0.0378	0.0425	0.0413
	95% CI Coverage	0.1180	0.7950	0.9660	0.8670	0.9230
	Median CI Length	0.2053	0.2069	0.2130	0.1858	0.2178
$N = 1000$ $T_0 = 250$	Median Abs. Err.	0.1712	0.0405	0.0378	0.0379	0.0395
	95% CI Coverage	0.1100	0.9350	0.9660	0.9420	0.9650
	Median CI Length	0.2047	0.2090	0.2130	0.1963	0.2238
$N = 1000$ $T_0 = 1000$	Median Abs. Err.	0.1710	0.0397	0.0378	0.0366	0.0367
	95% CI Coverage	0.1090	0.9460	0.9660	0.9540	0.9620
	Median CI Length	0.2045	0.2073	0.2130	0.2063	0.2114

Notes: Details are provided in the notes to Table 3.

5 Empirical Illustration

To evaluate the proposed method in settings that resemble real-world applications, we conduct an empirically calibrated simulation. We estimate the factors and loadings, fit the data nonparametrically, and then generate the panel by resampling from the estimated objects. This approach ensures that the generated panels reflect the underlying structure, heterogeneity, and noise levels of genuine data, while still allowing us full control over the sample size and potential smoothness.

5.1 Data

We calibrate the simulation to the Walmart weekly-sales panel used by [Abadie and Zhao \(2026\)](#), where the data are originally from [Prakash \(2023\)](#). The balanced panel records the weekly sales of $\tilde{N} = 45$ Walmart stores over $\tilde{T} = 143$ consecutive weeks, running from February 5, 2010, through October 26, 2012. The outcome variable is store-level weekly sales. After removing additive store and week fixed effects, the residualized panel retains a low-dimensional index and an empirical noise distribution; these are the features the calibrated DGP is designed to inherit. Note that since there is no treatment implemented in the actual data, the treatment mechanism for our simulation will be manually set up later. The data are standardized before the empirically calibrated panels are constructed, and we further denote this standardized outcome as $\tilde{Y}_{i,t}$.

5.2 Empirically Calibrated Data Generating Process

The DGP is calibrated nonparametrically in four steps: (i) estimate the additive fixed effects and the low-rank components; (ii) resample new individual-level factors and select the desired-length time-level loadings; (iii) fit the new factors back to outcomes nonparametrically and add resampled error terms; and (iv) assign treatment through a propensity score that depends on the latent features. A detailed description of each step is provided below.

Step (i) Extract factors: In a real-world scenario, although traditional two-way fixed effects often account for a significant portion of the variation, some variation is still explained by the nonlinear combination of remaining factors. To reflect this, we first remove additive two-way fixed effects $\bar{\alpha}_i$ and $\bar{\lambda}_t$ from the standardized outcome $\tilde{Y}_{i,t}$. Then with the remaining outcome $\bar{Y}_{i,t} := \tilde{Y}_{i,t} - \bar{\alpha}_i - \bar{\lambda}_t$, following [Bai \(2003\)](#), we estimate R factors and loadings, $\tilde{\alpha} \in \mathbb{R}^{\tilde{N} \times R}$ and $\tilde{\lambda} \in \mathbb{R}^{\tilde{T} \times R}$, and these extracted latent factors are the estimated low-dimensional indices

used to calibrate the subsequent nonparametric outcome model. In this simulation, we set our number of factors to be $R = 1$. Note that the additive fixed effects $\bar{\alpha}_i$ and $\bar{\lambda}_t$ are used only to isolate the remaining potentially nonlinear structure. As the original sales data are heavily driven by store fixed effects and seasonal effects, and the purpose of this empirically calibrated simulation is to study how the proposed method can further handle the nonlinear structure, we do not add those additive fixed effects back to the simulated panel.

Step (ii) Sample new factors: For a target panel of dimensions $N \times T$, new individual factors $\alpha = \{\alpha_i\}_{i \in \{1, \dots, N\}}$ are drawn by a kernel-density bootstrap²⁴ from the empirical distribution of $\tilde{\alpha}$. This produces a smooth analogue of the empirical distribution that supports arbitrary cross-sectional sample sizes N , while still preserving the empirical co-movement between factors when $R > 1$ is used.

As it is more difficult to extrapolate the time dimension, here we only consider the $T \leq \tilde{T} = 143$ setting. Rather than resampling the time dimension, we simply select T equally spaced calendar weeks from the original panel: simulated period t is mapped to the original week $\tau(t) := \text{round}(t \cdot \tilde{T}/T)$ for $t = 1, \dots, T$. The simulated time dimension therefore inherits the empirical time profile directly, including seasonality and trends, without a separately drawn time factor. Note that $\tau(T) = \tilde{T} = 143$, so the last simulated period corresponds to the same calendar week in every T setting. In sum, the new individual factors are $\alpha \in \mathbb{R}^{N \times R}$.

Step (iii) Fitting the outcome nonparametrically: With the empirically sampled factors, to generate the new outcome $Y_{i,t}(0)$ for $i \in \{1, \dots, N\}$ and $t \in \{1, \dots, T\}$, we then fit the outcome $\bar{Y}_{i,t}$ nonparametrically:

$$Y_{i,t}(0) = \underbrace{\frac{\sum_{j=1}^{\tilde{N}} K_h(\tilde{\alpha}_j - \alpha_i) \cdot \bar{Y}_{j,\tau(t)}}{\sum_{j=1}^{\tilde{N}} K_h(\tilde{\alpha}_j - \alpha_i)}}_{=: \hat{m}(\alpha_i, \tau(t))} + \underbrace{u_{i,t}}_{\text{sampled from residuals}}$$

In the kernel regression part, the product Epanechnikov kernel is used:

$$K_h(\mathbf{v}) = \prod_{r=1}^R \frac{3}{4} \left(1 - (v_r/h_r)^2\right) \cdot \mathbf{1}\{(v_r/h_r)^2 \leq 1\},$$

The bandwidth in each dimension follows a rule of thumb $h_r = c \cdot s_r \cdot \tilde{N}^{-1/(R+4)}$ for $r = 1, \dots, R$, where s_r is the sample standard deviation of the r -th loading, $\tilde{N} = 45$ is the number of *original* stores, and c is a tuning constant that indexes the calibration regimes below. The shocks

²⁴Here, we use the Epanechnikov kernel via R package `kernelboot::rmvk`, where each draw equals a uniformly resampled element of $\tilde{\alpha}$ plus Epanechnikov noise scaled by the package's default rule-of-thumb bandwidth.

$u_{i,t}$ are sampled uniformly and independently (with replacement) from the pooled fitted residuals $\bar{Y}_{\ell,k} - \hat{m}(\tilde{\alpha}_{\ell}, k)$ across all original stores ℓ and weeks k .

Step (iv) Treatment assignment mechanism and estimand: Since no treatment happened in the actual data, here, we generate a simple mechanism. Treatment is confined to a single period, the last one, with the first $T - 1 =: T_0$ periods serving as the pre-treatment time periods. Assignment is governed by a propensity score that is logistic in the latent individual factors $p_i = 1 / [1 + \exp(-a \cdot \sum_{r=1}^R \alpha_{i,r})]$ where we set $a = 2.5$ so that the propensity scores lie approximately in $[0.3, 0.7]$. In this simulation study, no treatment effect is injected into the outcome. The target estimand is the counterfactual control mean for the treated units in the period T , $\mathbb{E}[Y_{i,T}(0) \mid w_{i,T} = 1]$.

Two calibration regimes of smoothness: The constant c in the kernel regression controls how smooth the nonparametric structure is. A small c yields a small bandwidth, inducing fast singular-value decay. Here, we generate two settings for every sample size,

Setting 1 (fast decay): $c = 2.5$ Setting 2 (slow decay): $c = 6$

Figure 1 reports the singular value spectra of the two calibrated regimes from one replication alongside the residualized original data. In the figure, all spectra are normalized so that the

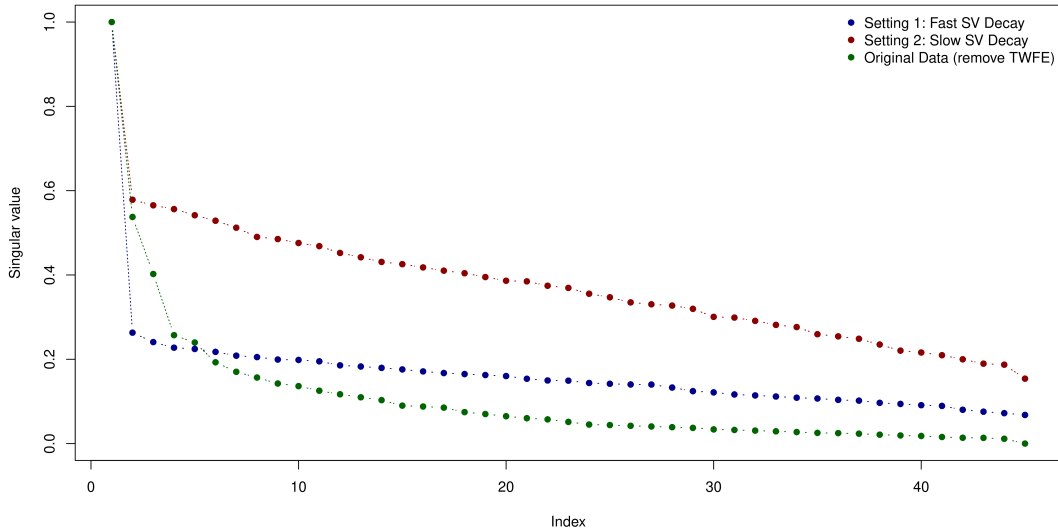


Figure 1: Singular Value Spectra

Notes: Each line is normalized by its own largest singular value. Settings 1 and 2 are single calibrated panels of dimension $\tilde{N} = 45$ by $\tilde{T} = 143$ from one replication. The green line represents the original Walmart panel after standardization and removal of additive store and week effects.

largest singular value equals one. Setting 1 falls sharply after the first component and then

declines slowly; the residualized Walmart panel falls more gradually over the first few components but continues to decline, so that beyond roughly the sixth component it sits below Setting 1. Setting 2 remains well above both across the entire spectrum. The mechanism is that a smaller calibration bandwidth produces a closer fit and hence smaller resampled residuals. Therefore, in Setting 1 the fitted component accounts for a larger share of the variation, whereas in Setting 2 the resampled noise dominates. We use these two settings to compare the estimators across panels whose singular values decay at visibly different rates.

5.3 Results

We vary the size of the panel dimensions over $N \in \{50, 100, 200\}$ and $T \in \{50, 100, 143\}$. Tables 6 and 7 report the results for the fast- and slow-decay regimes. In this empirical illustration, the bandwidth input for the proposed method is selected between 0.01 and 9.

The oracle benchmark that uses the true loading attains coverage between 0.921 and 0.957 under fast singular value decay and between 0.920 and 0.956 under slow singular value decay. The proposed estimator with the pseudo-distance performs very closely to this infeasible benchmark, with coverage between 0.923 and 0.954 and between 0.919 and 0.953, respectively. The configuration using five-fold cross-fitting and five-fold cross-validation again delivers coverage closer to the nominal level. The configuration without cross-fitting and with leave-one-out cross-validation yields shorter intervals and exhibits greater undercoverage, with coverage as low as 0.872; the coverage gap between the two configurations reaches five percentage points at $N = 50$ and narrows as N grows.

Under fast singular value decay, TWFE covers only 0.767–0.790 at $N = 200$, which suggests its bias may be large relative to its standard error. Local PCA does not suffer from this problem but undercovers in the smaller panels, with coverage between 0.877 and 0.890 at $N = 50$, and reaches approximately 0.934 at $N = 200$. Under slow singular value decay, all three feasible estimators perform much more similarly: at $N = 200$, TWFE covers 0.912–0.939 and local PCA 0.912–0.939. Taken together, the proposed estimator is close to the oracle benchmark across all reported designs, while the configuration without cross-fitting and with leave-one-out cross-validation generally produces shorter intervals and greater undercoverage, especially for smaller N .

Table 6: Empirically calibrated simulation results (fast singular-value decay)

		TWFE	Local PCA (Feng, 2026)	Proposed Method		
				W/ true α_i	Pseudo-Distance	
Cross-Fitting Bandwidth Selection				LOO LOO	None LOO	5-Fold 5-Fold
$N = 50$ $T = 50$	Median Abs. Err.	0.0224	0.0210	0.0201	0.0191	0.0200
	95% CI Coverage	0.9010	0.8820	0.9450	0.9030	0.9480
	Median CI Length	0.1023	0.1055	0.1124	0.0944	0.1146
$N = 50$ $T = 100$	Median Abs. Err.	0.0247	0.0220	0.0206	0.0202	0.0223
	95% CI Coverage	0.8490	0.8770	0.9210	0.8720	0.9230
	Median CI Length	0.1046	0.1041	0.1117	0.0950	0.1143
$N = 50$ $T = 143$	Median Abs. Err.	0.0218	0.0226	0.0198	0.0196	0.0214
	95% CI Coverage	0.8950	0.8900	0.9460	0.8880	0.9300
	Median CI Length	0.1016	0.1036	0.1126	0.0943	0.1157
$N = 100$ $T = 50$	Median Abs. Err.	0.0183	0.0157	0.0146	0.0148	0.0151
	95% CI Coverage	0.8470	0.9080	0.9250	0.9010	0.9240
	Median CI Length	0.0756	0.0764	0.0789	0.0704	0.0794
$N = 100$ $T = 100$	Median Abs. Err.	0.0189	0.0174	0.0150	0.0151	0.0152
	95% CI Coverage	0.8380	0.9110	0.9330	0.9070	0.9370
	Median CI Length	0.0755	0.0777	0.0783	0.0703	0.0806
$N = 100$ $T = 143$	Median Abs. Err.	0.0187	0.0154	0.0143	0.0139	0.0149
	95% CI Coverage	0.8550	0.9260	0.9400	0.9210	0.9420
	Median CI Length	0.0751	0.0777	0.0792	0.0716	0.0802
$N = 200$ $T = 50$	Median Abs. Err.	0.0154	0.0097	0.0092	0.0091	0.0096
	95% CI Coverage	0.7900	0.9340	0.9570	0.9360	0.9540
	Median CI Length	0.0548	0.0552	0.0563	0.0518	0.0558
$N = 200$ $T = 100$	Median Abs. Err.	0.0170	0.0099	0.0094	0.0095	0.0098
	95% CI Coverage	0.7700	0.9370	0.9480	0.9290	0.9440
	Median CI Length	0.0542	0.0552	0.0561	0.0520	0.0557
$N = 200$ $T = 143$	Median Abs. Err.	0.0173	0.0098	0.0100	0.0095	0.0098
	95% CI Coverage	0.7670	0.9340	0.9510	0.9280	0.9360
	Median CI Length	0.0536	0.0553	0.0560	0.0516	0.0558

Notes: 1,000 replications were conducted in each setting, where the estimand is $\mathbb{E}[Y_{i,T}(0) \mid w_{i,T} = 1]$, which is numerically approximated once per setting using 200,000 Monte Carlo draws of the noiseless calibrated regression function (values ≈ 0.008). LOO and 5-fold denote leave-one-out and five-fold procedures, respectively, for cross-fitting or bandwidth cross-validation as indicated by the corresponding row; the bandwidth is chosen from the grid by minimizing post-period squared prediction error. If the selected bandwidth yields an empty kernel neighborhood, the next larger grid value is used. "None" indicates no cross-fitting. Local PCA follows the replication package of Feng (2026) with K chosen by 5-fold cross-validation.

Table 7: Empirically calibrated simulation results (slow singular-value decay)

Cross-Fitting Bandwidth Selection		TWFE	Local PCA (Feng, 2026)	Proposed Method		
				W/ true α_i	Pseudo-Distance	
				LOO LOO	None LOO	5-Fold 5-Fold
$N = 50$ $T = 50$	Median Abs. Err.	0.0211	0.0248	0.0215	0.0204	0.0204
	95% CI Coverage	0.9360	0.9000	0.9420	0.9200	0.9440
	Median CI Length	0.1152	0.1229	0.1242	0.1092	0.1255
$N = 50$ $T = 100$	Median Abs. Err.	0.0248	0.0277	0.0248	0.0243	0.0251
	95% CI Coverage	0.8890	0.8700	0.9200	0.8820	0.9220
	Median CI Length	0.1131	0.1220	0.1252	0.1100	0.1254
$N = 50$ $T = 143$	Median Abs. Err.	0.0213	0.0249	0.0217	0.0215	0.0224
	95% CI Coverage	0.9330	0.8800	0.9480	0.9130	0.9490
	Median CI Length	0.1139	0.1200	0.1236	0.1091	0.1255
$N = 100$ $T = 50$	Median Abs. Err.	0.0159	0.0173	0.0160	0.0155	0.0160
	95% CI Coverage	0.9240	0.9090	0.9340	0.9100	0.9340
	Median CI Length	0.0846	0.0881	0.0873	0.0807	0.0874
$N = 100$ $T = 100$	Median Abs. Err.	0.0156	0.0175	0.0170	0.0164	0.0160
	95% CI Coverage	0.9180	0.9180	0.9320	0.9190	0.9340
	Median CI Length	0.0857	0.0895	0.0892	0.0832	0.0894
$N = 100$ $T = 143$	Median Abs. Err.	0.0162	0.0178	0.0158	0.0159	0.0166
	95% CI Coverage	0.9460	0.9140	0.9560	0.9390	0.9530
	Median CI Length	0.0857	0.0899	0.0897	0.0831	0.0898
$N = 200$ $T = 50$	Median Abs. Err.	0.0123	0.0122	0.0118	0.0114	0.0121
	95% CI Coverage	0.9390	0.9390	0.9470	0.9340	0.9430
	Median CI Length	0.0625	0.0636	0.0629	0.0602	0.0626
$N = 200$ $T = 100$	Median Abs. Err.	0.0116	0.0120	0.0107	0.0111	0.0114
	95% CI Coverage	0.9200	0.9300	0.9430	0.9310	0.9350
	Median CI Length	0.0613	0.0645	0.0625	0.0602	0.0625
$N = 200$ $T = 143$	Median Abs. Err.	0.0126	0.0132	0.0115	0.0118	0.0120
	95% CI Coverage	0.9120	0.9120	0.9210	0.9140	0.9190
	Median CI Length	0.0618	0.0641	0.0632	0.0606	0.0629

Notes: Details are provided in the notes to Table 6.

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.
- Abadie, A., Agarwal, A., and Shah, D. (2025). A causal inference framework for data rich environments. arXiv preprint arXiv:2504.01702.
- Abadie, A. and Zhao, J. (2026). Synthetic controls for experimental design. *The Review of Economics and Statistics*. Forthcoming.
- Abbring, J. H. and van den Berg, G. J. (2003). The nonparametric identification of treatment effects in duration models. *Econometrica*, 71(5):1491–1517.
- Agarwal, A., Dahleh, M., Shah, D., and Shen, D. (2023). Causal matrix completion. In Neu, G. and Rosasco, L., editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 3821–3826. PMLR.
- 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. W. (2022). Doubly robust identification for causal panel data models. *The Econometrics Journal*, 25(3):649–674.
- Arkhangelsky, D. and Imbens, G. W. (2024). Causal models for longitudinal and panel data: A survey. *The Econometrics Journal*, 27(3):C1–C61.
- 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.
- Athey, S. and Imbens, G. (2025). Identification of average treatment effects in nonparametric panel models. arXiv preprint arXiv:2503.19873.
- Auerbach, E. (2022). Identification and estimation of a partially linear regression model using network data. *Econometrica*, 90(1):347–365.

- Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, 71(1):135–171.
- 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.
- Beyhum, J. and Mugnier, M. (2024). Inference after discretizing time-varying unobserved heterogeneity. arXiv preprint arXiv:2412.07352.
- 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.
- 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. C. (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.
- Chen, X., Sant’Anna, P. H. C., and Xie, H. (2025). Efficient difference-in-differences and event study estimators. arXiv preprint arXiv:2506.17729.
- 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.
- Chernozhukov, V., Newey, W. K., and Singh, R. (2022). Debiased machine learning of global and local parameters using regularized Riesz representers. *The Econometrics Journal*, 25(3):576–601.
- 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.

- Chernozhukov, V., Wüthrich, K., and Zhu, Y. (2026). Debiasing and t -tests for synthetic control inference on average causal effects. *Journal of Political Economy*. Ahead of print. <https://doi.org/10.1086/742424>.
- Choi, J., Kwon, H., and Liao, Y. (2024). Inference for low-rank completion without sample splitting with application to treatment effect estimation. *Journal of Econometrics*, 240(1):105682.
- Choi, J. and Yuan, M. (2024). Matrix completion when missing is not at random and its applications in causal panel data models. *Journal of the American Statistical Association*, pages 1–15.
- 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.
- Dwivedi, R., Tian, K., Tomkins, S., Klasnja, P., Murphy, S., and Shah, D. (2025). Counterfactual inference in sequential experiments. *The Annals of Statistics*, 53(6):2380–2408.
- Feng, Y. (2024). Optimal estimation of large-dimensional nonlinear factor models. Working paper.
- Feng, Y. (2026). Causal inference in possibly 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.
- Freeman, H. and Kristensen, D. (2026). Inference on linear regressions with two-way unobserved heterogeneity. arXiv preprint arXiv:2605.06491.
- Ghanem, D., Sant’Anna, P. H. C., 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.
- Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2):315–331.

- Hoshino, T. and Yanagi, T. (2024). Estimating dyadic treatment effects with unknown confounders. arXiv preprint arXiv:2405.16547.
- Manski, C. F. and Pepper, J. V. (2018). How do right-to-carry laws affect crime rates? Coping with ambiguity using bounded-variation assumptions. *The Review of Economics and Statistics*, 100(2):232–244.
- Nowakowicz, K. (2024). Nonparametric network bootstrap. Working paper.
- Pesaran, M. H. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica*, 74(4):967–1012.
- Prakash, S. (2023). Walmart condensed sales data. Kaggle. <https://www.kaggle.com/ds/3471234>.
- Rambachan, A. and Roth, J. (2023). A more credible approach to parallel trends. *The Review of Economic Studies*, 90(5):2555–2591.
- Robins, J., Li, L., Tchetgen, E., and van der Vaart, A. (2008). Higher order influence functions and minimax estimation of nonlinear functionals. In Nolan, D. and Speed, T., editors, *Probability and Statistics: Essays in Honor of David A. Freedman*, volume 2 of *IMS Collections*, pages 335–421. Institute of Mathematical Statistics.
- Robins, J. M. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129.
- 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. C., 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. C. and Zhao, J. (2020). Doubly robust difference-in-differences estimators. *Journal of Econometrics*, 219(1):101–122.
- Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120.

- Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, 8(6):1348–1360.
- Sun, G. and Zhang, W. (2026). Flexible imputation of incomplete network data. arXiv preprint arXiv:2604.03171.
- Sun, L. and Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 225(2):175–199.
- Velez, A. (2026). Debiased machine learning with many cross-fitting folds. arXiv preprint arXiv:2411.01864.
- Wooldridge, J. M. (2025). Two-way fixed effects, the two-way Mundlak regression, and difference-in-differences estimators. *Empirical Economics*, 69:2545–2587.
- Xiong, R. and Pelger, M. (2023). Large dimensional latent factor modeling with missing observations and applications to causal inference. *Journal of Econometrics*, 233(1):271–301.
- Zeleneev, A. (2026). Identification and estimation of network models with nonparametric unobserved heterogeneity. arXiv preprint arXiv:2602.06885.
- Zhang, Y., Levina, E., and Zhu, J. (2017). Estimating network edge probabilities by neighbourhood smoothing. *Biometrika*, 104(4):771–783.

Online 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*

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

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

$$d_{i,j} = \sup_{\alpha_k \in \mathcal{A}} \left| \frac{1}{T_0} \sum_{t=1}^{T_0} \mathbb{E}[\mu^{(0)}(\alpha_k, \lambda_t)(\mu^{(0)}(\alpha_i, \lambda_t) - \mu^{(0)}(\alpha_j, \lambda_t)) | \alpha_i, \alpha_j, \alpha_k] \right|.$$

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

$$\begin{aligned} & \max_{i,j} |\hat{d}_{i,j} - \tilde{d}_{i,j}| \\ &= \max_{i,j} \left| \max_{k \notin \{i,j\}} \left| \frac{1}{T_0} \sum_{t=1}^{T_0} (\mu_{k,t}^{(0)} + u_{k,t})(\mu_{i,t}^{(0)} - \mu_{j,t}^{(0)} + u_{i,t} - u_{j,t}) \right| - \max_{k \notin \{i,j\}} \left| \frac{1}{T_0} \sum_{t=1}^{T_0} \mu_{k,t}^{(0)}(\mu_{i,t}^{(0)} - \mu_{j,t}^{(0)}) \right| \right| \\ &\leq \max_{i,j,k: k \notin \{i,j\}} \left| \frac{1}{T_0} \sum_{t=1}^{T_0} [u_{k,t}(\mu_{i,t}^{(0)} - \mu_{j,t}^{(0)}) + \mu_{k,t}^{(0)}(u_{i,t} - u_{j,t}) + u_{k,t}(u_{i,t} - u_{j,t})] \right|. \end{aligned} \quad (11)$$

By (2) and Assumption 4 (iii), there exist constants C , a , and b , so that for any $\eta > 0$,

$$\mathbb{P} \left(\left| \frac{1}{T_0} \sum_{t=1}^{T_0} [(u_{k,t}(\mu_{i,t}^{(0)} - \mu_{j,t}^{(0)}) + \mu_{k,t}^{(0)}(u_{i,t} - u_{j,t}) + u_{k,t}(u_{i,t} - u_{j,t}))] \right| > \eta \right) \leq C \exp \left(-\frac{T_0 \eta^2}{a + b\eta} \right). \quad (12)$$

To see this, note that the independence conditions in Assumption 4 (iii) imply that for any $t \leq T_0$ and any r and l with $l \neq r$ in $\{i, j, k\}$, $u_{r,t} \perp\!\!\!\perp \alpha_l, \{\lambda_s\}_{s \neq t} | \alpha_r, \lambda_t$,²⁵ which in turn implies

²⁵This and the other conditional independencies in this paragraph can be derived from the two conditional independence restrictions in 4 (iii) by repeated application of the rule $A \perp\!\!\!\perp B, C | D \implies (A \perp\!\!\!\perp B | C, D \text{ and } A \perp\!\!\!\perp B | D)$, and the rule $(A \perp\!\!\!\perp B | C, D \text{ and } A \perp\!\!\!\perp C | D) \implies A \perp\!\!\!\perp B, C | D$, for any random variables A, B, C, D .

$\mathbb{E}[u_{r,t}|\alpha_l, \alpha_r, \{\lambda_s\}_{s=1}^{T_0}] = \mathbb{E}[u_{r,t}|\alpha_r, \lambda_t]$ and this is zero by (2), and so $\mathbb{E}[u_{r,t}|\alpha_l, \alpha_r, \{\lambda_s\}_{s=1}^{T_0}] = 0$. Additionally, by Assumption 4 (iii), $u_{r,t} \perp\!\!\!\perp u_{l,t} \mid \alpha_r, \alpha_l, \{\lambda_s\}_{s=1}^{T_0}$. Moreover, by Assumption 4 (iv), $\mu_{i,t}^{(0)}$, $\mu_{j,t}^{(0)}$, and $\mu_{k,t}^{(0)}$ are bounded. By Assumption 4 (iii), for each $t \leq T_0$ and $r, l \in \{i, j, k\}$, $(u_{r,t}, u_{l,t}) \perp\!\!\!\perp \{(u_{r,s}, u_{l,s})\}_{s=1}^{t-1} \mid \alpha_r, \alpha_l, \{\lambda_s\}_{s=1}^{T_0}$, and $u_{r,t}$ is sub-Gaussian conditional on α_r, α_l , and $\{\lambda_s\}_{s=1}^{T_0}$. Thus, from the Bernstein inequalities in Lemma A.8, we get that there exist constants C , a , and b , so that for any $\eta > 0$,

$$\begin{aligned} \mathbb{P}\left(\left|\frac{1}{T_0} \sum_{t=1}^{T_0} u_{k,t} \mu_{r,t}^{(0)}\right| > \frac{\eta}{6} \mid \alpha_r, \alpha_k, \{\lambda_t\}_{t=1}^{T_0}\right) &\leq \frac{C}{6} \exp\left(-\frac{T_0 \eta^2}{a + b\eta}\right), \quad r = i, j \\ \mathbb{P}\left(\left|\frac{1}{T_0} \sum_{t=1}^{T_0} \mu_{k,t}^{(0)} u_{r,t}\right| > \frac{\eta}{6} \mid \alpha_r, \alpha_k, \{\lambda_t\}_{t=1}^{T_0}\right) &\leq \frac{C}{6} \exp\left(-\frac{T_0 \eta^2}{a + b\eta}\right), \quad r = i, j \\ \mathbb{P}\left(\left|\frac{1}{T_0} \sum_{t=1}^{T_0} u_{k,t} u_{r,t}\right| > \frac{\eta}{6} \mid \alpha_r, \alpha_k, \{\lambda_t\}_{t=1}^{T_0}\right) &\leq \frac{C}{6} \exp\left(-\frac{T_0 \eta^2}{a + b\eta}\right), \quad r = i, j \end{aligned}$$

These inequalities hold for some a , b , and C by the Bernstein bounds in Lemma A.8. Applying the law of iterated expectations to integrate out the factors, and then using the triangle inequality and the union bound yields (12).

By taking the union bound and (11), we then get from (12) that

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

Set $\eta = \sqrt{\frac{Ma \log(N)}{T_0}}$ for some fixed $M > 3$, in which case by $\frac{\log(N)}{T_0} \rightarrow 0$, we have $\eta \rightarrow 0$. Then by the above inequality

$$\mathbb{P}\left(\max_{i,j} |\hat{d}_{i,j} - \tilde{d}_{i,j}| > \sqrt{\frac{Ma \log(N)}{T_0}}\right) \leq C \exp\left(\frac{(3-M)a + 3b\eta}{a + b\eta} \log(N)\right) \rightarrow 0,$$

and so we conclude that

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

From $\tilde{d}_{i,j}$ to $d_{i,j}$ By the triangle inequality and reverse triangle inequality, the maximum difference between the population and oracle distances satisfies

$$\begin{aligned} &\max_{i,j} |d_{i,j} - \tilde{d}_{i,j}| \\ &\leq \max_{i,j,k: k \notin \{i,j\}} \left| \frac{1}{T_0} \sum_{t=1}^{T_0} \mu_{k,t}^{(0)} (\mu_{i,t}^{(0)} - \mu_{j,t}^{(0)}) - \mathbb{E}\left[\frac{1}{T_0} \sum_{t=1}^{T_0} \mu_{k,t}^{(0)} (\mu_{i,t}^{(0)} - \mu_{j,t}^{(0)}) \mid \alpha_i, \alpha_j, \alpha_k\right] \right| \end{aligned} \quad (13)$$

$$\begin{aligned} &+ \max_{i,j} \left[\sup_{\alpha \in \mathcal{A}} \left| \mathbb{E}\left[\frac{1}{T_0} \sum_{t=1}^{T_0} \mu^{(0)}(\alpha, \lambda_t) (\mu_{i,t}^{(0)} - \mu_{j,t}^{(0)}) \mid \alpha_i, \alpha_j\right] \right| \right. \\ &\quad \left. - \max_{k: k \notin \{i,j\}} \left| \mathbb{E}\left[\frac{1}{T_0} \sum_{t=1}^{T_0} \mu_{k,t}^{(0)} (\mu_{i,t}^{(0)} - \mu_{j,t}^{(0)}) \mid \alpha_i, \alpha_j, \alpha_k\right] \right| \right]. \end{aligned} \quad (14)$$

We first derive the rate for (13). By Assumption 4, conditions (i) and (iv), the vectors $(\mu_{i,t}^{(0)}, \mu_{j,t}^{(0)}, \mu_{k,t}^{(0)})$ are uniformly bounded and independent across t , conditional on $\alpha_i, \alpha_j, \alpha_k$. Therefore, by the Bernstein inequality for bounded random variables (Lemma A.5), there are positive constants C, a , and b , uniform in i, j, k, N, T_0 , such that, for any $\eta > 0$,

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

Integrating out the individual factors gives

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

Applying the union bound over at most N^3 triples yields

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

Set $\eta = \sqrt{Ma \log(N)/T_0}$ for some fixed $M > 3$. Since $\log(N)/T_0 \rightarrow 0$, we have $\eta \rightarrow 0$, and the last bound becomes

$$C \exp \left[\left(3 - \frac{Ma}{a + b\eta} \right) \log(N) \right] \rightarrow 0.$$

Consequently,

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

We next derive the rate for (14). The terms with $i = j$ are zero, so it suffices to consider distinct pairs. By the reverse triangle inequality, boundedness of $\mu^{(0)}$, and the Lipschitz con-

dition in Assumption 4(iv), there is a constant C such that

$$\begin{aligned}
& \max_{i \neq j} \left[\sup_{\alpha \in \mathcal{A}} \left| \mathbb{E} \left[\frac{1}{T_0} \sum_{t=1}^{T_0} \mu^{(0)}(\alpha, \lambda_t) (\mu_{i,t}^{(0)} - \mu_{j,t}^{(0)}) \mid \alpha_i, \alpha_j \right] \right| \right. \\
& \quad \left. - \max_{k: k \notin \{i,j\}} \left| \mathbb{E} \left[\frac{1}{T_0} \sum_{t=1}^{T_0} \mu_{k,t}^{(0)} (\mu_{i,t}^{(0)} - \mu_{j,t}^{(0)}) \mid \alpha_i, \alpha_j, \alpha_k \right] \right| \right] \\
& \leq \max_{i \neq j} \sup_{\alpha \in \mathcal{A}} \min_{k \notin \{i,j\}} \left| \mathbb{E} \left[\frac{1}{T_0} \sum_{t=1}^{T_0} (\mu^{(0)}(\alpha, \lambda_t) - \mu^{(0)}(\alpha_k, \lambda_t)) (\mu_{i,t}^{(0)} - \mu_{j,t}^{(0)}) \mid \alpha_i, \alpha_j, \alpha_k \right] \right| \\
& \leq C \max_{i \neq j} \sup_{\alpha \in \mathcal{A}} \min_{k \notin \{i,j\}} \|\alpha - \alpha_k\|.
\end{aligned}$$

Define

$$R_N^{(-2)} := \max_{i \neq j} \sup_{\alpha \in \mathcal{A}} \min_{k \notin \{i,j\}} \|\alpha - \alpha_k\|.$$

By the small-ball condition in Assumption 4(i), for all sufficiently small $\eta > 0$, there is a deterministic η -covering of $\mathcal{A}$ with centres $\{\bar{\alpha}_\ell\}_{\ell=1}^J \subset \mathcal{A}$ and $J \leq C\eta^{-d_\alpha}$. The triangle inequality therefore gives

$$R_N^{(-2)} \leq \eta + \max_{i \neq j} \max_{\ell \in [J]} \min_{k \notin \{i,j\}} \|\bar{\alpha}_\ell - \alpha_k\|.$$

The same small-ball condition implies that there exist $\delta > 0$ and $\epsilon_0 > 0$ such that, for every $\alpha \in \mathcal{A}$ and $0 < \epsilon \leq \epsilon_0$,

$$\mathbb{P}(\|\alpha - \alpha_k\| > \epsilon) \leq 1 - \delta\epsilon^{d_\alpha}.$$

For each fixed pair $i \neq j$ and each $\ell \in [J]$, independence of the individual factors gives

$$\begin{aligned}
\mathbb{P}\left(\min_{k \notin \{i,j\}} \|\bar{\alpha}_\ell - \alpha_k\| > \epsilon\right) &= \prod_{k \notin \{i,j\}} \mathbb{P}(\|\bar{\alpha}_\ell - \alpha_k\| > \epsilon) \\
&\leq (1 - \delta\epsilon^{d_\alpha})^{N-2}.
\end{aligned}$$

Taking a union bound over both the excluded pairs and the covering centres, we obtain

$$\begin{aligned}
\mathbb{P}\left(R_N^{(-2)} > \eta + \epsilon\right) &\leq \mathbb{P}\left(\max_{i \neq j} \max_{\ell \in [J]} \min_{k \notin \{i,j\}} \|\bar{\alpha}_\ell - \alpha_k\| > \epsilon\right) \\
&\leq \sum_{i \neq j} \sum_{\ell=1}^J \mathbb{P}\left(\min_{k \notin \{i,j\}} \|\bar{\alpha}_\ell - \alpha_k\| > \epsilon\right) \\
&\leq CN^2 \eta^{-d_\alpha} (1 - \delta\epsilon^{d_\alpha})^{N-2}.
\end{aligned}$$

Set

$$\eta_N = \frac{1}{N-2}, \quad \epsilon_N = \left(\frac{M \log(N-2)}{\delta(N-2)} \right)^{1/d_\alpha},$$

for some fixed $M > d_\alpha + 2$. For all sufficiently large N , these radii lie within the ranges required above, and

$$\begin{aligned} \mathbb{P}\left(R_N^{(-2)} > \eta_N + \epsilon_N\right) &\leq CN^2(N-2)^{d_\alpha} \left(1 - \frac{M \log(N-2)}{N-2}\right)^{N-2} \\ &\leq CN^2(N-2)^{d_\alpha - M} \longrightarrow 0. \end{aligned}$$

Since $\eta_N = o(\epsilon_N)$ for $d_\alpha \geq 1$, it follows that

$$\max_{i \neq j} \sup_{\alpha \in \mathcal{A}} \min_{k \notin \{i, j\}} \|\alpha - \alpha_k\| = O_p \left(\left(\frac{\log(N)}{N} \right)^{1/d_\alpha} \right).$$

Combining the bounds for (13) and (14) with the preceding bound for $\max_{i,j} |\hat{d}_{i,j} - \tilde{d}_{i,j}|$, and applying the triangle inequality, we conclude that

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

□

Lemma A.2. Under Assumption 4 (vi), there exists a $c > 0$ so that for all sufficiently large T_0 and $\alpha_1, \alpha_2 \in \mathcal{A}$:

$$\sup_{\alpha \in \mathcal{A}} \left| \int_{\lambda \in \text{supp}(\lambda)} \mu^{(0)}(\alpha, \lambda) (\mu^{(0)}(\alpha_1, \lambda) - \mu^{(0)}(\alpha_2, \lambda)) d\bar{\pi}_{T_0}(\lambda) \right| \geq c \|\alpha_1 - \alpha_2\|.$$

Proof of Lemma A.2. By Assumption 4 (vi), there exists a constant $c_2 > 0$ such that

$$\begin{aligned} &c_2 \sqrt{\int_{\lambda \in \text{supp}(\lambda)} (\mu^{(0)}(\alpha_1, \lambda) - \mu^{(0)}(\alpha_2, \lambda))^2 d\bar{\pi}_{T_0}(\lambda)} \\ &\leq \sup_{\alpha \in \mathcal{A}} \left| \int_{\lambda \in \text{supp}(\lambda)} \mu^{(0)}(\alpha, \lambda) (\mu^{(0)}(\alpha_1, \lambda) - \mu^{(0)}(\alpha_2, \lambda)) d\bar{\pi}_{T_0}(\lambda) \right|, \end{aligned}$$

and the object on the LHS is bounded below by $c_2 \sqrt{c_1} \|\alpha_1 - \alpha_2\|$. Thus letting $c = c_2 \sqrt{c_1} > 0$ gives the result. □

Lemma A.3. Suppose Assumption 4 holds and that $\frac{1}{h} \sqrt{\frac{\log(N)}{T_0}}$ and $\frac{\log(N)}{Nh^{d_\alpha}}$ go to zero. Then 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

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

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

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

By supposition, $\frac{1}{h} \sqrt{\frac{\log(N)}{T_0}}$ and $\frac{\log(N)}{Nh^{d_\alpha}}$ go to zero, and so for any fixed $\delta > 0$, with probability approaching one, we have $\frac{\max_{i,j} |\hat{d}_{i,j} - d_{i,j}|}{h} \leq \frac{1}{2} \delta C$. Hence, for any fixed $\delta > 0$, with probability approaching one, for all i, j , it follows that

$$\|\alpha_i - \alpha_j\| > \delta h \implies \frac{\hat{d}_{i,j}}{h} > \frac{1}{2} \delta C$$

Assumption 4 (v) states that the kernel is zero everywhere other than a compact (and therefore bounded) set. 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, fixing some $\delta > \frac{2c}{C}$ gives the desired result. $\square$

Lemma A.4. Suppose Assumption 4 holds, $\frac{\log N}{Nh^{d_\alpha}} \rightarrow 0$, and $h \rightarrow 0$. Then for any fixed $\delta > 0$, there are constants $0 < \underline{c} < \bar{c} < \infty$ and $0 < \tilde{c}$, such that for sufficiently large T_0 , with probability approaching one and for all $i \in [N]$, it follows that

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

and

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

and in addition, $\tilde{c} \leq \frac{1}{Nh^{d_\alpha}} \sum_{j=1}^N (1 - w_{j,t}) K(d_{i,j}/h)$ and $\tilde{c} \leq \frac{1}{Nh^{d_\alpha}} \sum_{j=1}^N K(d_{i,j}/h)$.

Proof of Lemma A.4. Define $X_{i,j} = \mathbf{1} \{ \|\alpha_i - \alpha_j\| \leq \delta h \}$ and $X_{i,j}^{(w)} = (1 - w_{j,t}) \mathbf{1} \{ \|\alpha_i - \alpha_j\| \leq \delta h \}$. By the upper bound in Assumption 4 (ii), for some $c > 0$, $\mathbb{E}_t[1 - w_{j,t} | \alpha_j] = 1 - p_{j,t} \geq c$ almost

surely. Using the fact that $w_{j,t}$ and α_j are jointly independent of α_i and that $\{\alpha_j\}_{j=1}^N$ are i.i.d. and independent of λ_t (by Assumptions 4 (iii) and 4 (i)), we have

$$\begin{aligned}\mathbb{E}_t[X_{i,j}^{(w)} \mid \alpha_i] &= \mathbb{E}_t[\mathbb{E}_t[1 - w_{j,t}|\alpha_j] \mathbf{1}\{\|\alpha_i - \alpha_j\| \leq \delta h\} \mid \alpha_i] \\ &\geq c \cdot \mathbb{E}_t[\mathbf{1}\{\|\alpha_i - \alpha_j\| \leq \delta h\} \mid \alpha_i] \\ &= c \cdot \mathbb{P}(\|\alpha_i - \alpha_j\| \leq \delta h \mid \alpha_i).\end{aligned}$$

To upper bound the quantity note that

$$\begin{aligned}\mathbb{E}_t[X_{i,j}^{(w)} \mid \alpha_i] &\leq \mathbb{E}_t[\mathbf{1}\{\|\alpha_i - \alpha_j\| \leq \delta h\} \mid \alpha_i] \\ &= \mathbb{P}(\|\alpha_i - \alpha_j\| \leq \delta h \mid \alpha_i).\end{aligned}$$

With Assumption 4 (i), we then see that for any fixed δ and sufficiently small $h > 0$, there are constants $0 < c_- < c_+ < \infty$ such that

$$c_- h^{d_\alpha} \leq \mathbb{E}_t[X_{i,j}^{(w)} \mid \alpha_i] \leq c_+ h^{d_\alpha}.$$

By Assumption 4 (i), $\mathbb{E}_t[X_{i,j} \mid \alpha_i] = \mathbb{P}(\|\alpha_i - \alpha_j\| \leq \delta h \mid \alpha_i)$ and so we see that the above also holds for $X_{i,j}^{(w)}$ replaced by $X_{i,j}$. Note that the variables $X_{i,j}^{(w)}$ for $j \neq i$ are independent Bernoulli conditional on α_i and λ_t , and likewise for $X_{i,j}$, and so the above also demonstrates that they have conditional variance below $c_+ h^{d_\alpha}$. So applying the Bernstein inequality in Lemma A.7, it follows that, for a sufficiently small fixed $a > 0$ and setting $C = \frac{a^2}{2(c_+ + a/3)}$,

$$\begin{aligned}&\mathbb{P}_t\left(\max_{i \in [N]} \left| \frac{1}{N} \sum_{j \neq i} (X_{i,j}^{(w)} - \mathbb{E}_t[X_{i,j}^{(w)} \mid \alpha_i]) \right| > ah^{d_\alpha}\right) \\ &\leq 2N \cdot \exp(-CNh^{d_\alpha}) \\ &= 2 \cdot \exp(\log(N) - CNh^{d_\alpha}),\end{aligned}$$

and similarly with $X_{i,j}^{(w)}$ replaced by $X_{i,j}$. The RHS goes to zero if $\frac{\log N}{Nh^{d_\alpha}} \rightarrow 0$, which holds by supposition, and so with probability approaching one, for all $i \in [N]$,

$$\left(\frac{N-1}{N}c_- - a\right)h^{d_\alpha} \leq \frac{1}{N} \sum_{j \neq i} X_{i,j}^{(w)} \leq (c_+ + a)h^{d_\alpha},$$

in which case, including the $j = i$ term in the sum,

$$\left(\frac{N-1}{N}c_- - a\right)h^{d_\alpha} \leq \frac{1}{N} \sum_{j=1}^N X_{i,j}^{(w)} \leq (c_+ + a)h^{d_\alpha} + 1/N,$$

and the same steps apply with $X_{i,j}^{(w)}$ replaced by $X_{i,j}$. By supposition $1/N = o(h^{d_\alpha})$, and so by setting a sufficiently small, we get the first two results in the Lemma.

Now for the final two statements in Lemma A.4, by Assumption 4 (iv), it follows that $\mu^{(0)}(\cdot, \lambda)$ is Lipschitz continuous and bounded (uniformly over λ in both cases). Thus, for some constant $C < \infty$, we have

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

Since Assumption 4 (v) requires the kernel to be weakly positive, strictly positive at zero, and Lipschitz continuous, there are constants $a, b > 0$ 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

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

and

$$b \cdot \frac{1}{N} \sum_{j=1}^N \mathbf{1} \left\{ \|\alpha_i - \alpha_j\| \leq \frac{a}{C} \cdot h \right\} \leq \frac{1}{N} \sum_{j=1}^N K(d_{i,j}/h).$$

We already established that for any fixed δ , there is a constant $\underline{c} > 0$ such that $\underline{c}h^{d_a} \leq \frac{1}{N} \sum_{j=1}^N (1 - w_{j,t}) \mathbf{1} \left\{ \|\alpha_i - \alpha_j\| \leq \delta h \right\}$ and $\underline{c}h^{d_a} \leq \frac{1}{N} \sum_{j=1}^N \mathbf{1} \left\{ \|\alpha_i - \alpha_j\| \leq \delta h \right\}$ 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 ,

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

and similarly $\tilde{c} \leq \frac{1}{Nh^{d_a}} \sum_{j=1}^N K(d_{i,j}/h)$. □

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, replacing $u_{j,t}$ by $\epsilon_{j,t}$, dropping the restriction to $\{j : w_{j,t} = 0\}$ in the kernel sums, using that $\epsilon_{j,t} \in [-1, 1]$ is bounded and hence sub-Gaussian, and removing conditioning on current treatments throughout the argument.

By Lemma A.3, there exists a $\delta < \infty$ such that with probability approaching one, for all $i, j \in [N]$, we have

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

and by Lemma A.2 the same holds with $\hat{d}_{i,j}$ replaced by $d_{i,j}$. Fix such a δ . In the event that these equalities hold, we can then decompose $\hat{\mu}_{i,t}^{(0)} - \mu_{i,t}^{(0)}$ as follows:

$$\begin{aligned} \hat{\mu}_{i,t}^{(0)} - \mu_{i,t}^{(0)} &= \frac{\sum_{j:w_{j,t}=0} \mathbf{1}\{\|\alpha_i - \alpha_j\| \leq \delta h\} K(\hat{d}_{i,j}/h) \cdot \mu_{j,t}^{(0)}}{\sum_{j:w_{j,t}=0} \mathbf{1}\{\|\alpha_i - \alpha_j\| \leq \delta h\} K(\hat{d}_{i,j}/h)} - \mu_{i,t}^{(0)} \\ &\quad + \frac{\frac{1}{h^{d_\alpha} N} \sum_{j:w_{j,t}=0} \mathbf{1}\{\|\alpha_i - \alpha_j\| \leq \delta h\} K(\hat{d}_{i,j}/h) \cdot u_{j,t}}{\frac{1}{h^{d_\alpha} N} \sum_{j:w_{j,t}=0} K(\hat{d}_{i,j}/h)}. \end{aligned} \quad (15)$$

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

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

We now consider the term in the second line (15). First consider the numerator. By Lemma A.4, there exist constants $0 < \underline{c} < \bar{c} < \infty$ such that, with probability approaching one, for all i , $\underline{c} h^{d_\alpha} \leq \frac{1}{N} \sum_{j:w_{j,t}=0} \mathbf{1}\{\|\alpha_i - \alpha_j\| \leq \delta h\} \leq \bar{c} h^{d_\alpha}$. Call the event that this holds E_h , then we have

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

By applying union bound to the first probability on the second line above, we obtain

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

Suppose that this E_h holds, it follows that

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

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 (iii), $\{u_{j,t}\}_{j=1}^N$ 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.8, it follows that for constants C, a , and b , for any $\eta > 0$,

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

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

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

And so, combining with our earlier result we get

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

By setting $\eta = \bar{c} \sqrt{\frac{Ma \log(N)}{ch^{d_\alpha} N}}$ in the above inequality for a fixed $M > 1$, and noting that $\eta \rightarrow 0$ and $\mathbb{P}_t(E_h) \rightarrow 1$, we obtain

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

and therefore we have

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

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

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

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

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

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

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

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

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

Therefore, combining the bias bound $L_0 \delta h$ on the first line of the decomposition with the numerator and denominator bounds just established, we conclude that

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

□

A.3 Proof of Theorem 2

Proof of Theorem 2. Recall the definition of the ATT estimate,

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

Using $(1 - w_{i,t})Y_{i,t} = (1 - w_{i,t})\mu_{i,t}^{(0)} + (1 - w_{i,t})u_{i,t}$ and $w_{i,t} = p_{i,t} + \epsilon_{i,t}$, we can express this as

$$\begin{aligned}\hat{\text{ATT}}_t &= \frac{1}{N_{1,t}} \sum_{i=1}^N \left(Y_{i,t} w_{i,t} - \frac{(1 - w_{i,t})Y_{i,t} p_{i,t} + (w_{i,t} - p_{i,t})\mu_{i,t}^{(0)}}{1 - p_{i,t}} \right) \\ &+ \frac{1}{N_{1,t}} \sum_{i=1}^N \frac{1}{1 - \hat{p}_{i,t}} (p_{i,t} - \hat{p}_{i,t})(\mu_{i,t}^{(0)} - \hat{\mu}_{i,t}^{(0)}) \\ &+ \frac{1}{N_{1,t}} \sum_{i=1}^N \epsilon_{i,t} \left(\frac{\mu_{i,t}^{(0)} - \hat{\mu}_{i,t}^{(0)}}{1 - \hat{p}_{i,t}} \right) \\ &+ \frac{1}{N_{1,t}} \sum_{i=1}^N \frac{(1 - w_{i,t})u_{i,t}}{(1 - p_{i,t})} \left(\frac{p_{i,t} - \hat{p}_{i,t}}{1 - \hat{p}_{i,t}} \right).\end{aligned}$$

We tackle each line separately.

By the upper bound in Assumption 4 (ii), for some $c > 0$, we have that $\frac{1}{1 - p_{i,t}} \leq 1/c$ with probability 1. And by supposition, we have $\max_{i \in [N]} |\hat{p}_{i,t} - p_{i,t}| = o_p(1)$ and so together with the previous bound we see that $\max_{i \in [N]} \left| \frac{1}{1 - \hat{p}_{i,t}} - \frac{1}{1 - p_{i,t}} \right| = o_p(1)$ and thus $\max_{i \in [N]} \frac{1}{1 - \hat{p}_{i,t}} = O_p(1)$. Additionally, by supposition $\max_{i \in [N]} |(p_{i,t} - \hat{p}_{i,t})(\mu_{i,t}^{(0)} - \hat{\mu}_{i,t}^{(0)})| = o_p(N^{-1/2})$. With the lower bound in Assumption 4 (ii) and the law of large numbers, $\frac{N}{N_{1,t}}$ is bounded above uniformly with probability approaching 1. Hence, it follows that

$$\frac{1}{N_{1,t}} \sum_{i=1}^N \frac{1}{1 - \hat{p}_{i,t}} (p_{i,t} - \hat{p}_{i,t})(\mu_{i,t}^{(0)} - \hat{\mu}_{i,t}^{(0)}) = o_p(N^{-1/2}).$$

Define $\mathcal{D}_{k,t} := \left(\left\{ \left\{ (Y_{j,s}(0), w_{j,s}) \right\}_{s=1}^t \right\}_{j \in \mathcal{I}_{-k}}, \left\{ \{Y_{i,s}\}_{s=1}^{T_0} \right\}_{i \in \mathcal{I}_k} \right)$. By Assumptions 2 and 4 (iii), for $t > T_0$ and $i \in \mathcal{I}_k$, we have $\epsilon_{i,t} \perp \mathcal{D}_{k,t} \mid \{\alpha_\ell\}_{\ell=1}^N, \lambda_t$ and this implies

$$\mathbb{E}[\epsilon_{i,t} \mid \mathcal{D}_{k,t}, \{\alpha_\ell\}_{\ell=1}^N, \lambda_t] = 0.$$

Thus, by the law of iterated expectations, and using the fact that $\hat{\mu}_{i,t}^{(0)}$ and $\hat{p}_{i,t}$ are non-stochastic given $\mathcal{D}_{k,t}$, we have

$$\begin{aligned}\mathbb{E} \left[\epsilon_{i,t} \left(\frac{\mu_{i,t}^{(0)} - \hat{\mu}_{i,t}^{(0)}}{1 - \hat{p}_{i,t}} \right) \mid \mathcal{D}_{k,t}, \{\alpha_\ell\}_{\ell=1}^N, \lambda_t \right] &= \mathbb{E}[\epsilon_{i,t} \mid \mathcal{D}_{k,t}, \{\alpha_\ell\}_{\ell=1}^N, \lambda_t] \left(\frac{\mu_{i,t}^{(0)} - \hat{\mu}_{i,t}^{(0)}}{1 - \hat{p}_{i,t}} \right) \\ &= 0.\end{aligned}$$

Also, by assumptions 2, 3, and 4 (iii), for any $i, j \in \mathcal{I}_k$ with $i \neq j$, we have $\epsilon_{i,t} \perp \epsilon_{j,t} \mid \{\alpha_\ell\}_{\ell=1}^N, \lambda_t, \mathcal{D}_{k,t}$.

Using this and the fact that $\epsilon_{i,t} \in [-1, 1]$ for all i , it follows that

$$\begin{aligned}\mathbb{E} \left[\epsilon_{i,t} \epsilon_{j,t} \left(\frac{\mu_{i,t}^{(0)} - \hat{\mu}_{i,t}^{(0)}}{1 - \hat{p}_{i,t}} \right) \left(\frac{\mu_{j,t}^{(0)} - \hat{\mu}_{j,t}^{(0)}}{1 - \hat{p}_{j,t}} \right) \mid \mathcal{D}_{k,t}, \{\alpha_\ell\}_{\ell=1}^N, \lambda_t \right] &= \mathbb{E}[\epsilon_{i,t} \epsilon_{j,t} \mid \mathcal{D}_{k,t}, \{\alpha_\ell\}_{\ell=1}^N, \lambda_t] \left(\frac{\mu_{i,t}^{(0)} - \hat{\mu}_{i,t}^{(0)}}{1 - \hat{p}_{i,t}} \right) \left(\frac{\mu_{j,t}^{(0)} - \hat{\mu}_{j,t}^{(0)}}{1 - \hat{p}_{j,t}} \right) \\ &\leq \mathbf{1}\{j = i\} \cdot \left(\frac{\mu_{i,t}^{(0)} - \hat{\mu}_{i,t}^{(0)}}{1 - \hat{p}_{i,t}} \right)^2.\end{aligned}$$

Using the results above and Markov's inequality, we have

$$\mathbb{P}\left(\left|\frac{1}{\sqrt{|\mathcal{I}_k|}} \sum_{i \in \mathcal{I}_k} \epsilon_{i,t} \left(\frac{\mu_{i,t}^{(0)} - \hat{\mu}_{i,t}^{(0)}}{1 - \hat{p}_{i,t}}\right)\right| > \epsilon \mid \mathcal{D}_{k,t}, \{\alpha_i\}_{i=1}^N, \lambda_t\right) \leq \frac{\frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} \left(\frac{\mu_{i,t}^{(0)} - \hat{\mu}_{i,t}^{(0)}}{1 - \hat{p}_{i,t}}\right)^2}{\epsilon^2} = o_p(1),$$

where the final equality uses $\max_i |\hat{\mu}_{i,t}^{(0)} - \mu_{i,t}^{(0)}| = o_p(1)$ together with $\max_i 1/(1 - \hat{p}_{i,t}) = O_p(1)$.

Hence, by using $|\mathcal{I}_k| \asymp N$, it follows that

$$\frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} \epsilon_{i,t} \left(\frac{\mu_{i,t}^{(0)} - \hat{\mu}_{i,t}^{(0)}}{1 - \hat{p}_{i,t}}\right) = o_p(N^{-1/2}).$$

Since the number of folds $\mathcal{K}$ is fixed, summing over $k \in [\mathcal{K}]$ and using $N/N_{1,t} = O_p(1)$ gives

$$\frac{1}{N_{1,t}} \sum_{i=1}^N \epsilon_{i,t} (\mu_{i,t}^{(0)} - \hat{\mu}_{i,t}^{(0)}) / (1 - \hat{p}_{i,t}) = o_p(N^{-1/2}).$$

Similarly, by Assumptions 2 and 4 (iii), we have $u_{i,t} \perp \mathcal{D}_{k,t} \mid \{\alpha_l\}_{l=1}^N, \lambda_t, w_{i,t} = 0$, and for any $i, j \in \mathcal{I}_k$ with $i \neq j$,

$$u_{i,t} \perp u_{j,t} \mid \{\alpha_l\}_{l=1}^N, \lambda_t, \mathcal{D}_{k,t}, w_{i,t} = w_{j,t} = 0.$$

Therefore, by the law of iterated expectations, alongside the fact that $\hat{p}_{i,t}$ is non-stochastic given $\mathcal{D}_{k,t}$, it follows that

$$\begin{aligned} \mathbb{E}\left[\frac{u_{i,t}}{(1 - p_{i,t})} \left(\frac{p_{i,t} - \hat{p}_{i,t}}{1 - \hat{p}_{i,t}}\right) \mid \mathcal{D}_{k,t}, \{\alpha_l\}_{l=1}^N, \lambda_t, w_{i,t} = 0\right] &= \mathbb{E}\left[u_{i,t} \mid \{\alpha_l\}_{l=1}^N, \lambda_t, w_{i,t} = 0\right] \frac{1}{1 - p_{i,t}} \left(\frac{p_{i,t} - \hat{p}_{i,t}}{1 - \hat{p}_{i,t}}\right) \\ &= 0 \end{aligned}$$

Moreover, we have

$$\begin{aligned} &\mathbb{E}\left[\frac{u_{i,t}}{(1 - p_{i,t})} \frac{u_{j,t}}{(1 - p_{j,t})} \left(\frac{p_{i,t} - \hat{p}_{i,t}}{1 - \hat{p}_{i,t}}\right) \left(\frac{p_{j,t} - \hat{p}_{j,t}}{1 - \hat{p}_{j,t}}\right) \mid \mathcal{D}_{k,t}, \{\alpha_k\}_{k=1}^N, \lambda_t, w_{i,t} = w_{j,t} = 0\right] \\ &= \mathbb{E}\left[\frac{u_{i,t}}{(1 - p_{i,t})} \frac{u_{j,t}}{(1 - p_{j,t})} \mid \mathcal{D}_{k,t}, \{\alpha_k\}_{k=1}^N, \lambda_t, w_{i,t} = w_{j,t} = 0\right] \left(\frac{p_{i,t} - \hat{p}_{i,t}}{1 - \hat{p}_{i,t}}\right) \left(\frac{p_{j,t} - \hat{p}_{j,t}}{1 - \hat{p}_{j,t}}\right) \\ &\leq \mathbf{1}\{j = i\} \cdot \mathbb{E}\left[u_{i,t}^2 \mid w_{i,t} = 0, \{\alpha_k\}_{k=1}^N, \lambda_t\right] \frac{1}{(1 - p_{i,t})^2} \left(\frac{p_{i,t} - \hat{p}_{i,t}}{1 - \hat{p}_{i,t}}\right)^2 \\ &\leq \mathbf{1}\{j = i\} \cdot \bar{\sigma}_Y^2 \cdot \frac{1}{(1 - p_{i,t})^2} \left(\frac{p_{i,t} - \hat{p}_{i,t}}{1 - \hat{p}_{i,t}}\right)^2, \end{aligned}$$

for some finite constant $\bar{\sigma}_Y^2$, which exists by Assumption 4 (iii) (by standard properties of sub-Gaussian random variables $u_{i,t}$ has conditional variance at most 2δ where δ is the 'variance proxy' referred to in the assumption). Thus, by Markov's inequality, it follows that

$$\mathbb{P}\left(\left|\frac{1}{\sqrt{|\mathcal{I}_k|}} \sum_{i \in \mathcal{I}_k} (1 - w_{i,t}) \frac{u_{i,t}}{1 - p_{i,t}} \left(\frac{p_{i,t} - \hat{p}_{i,t}}{1 - \hat{p}_{i,t}}\right)\right| > \epsilon \mid \mathcal{D}_{k,t}, \{\alpha_i\}_{i=1}^N, \lambda_t\right) \leq \frac{\frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} \bar{\sigma}_Y^2 \frac{1}{(1 - p_{i,t})^2} \left(\frac{p_{i,t} - \hat{p}_{i,t}}{1 - \hat{p}_{i,t}}\right)^2}{\epsilon^2} = o_p(1)$$

The conditional probability is $o_p(1)$ which implies that the unconditional probability is $o(1)$. Hence, we have

$$\frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} (1 - w_{i,t}) \frac{u_{i,t}}{(1 - p_{i,t})} \left(\frac{p_{i,t} - \hat{p}_{i,t}}{1 - \hat{p}_{i,t}} \right) = o_p(N^{-1/2}).$$

By combining all previous results, we have shown that

$$\sqrt{N} \hat{\text{ATT}}_t = \frac{\sqrt{N}}{N_{1,t}} \sum_{i=1}^N \left(Y_{i,t} w_{i,t} - \frac{(1 - w_{i,t}) Y_{i,t} p_{i,t} + (w_{i,t} - p_{i,t}) \mu_{i,t}^{(0)}}{1 - p_{i,t}} \right) + o_p(1).$$

The remaining steps are a standard application of asymptotic theory for the ATT estimate with known nuisance functions, and so we omit them here. $\square$

A.4 Bernstein Inequalities

In this section, we specify the Bernstein inequalities, which we refer to in the proofs.

A.4.1 Bernstein inequality for bounded random variables

Lemma A.5 (Bernstein inequality for bounded random variables; 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 $\epsilon > 0$*

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

A.4.2 Bernstein inequality for unbounded random variables

Lemma A.6 (Moments of a sub-Gaussian random variable). *Let Z be a mean zero random variable satisfying $\mathbb{E}[e^{\lambda Z}] \leq e^{v\lambda^2/2}$ for all $\lambda \in \mathbb{R}$, for some $v > 0$. Then for every integer $q \geq 1$, we have $\mathbb{E}[Z^{2q}] \leq q!(4v)^q$.*

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

Lemma A.7 (Bernstein inequality for unbounded random variables). *Let $Z_1, \dots, Z_n$ be independent random variables. Assume that there exist some positive constants v and c such that $\frac{1}{n} \sum_{i=1}^n \mathbb{E}[Z_i^2] \leq v$ and for all integers $q \geq 3$,*

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

Then for all $\epsilon > 0$,

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

Proof of Lemma A.7. Apply the one-sided bound in Corollary 2.11 in [Boucheron et al. \(2013\)](#) to $Z_1, \dots, Z_n$ and to $-Z_1, \dots, -Z_n$, with variance parameter $n\nu$ and threshold $n\epsilon$. The moment assumptions hold for both signs, so the union bound gives the stated two-sided inequality, with the factor of two accounting for the two tails. $\square$

Lemma A.8 (Bernstein inequality for sub-Gaussian random variables). *Let $Z_1, \dots, Z_n$ be mean-zero independent random variables. Assume that there exists some $\nu > 0$ such that $\mathbb{E}[e^{\lambda Z_i}] \leq e^{\nu \lambda^2/2}$ for all $\lambda \in \mathbb{R}$ and for all $i \in \{1, \dots, n\}$. Fix $\bar{\beta} \in (0, \infty)$. Then there exist positive constants C, a , and b , depending only on ν and $\bar{\beta}$, such that for all constants $\beta_1, \dots, \beta_n$ satisfying $\max_i |\beta_i| < \bar{\beta}$ and for all $\epsilon > 0$,*

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

and

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

Additionally, let $W_1, \dots, W_n$ be mean-zero random variables such that the pairs (Z_i, W_i) are independent across i , $\mathbb{E}[e^{\lambda W_i}] \leq e^{\nu \lambda^2/2}$ for all $\lambda \in \mathbb{R}$ and all i , and $\mathbb{E}[Z_i W_i] = 0$ for all i . Then, after enlarging C, a , and b if necessary, for all $\epsilon > 0$,

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

Proof of Lemma A.8. The first inequality follows from the sub-Gaussian tail bound in Section 2.3 of [Boucheron et al. \(2013\)](#), since $n^{-1} \sum_{i=1}^n \beta_i Z_i$ has variance factor at most $\nu \bar{\beta}^2/n$.

For the second inequality, Lemma A.6 gives $\mathbb{E}[|\beta_i Z_i^2|^q] \leq q!(4\nu \bar{\beta})^q$ for every integer $q \geq 2$, so Lemma A.7 applies to $\beta_i Z_i^2$.

For the third inequality, Cauchy–Schwarz and Lemma A.6 give

$$\mathbb{E}[|Z_i W_i|^q] \leq (\mathbb{E}[Z_i^{2q}] \mathbb{E}[W_i^{2q}])^{1/2} \leq q!(4\nu)^q$$

for every integer $q \geq 2$. Apply Lemma A.7 to the independent, mean-zero products $Z_i W_i$. The constants in all three bounds can be chosen to depend only on ν and $\bar{\beta}$. $\square$