Causal Inference in Possibly Nonlinear Factor Models [*]

Yingjie Feng[†]

April 30, 2024

Abstract

This paper develops a general causal inference method for treatment effects models with noisily measured confounders. The key feature is that a large set of noisy measurements are linked with the underlying latent confounders through an unknown, possibly nonlinear factor structure. The main building block is a local principal subspace approximation procedure that combines K-nearest neighbors matching and principal component analysis. Estimators of many causal parameters, including average treatment effects and counterfactual distributions, are constructed based on doubly-robust score functions. Large-sample properties of these estimators are established, which only require relatively mild conditions on informativeness of noisy measurements and local principal subspace approximation. The results are illustrated with an empirical application studying the effect of political connections on stock returns of financial firms, and a Monte Carlo experiment.

Keywords: causal inference, latent confounders, noisy measurements, nonlinear factor model, heterogeneous treatment effects, doubly-robust estimator, high-dimensional data

JEL: C14, C21, C23, C38

[*] We especially thank Matias Cattaneo for helpful comments and suggestions. We also thank Sebastian Calonico, Richard Crump, Jianqing Fan, Max Farrell, Andreas Hagemann, Michael Jansson, Lutz Kilian, Xinwei Ma, Kenichi Nagasawa, Rocío Titiunik, Gonzalo Vazquez-Bare, and Jingshen Wang for their valuable feedback. Feng gratefully acknowledges financial support from the National Natural Science Foundation of China (NSFC) through grants 72203122, 72133002, and 72250064.

[†] Department of Economics, School of Economics and Management, Tsinghua University, Beijing 100084, China.

1 Introduction

Understanding effects of policy interventions is central in many disciplines. When observational data are used, researchers usually confront the challenge that the treatment is nonrandomly assigned based on some characteristics that are *not* directly observed. The confounding effects of these variables (confounders) make it difficult to uncover the true causal relation between the outcome and the treatment. Commonly used econometric methods that assume selection on observables are inappropriate in this situation. This paper considers a treatment effects model in which a *large* set of observed covariates, as the noisy measurements of the underlying confounders, are available. The key assumption is that the observed measurements and unobserved confounders are linked via an unknown, possibly nonlinear factor model. The former, though not affecting the potential outcome and the treatment assignment directly, provide information on the latter, thus making it possible to resolve the confounding issue.

As an example, consider the effect of a scholarship on academic performance of newly admitted college students. One may be concerned about the confounding effect of the latent precollege ability, since it may correlate with both a student's likelihood of getting a scholarship and her future academic performance. If the researcher observes the same student taking multiple tests in different subjects or time periods at the precollege stage, these past test scores may play the role of the noisy measurements of the unobserved ability. The nonlinear factor structure allows for a flexible latent relationship between ability and test outcomes, which may vary across subjects or time in a complex way. The idea of using noisy proxies to measure ability or skills is the key to many econometric analyses. It forms the foundation of estimating skill production function, evaluating early childhood investment policies and identifying causal effects of education on earnings, health or other outcomes (see, e.g., Cunha et al., 2010; Heckman et al., 2018).

In this paper we develop a novel inference procedure for counterfactual analysis, which builds on a local principal component analysis (PCA) method developed in Feng (2023) to "extract" information on latent confounders from the noisy measurements. We first find the K-nearest neighbors (K-NN) for each unit based on the observed measurements, with K diverging as the sample size grows. If different values of latent confounders can induce non-negligible differences in many observed measurements, then the K nearest neighbors, appropriately measured by the noisy measurements, should also be close in terms of the latent confounders. Next, within each local neighborhood formed by the K matches, under mild regularity conditions, the underlying possibly nonlinear factor structure is approximated by a linear factor structure, which can be estimated by principal component analysis.

As in linear factor models (Bai, 2003), the values of the latent confounders cannot be exactly recovered without additional normalizations. Nevertheless, employing the results in Feng (2023), we can show that the nearest neighbors and estimated local factor loadings from local PCA characterize the latent confounders and suffice to restore unconfoundedness in our treatment effects analysis. Specifically, we propose a local quasi-maximum likelihood method to estimate the conditional means of potential outcomes and the conditional treatment probabilities (generalized propensity scores), which form the basis of regression imputation and propensity score weighting estimators of treatment effects. The local region used in such estimation is defined by nearest neighbors, and the extracted local factor loadings play the role of generated regressors that provide further approximation to unknown conditional expectation functions of interest. The number of nearest neighbors, the main tuning parameter of our procedure, implicitly governs the "bandwidth" of the estimation and determines the consistency of final estimators, whereas the number of local factor loadings extracted is analogous to the degree of the basis in local polynomial regression and can be taken as fixed in practice.

In contrast with standard nonparametric regression analysis, the conditioning variables in this scenario are *indirectly* obtained from the observed measurements, and the noise in their factor structure restricts one's ability to select a bandwidth. Using a small number of nearest neighbors does *not* necessarily lead to a small bandwidth and thus is not helpful for further bias reduction, which differs from other matching techniques based on a fixed number of matches (e.g., Abadie and Imbens, 2006). Consequently, the possibly large smoothing bias of the nonparametric ingredients may render the final inference on causal parameters invalid. To deal with this issue, we follow the Neyman-orthogonalization strategy that has been extensively applied in the recent double/debiased machine learning literature (Belloni et al., 2014; Farrell, 2015; Chernozhukov et al., 2018, 2022). In treatment effects models, the widely used doubly-robust scores (Robins and Rotnitzky, 1995; Cattaneo, 2010) are estimating equations constructed based on the efficient influence function and are automatically Neyman orthogonal. Taking advantage of this property, we can conduct valid inference under mild restrictions on the number of nearest neighbors.

Leveraging these ideas, we develop a novel estimation and inference procedure for a large class of estimands, including counterfactual distributions and functionals thereof, and provide the basis for analyzing many causal quantities of interest such as average, quantile, and distributional treatment effects. It has several appealing features in theory and practice. First, as a dimension reduction technique, the proposed method allows users to obtain low-dimensional information on latent confounders from large-dimensional noisy measurements. It only requires some but *not all* measurements to be informative about latent confounders, and it is *unnecessary* to know their identities a priori (see Remark 4.1 below). Second, the proposed method does *not* impose a functional form assumption on the relationship between latent confounders and noisy measurements, thus making the final inference more robust. In particular, the nonlinearity of this relationship is allowed but not assumed, and the classical linear factor model can be covered as a special case. Third, our theory builds on a generic choice of the "distance" to define nearest neighbors, accommodating and extending previous matching strategies suggested in the panel and network data literature (e.g. Zhang et al., 2017; Bonhomme et al., 2022). Our theoretical and numerical results also show that the proposed method based on local PCA provides more flexible approximation to the nonlinear factor structure and delivers more robust inference results, compared with local constant smoothing based solely on matching techniques. Finally, the output of local PCA can be readily taken as input to many classical econometric estimation procedures such as local polynomial kernel regression (Fan and Gijbels, 1996), which may be useful in other contexts such as diffusion index forecasts (Stock and Watson, 2002; Bai and Ng, 2006), where prediction based on noisily measured variables is of interest.

The paper is organized as follows. The rest of this section discusses the related literature. In Section 2, we set up a multi-valued treatment effects model and describe the nonlinear factor structure of the large-dimensional measurements of latent confounders. Section 3 gives a detailed description of the entire estimation procedure, accompanied by a step-by-step empirical illustration using the data of Acemoglu et al. (2016). Section 4 presents the main theoretical results and some Monte Carlo evidence. Section 5 extends our theory to uniform inference on counterfactual distributions. Section 6 concludes. The Supplemental Appendix contains all theoretical proofs and additional technical results. Replications of the simulation study and empirical illustration are available at https://github.com/yingjieum/replication-Feng_2024.

1.1 Related Literature

This paper contributes to several strands of literature. First, the observed covariates may be viewed as an array of noisy measurements of the latent confounders, and thus the theoretical framework in this paper is closely related to nonlinear models with measurement errors. Much effort has been devoted to the identification of such models (see Schennach, 2016 for a review). For example, factor models can be utilized to construct repeated measurements of unobserved variables, which allows for the identification of their distribution under suitable normalizations. A general treatment following this strategy is available in Cunha et al. (2010), using and extending results in Hu and Schennach (2008). This paper takes a different route. A large-dimensional nonlinear factor model is exploited to directly extract the geometric relation among different units in terms of the latent variables, which is then used to control for their confounding effects in the treatment effects analysis. Some measurements are allowed to be uninformative about the latent confouders, and to identify the causal effect of interest, it is unnecessary to recover the exact values (or distributions) of latent confounders. Conceptually, the extracted information from the observables plays a similar role as a control function, conditional on which the treatment assignment is no longer confounded. See, e.g., Altonji and Mansfield (2018), Miao et al. (2018), and Nagasawa (2022) for causal effects identification that apply the idea of using noisy proxies to control for unobservables.

Second, this study contributes to the existing literature on causal inference and program evaluation (see Abadie and Cattaneo, 2018 for a review). In particular, it is connected with the fast-growing literature on synthetic control (see Abadie, 2021 and references therein) and staggered adoption designs (Athey and Imbens, 2022). The classical synthetic control method and many variants thereof are often motivated by assuming a linear factor structure for the pre-treatment data. By contrast, this paper allows for a possibly nonlinear factor structure and does not rely on the strong assumption of linear factor models. Using the extracted information on latent confounders, we derive formal large-sample properties of the proposed estimators under mild side conditions, which can be applied to (but is not restricted to) synthetic control problems with disaggregated data.

Third, the local PCA method, the technical building block of this paper, is developed in Feng (2023), which builds on and extends results on large-dimensional factor analysis and panel data models with fixed effects (Bai, 2009; Bai and Wang, 2016; Wang and Fan, 2017). See more references and related discussions therein. However, unlike this paper, Feng (2023) focuses on the optimal matrix estimation problem in the nonlinear factor setting. A recent study by Bonhomme et al. (2022) develops two-step grouped fixed-effects estimators that discretize latent heterogeneity by K-means clustering, which relies on a specific injective moment condition to ensure the informativeness of the measurements. By contrast, this paper relies on a more general informativeness requirement (see Assumption 3 and Remark

4.1 for details) and can achieve more flexible approximation of smooth functions of latent features via the local principal subspace approximation strategy. The intermediate result (Theorem 4.1) also characterizes the uniform convergence of nonparametric estimators of individual-specific features and may be of independent interest.

Finally, we also note that the idea of locally approximating a latent nonlinear "surface" embedded in a high-dimensional space has been widely used in the modern machine learning literature (e.g., Zhang and Zha, 2004; Arias-Castro et al., 2017). These methods are used to construct a global nonlinear structure that preserve the local geometry of the data for the purpose of classification, clustering or data visualization. Unlike these studies, this paper focuses on estimation and inference of causal parameters in the treatment effects model rather than recovering the latent surface.

2 Treatment Effects Model with Latent Variables

Suppose that a random sample $\{(y_i, t_i, \boldsymbol{x}_i, \boldsymbol{z}_i)\}_{i=1}^n$ is available, where $y_i \in \mathbb{R}$ is the outcome of interest, $t_i \in \mathcal{J} = \{0, \cdots, J\}$ denotes the multi-valued treatment status, and $\boldsymbol{x}_i \in \mathbb{R}^p$ and $\boldsymbol{z}_i \in \mathbb{R}^{\mathsf{d}_z}$ are vectors of covariates. $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ play different roles in later analysis: $\boldsymbol{x}_i$ is a vector of noisy measurements of some *unobserved* confounders $\boldsymbol{\alpha}_i \in \mathbb{R}^{\mathsf{d}_\alpha}$, whereas $\boldsymbol{z}_i$ itself is a vector of *observed* confounders that can be controlled for directly. Some covariates may be used for both purposes simultaneously, making $\boldsymbol{z}_i$ and $\boldsymbol{x}_i$ share some variables in common. The asymptotic theory in this paper is developed assuming the sample size n and the number of noisy measurements p simultaneously increase to infinity whereas the number of confounders $\mathsf{d}_\alpha + \mathsf{d}_z$ are fixed.

We follow the standard potential outcomes framework. Let $y_i(t)$ denote the potential outcome of unit i at treatment level $t \in \mathcal{J}$. The observed outcome can be written as $y_i = \sum_{t=0}^J d_i(t) y_i(t)$ where $d_i(t) = \mathbb{1}(t_i = t)$ is an indicator for each treatment level $t \in \mathcal{J}$. The key challenge for causal analysis is the missing data issue. For example, when t_i is binary

($t_i \in \{0,1\}$), the average treatment effects on the treated (ATT) relies on $\mathbb{E}[y_i(0)|t_i = 1]$, but $y_i(0)$ is unobserved for the treated group. This hurdle is often overcome by imposing an unconfoundedness condition so that the treatment assignment becomes independent of potential outcomes after conditioning on a set of observed covariates. By contrast, this paper assumes that

$$y_i(t) \perp\!\!\!\perp d_i(t') \,|\, \boldsymbol{\alpha}_i, \boldsymbol{z}_i \quad \forall t, t' \in \mathcal{J}. \tag{2.1}$$

While $\boldsymbol{z}_i$ is observed, $\boldsymbol{\alpha}_i$ is unobserved and thus cannot be directly controlled for in causal inference. However, as described later in Section 2.1, the unconfoundedness can be restored when we have a vector of noisy measurements $\boldsymbol{x}_i$ of $\boldsymbol{\alpha}_i$.

For each treatment level $t \in \mathcal{J}$, the outcome of interest is characterized by a possibly nonlinear, reduced-form model:

$$y_i(t) = \zeta_{i,t} + \epsilon_{i,t}, \qquad \zeta_{i,t} = \psi_{\mathsf{y}}(\mu_t(\boldsymbol{\alpha}_i) + \boldsymbol{z}_i'\boldsymbol{\beta}_t), \qquad \mathbb{E}[\epsilon_{i,t}|\boldsymbol{z}_i, \boldsymbol{\alpha}_i] = 0, \tag{2.2}$$

where $\zeta_{i,t}$ is the conditional expectation of the potential outcome at treatment level t given the observed $\boldsymbol{z}_i$ and unobserved $\boldsymbol{\alpha}_i$, and $\psi_{\mathsf{y}}^{-1}(\cdot) : \mathbb{R} \mapsto \mathbb{R}$ is a (known) link function associated with the outcome equation. Let $\boldsymbol{\epsilon}_i = (\epsilon_{i,0}, \cdots, \epsilon_{i,J})$.

On the other hand, given a (known) link function $\boldsymbol{\psi}_{\mathsf{p}}^{-1}(\cdot) : (0,1)^{J+1} \mapsto \mathbb{R}^J$ associated with the treatment equation, setting $t = 0$ as the base level, the assignment mechanism is described by

$$\boldsymbol{d}_i = \boldsymbol{\varrho}_i + \boldsymbol{v}_i, \qquad \boldsymbol{\varrho}_i = \boldsymbol{\psi}_{\mathsf{p}}(\boldsymbol{e}(\boldsymbol{\alpha}_i) + \boldsymbol{\Gamma}\boldsymbol{z}_i), \qquad \mathbb{E}[\boldsymbol{v}_i|\boldsymbol{z}_i, \boldsymbol{\alpha}_i] = 0, \tag{2.3}$$

where $\boldsymbol{d}_i = (d_i(0), \cdots, d_i(J))'$, $\boldsymbol{\varrho}_i = (\varrho_{i,0}, \cdots, \varrho_{i,J})'$, $\boldsymbol{v}_i = (v_{i,0}, \cdots, v_{i,J})'$, $\boldsymbol{e}(\cdot) = (e_1(\cdot), \cdots, e_J(\cdot))'$, and $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_J)'$ with $\boldsymbol{\gamma}_t \in \mathbb{R}^{\mathsf{d_z}}$ for each $t = 1, \cdots, J$. Thus, each $\varrho_{i,t}$ is the conditional probability of treatment level t, which would be the usual propensity score if $\boldsymbol{\alpha}_i$ were observed.

Note that $\boldsymbol{z}_i$ and $\boldsymbol{\alpha}_i$ are assumed to enter the two equations simultaneously and thus play the role of confounders, which is consistent with the conditional independence assumption (2.1). For simplicity, $\zeta_{i,t}$ and $\varrho_{i,t}$ are assumed to take generalized partially linear forms: $\boldsymbol{\alpha}_i$ enters the model nonparametrically through the unknown functions $\mu_t(\cdot)$ and $e_t(\cdot)$, whereas $\boldsymbol{z}_i$ enters the model in an additive-separable way. This specification allows us to flexibly control for the unobserved confounders, which is the focus of this paper, but still maintain practical tractability, compared to fully nonparametric models.

2.1 Structure of Large-Dimensional Measurements

The observed covariates $\boldsymbol{x}_i = (x_{i1}, \cdots, x_{ip})'$ play the role of noisy measurements of latent confounders $\boldsymbol{\alpha}_i$. In general, we can consider a covariates-adjusted nonlinear factor model for $\boldsymbol{x}_i$:

$$\boldsymbol{x}_i = \boldsymbol{W}_i\boldsymbol{\vartheta} + \boldsymbol{\eta}(\boldsymbol{\alpha}_i) + \boldsymbol{u}_i, \quad \mathbb{E}[\boldsymbol{u}_i|\mathcal{F}, \{\boldsymbol{W}_i\}_{i=1}^n] = 0, \quad 1 \leq i \leq n, \tag{2.4}$$

where $\boldsymbol{W}_i = (\boldsymbol{w}_{i,1}, \cdots, \boldsymbol{w}_{i,\mathsf{d}_w}) \in \mathbb{R}^{p\times \mathsf{d}_w}$ is a matrix of covariates with the slope parameter $\boldsymbol{\vartheta} \in \mathbb{R}^{\mathsf{d}_w}$, $\boldsymbol{\eta}(\cdot) = (\eta_1(\cdot), \cdots, \eta_p(\cdot))' : \mathbb{R}^{\mathsf{d}_\alpha} \mapsto \mathbb{R}^p$ is a vector of latent functions, and $\boldsymbol{u}_i = (u_{i1}, \cdots, u_{ip})'$ is a vector of idiosyncratic errors. We let $\mathcal{F}$ be the σ-field generated by unobserved random elements $\{\boldsymbol{\alpha}_i\}_{i=1}^n$ and $\boldsymbol{\eta}(\cdot)$. The regressors included in $\boldsymbol{W}_i$ need to be sufficiently *high-rank* (enough variation across both dimensions) for identification of $\boldsymbol{\vartheta}$.

Since incorporating $\boldsymbol{W}_i$ is notationally cumbersome and less relevant to the core idea of this paper, we consider a simplified model with $\boldsymbol{\vartheta} = \boldsymbol{0}$ hereafter:

$$\boldsymbol{x}_i = \boldsymbol{\eta}(\boldsymbol{\alpha}_i) + \boldsymbol{u}_i, \quad \mathbb{E}[\boldsymbol{u}_i|\mathcal{F}] = 0. \tag{2.5}$$

Define $p \times n$ matrices $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)$, $\boldsymbol{H} = (\boldsymbol{\eta}(\boldsymbol{\alpha}_1), \cdots, \boldsymbol{\eta}(\boldsymbol{\alpha}_n))$ and $\boldsymbol{U} = (\boldsymbol{u}_1, \cdots, \boldsymbol{u}_n)$. Then, (2.5) can be written in matrix form: $\boldsymbol{X} = \boldsymbol{H} + \boldsymbol{U}$. The Supplemental Appendix describes how the estimation procedure in Section 3 below can be adjusted when additional covariates $\boldsymbol{W}_i$ are included, and formal large-sample theory for this general case is available in Feng (2023).

Throughout the paper, the latent variables $\{\boldsymbol{\alpha}_i\}_{i=1}^n$ and the latent functions $\boldsymbol{\eta}(\cdot)$ are understood as random elements, but our analysis is conducted *conditional* on them. In this sense, they are analogous to *fixed effects* in the panel data literature. This setup indeed encompasses many panel data models as special cases. For instance, if we assume $\eta_l(\boldsymbol{\alpha}_i) = \boldsymbol{\varpi}_l'\boldsymbol{\alpha}_i$ for some $\boldsymbol{\varpi}_l \in \mathbb{R}^{\mathsf{d}_\alpha}$, Equation (2.5) reduces to the classical linear factor model (Bai, 2003). Instead of restricting the latent mean structure $\boldsymbol{H}$ to be *exactly* low-rank, (2.5) allows $\boldsymbol{H}$ to be full rank due to the potential nonlinearity of the latent functions $\boldsymbol{\eta}(\cdot)$, while the variation of the large-dimensional $\boldsymbol{X}$ may still be explained by a few low-dimensional components in a possibly nonlinear way.

2.2 Notation

Matrices. For a vector $\boldsymbol{v} \in \mathbb{R}^{\mathsf{d}}$, $\|\boldsymbol{v}\| = \sqrt{\boldsymbol{v}'\boldsymbol{v}}$ is the Euclidean norm of $\boldsymbol{v}$, and for an $m \times n$ matrix $\boldsymbol{A}$, $\|\boldsymbol{A}\|_{\max} = \max_{1\leq i\leq m, 1\leq j\leq n} |a_{ij}|$ is the entrywise sup-norm of $\boldsymbol{A}$. $s_{\max}(\boldsymbol{A})$ and $s_{\min}(\boldsymbol{A})$ denote the largest and smallest singular values of $\boldsymbol{A}$ respectively. Moreover, $\boldsymbol{A}_{i\cdot}$ and $\boldsymbol{A}_{\cdot j}$ denote the ith row and the jth column of $\boldsymbol{A}$ respectively, and $\mathbf{1}_{\mathsf{d}}$ denotes a d-vector of ones.

Asymptotics. For sequences of numbers or random variables, $a_n \lesssim b_n$ denotes $\limsup_n |a_n/b_n|$ is finite, and $a_n \lesssim_{\mathbb{P}} b_n$ denotes $\limsup_{\varepsilon\to\infty} \limsup_n \mathbb{P}[|a_n/b_n| \geq \varepsilon] = 0$. $a_n = o(b_n)$ implies $a_n/b_n \to 0$, and $a_n = o_{\mathbb{P}}(b_n)$ implies that $a_n/b_n \to_{\mathbb{P}} 0$, where $\to_{\mathbb{P}}$ denotes convergence in probability. $a_n \asymp b_n$ implies that $a_n \lesssim b_n$ and $b_n \lesssim a_n$. $\rightsquigarrow$ denotes convergence in distribution. Moreover, for possibly random sequence $\{a_{i,n}\}_{i\in[n]}$ and a non-random strictly positive sequence $\{r_{i,n}\}_{i\in[n]}$, we write $a_{i,n} = \bar{O}_{\mathbb{P}}(r_{i,n})$ if $\max_{i\in[n]} |a_{i,n}/r_{i,n}| \lesssim_{\mathbb{P}} 1$, and $a_{i,n} = \bar{o}_{\mathbb{P}}(r_{i,n})$ if $\max_{i\in[n]} |a_{i,n}/r_{i,n}| \to_{\mathbb{P}} 0$.

Others. For two numbers a and b, $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. For a finite set $\mathcal{S}$, $|\mathcal{S}|$ denotes its cardinality. For an integer $m > 0$, $[m] = \{1, 2, \cdots, m\}$. For a sequence $\{a_{i,n} : i \in [n]\}$, $\mathbb{E}_n[a_{i,n}] = \frac{1}{n}\sum_{i=1}^n a_{i,n}$.

3 Outline of Estimation Procedure

This section describes the main procedure for treatment effects analysis, which consists of three steps. First, given the structure (2.5), relevant information on $\boldsymbol{\alpha}_i$ is extracted from $\boldsymbol{x}_i$. Second, the conditional means $\{\zeta_{i,t}\}_{i=1}^n$ of potential outcomes and conditional treatment probabilities $\{\varrho_{i,t}\}_{i=1}^n$ are estimated by a local quasi-maximum likelihood method where the extracted information from the first step plays the role of kernel functions and generated regressors. Third, estimators of causal estimands of interest are constructed based on doubly-robust score functions. See Algorithm 1 for a short summary. The main tuning parameter in this procedure is the number of nearest neighbors K, which governs the "bandwidth" of the nonparametric estimation in the second step, while the number of (local) principal components to be extracted can be taken as fixed.

In addition to methodological discussions, each step below will be accompanied by an empirical illustration using the data of Acemoglu et al. (2016), which analyzes the effect of the announcement of the appointment of Tim Geithner as Treasury Secretary on November 21, 2008 on stock returns of financial firms that were connected to him. This study can be viewed as an example of the synthetic control design in the program evaluation literature (see Abadie, 2021 for a review). Specifically, the treatment of interest is the appointment of Geithner, which starts at a particular date (referred to as "event day 0" hereafter). All firms remain untreated prior to the appointment. Starting at event day 0, a subgroup of firms that are connected to Geithner are treated ($t_i = 1$), while the other group remains untreated ($t_i = 0$). Variables used in this analysis and the parameter of interest are listed in the following.

- Potential outcomes $y_i(1)$ and $y_i(0)$: the cumulative stock returns of firm i from date 0 to date 1 that would be observed with and without Geithner connections;
- Noisy measurements $\boldsymbol{x}_i$: the daily stock returns of firm i prior to the Geithner announcement;

- Additional controls $\boldsymbol{z}_i$: the size (log of total assets), profitability (return on equity), and leverage (total debt to total capital) of firm i as of 2008;
- Parameter of interest $\mathbb{E}[y_i(1) - y_i(0)|t_i = 1]$: the average cumulative abnormal returns of firms connected to Geithner from date 0 to date 1.

The sample consists of 583 firms in total ($n = 583$) and 22 of them are treated ("connected to Geithner"). To be comparable with the results in Acemoglu et al. (2016), the observed measurements $\boldsymbol{x}_i$ only include the daily stock returns for 250 days that ends 30 days prior to the Geithner announcement ($p = 250$). In this application, the "pre-treatment" returns can be viewed as noisy measurements of some unobserved firm characteristics $\boldsymbol{\alpha}_i$, such as risk preference or management level.

3.1 Step 1: Latent Variables Extraction

The goal is to extract information on latent confounders $\boldsymbol{\alpha}_i$ by employing the structure (2.5). The main ideas are sketched below.

Row-wise Splitting. To guarantee desired theoretical properties of local PCA, Feng (2023) recommends users separate the K-NN matching and principal component analysis. Specifically, split the row index set $\mathcal{R} = [p]$ of $\boldsymbol{X}$ into two non-overlapping subsets: $\mathcal{R} = \mathcal{R}^\dagger \cup \mathcal{R}^\ddagger$ with $p^\dagger = |\mathcal{R}^\dagger|$, $p^\ddagger = |\mathcal{R}^\ddagger|$ and $p^\dagger \asymp p^\ddagger \asymp p$. Accordingly, the data matrix $\boldsymbol{X}$ is divided into two submatrices $\boldsymbol{X}^\dagger$ and $\boldsymbol{X}^\ddagger$ with row indices in $\mathcal{R}^\dagger$ and $\mathcal{R}^\ddagger$ respectively. $\boldsymbol{U}^\dagger$ and $\boldsymbol{U}^\ddagger$ are defined similarly. In principle, the splitting is only used to make the two portions of data $\boldsymbol{X}^\dagger$ and $\boldsymbol{X}^\ddagger$ approximately independent (conditionally on the latent variables and latent functions). Under Assumption 2(e) imposed below, this goal can be easily achieved by, for example, randomly splitting the row index set. When measurements are collected over a time series dimension, one may, for example, take the first half of time periods for K-NN and the second half for PCA.

K-Nearest Neighbors Matching. This step makes use of the subsample labeled by $\dagger$,

Algorithm 1 (Causal inference with latent confounders)

Step 1: Latent Variables Extraction

Input: covariate matrix $\boldsymbol{X} \in \mathbb{R}^{p\times n}$, tuning parameters K, d_i

Output: $\{\mathcal{N}_i\}_{i=1}^n$, $\{\widehat{\boldsymbol{\Lambda}}_{\langle i\rangle}\}_{i=1}^n$

Row-wise split $\boldsymbol{X}$ into two submatrices $\boldsymbol{X}^\dagger \in \mathbb{R}^{T^\dagger\times n}$ and $\boldsymbol{X}^\ddagger \in \mathbb{R}^{T^\ddagger\times n}$

For $i \in [n]$,

(1) use $\boldsymbol{X}^\dagger$ to obtain the set $\mathcal{N}_i$ of the K nearest neighbors of unit i based on distance $\rho(\cdot,\cdot)$:

$$\mathcal{N}_i = \Big\{ j_k(i) : \sum_{\ell=1}^n \mathbb{1}\Big(\rho(\boldsymbol{X}^\dagger_{\cdot i}, \boldsymbol{X}^\dagger_{\cdot \ell}) \le \rho(\boldsymbol{X}^\dagger_{\cdot i}, \boldsymbol{X}^\dagger_{\cdot j_k(i)})\Big) \le K,\ 1 \le k \le K \Big\}$$

(2) use $\boldsymbol{X}_{\langle i\rangle} = (\boldsymbol{X}^\ddagger_{\cdot j_1(i)}, \cdots, \boldsymbol{X}^\ddagger_{\cdot j_K(i)})$ to obtain the local factor loading $\widehat{\boldsymbol{\Lambda}}_{\langle i\rangle}$ by local PCA:

$$(\widehat{\boldsymbol{F}}_{\langle i\rangle}, \widehat{\boldsymbol{\Lambda}}_{\langle i\rangle}) = \mathop{\arg\min}_{\tilde{\boldsymbol{F}}_{\langle i\rangle} \in \mathbb{R}^{T^\ddagger\times \mathsf{d}_i},\, \tilde{\boldsymbol{\Lambda}}_{\langle i\rangle} \in \mathbb{R}^{K\times \mathsf{d}_i}} \operatorname{Tr}\Big[\Big(\boldsymbol{X}_{\langle i\rangle} - \tilde{\boldsymbol{F}}_{\langle i\rangle}\tilde{\boldsymbol{\Lambda}}'_{\langle i\rangle}\Big)\Big(\boldsymbol{X}_{\langle i\rangle} - \tilde{\boldsymbol{F}}_{\langle i\rangle}\tilde{\boldsymbol{\Lambda}}'_{\langle i\rangle})'\Big]$$

Step 2: Local Quasi-Maximum Likelihood Estimation (QMLE)

Input: dependent variables: $\{y_i\}_{i=1}^n$, $\{d_i(t)\}_{i=1}^n$; independent variables: $\{\boldsymbol{z}_i\}_{i=1}^n$, $\{\widehat{\boldsymbol{\Lambda}}_{\langle i\rangle}\}_{i=1}^n$;

neighborhoods: $\{\mathcal{N}_i\}_{i=1}^n$

Output: fitted values $\{\widehat{\zeta}_{i,t}\}_{i=1}^n$ and $\{\widehat{\boldsymbol{\varrho}}_i\}_{i=1}^n$

(1) For each $i \in [n]$ and $t \in \mathcal{J}$, estimate $\zeta_{i,t}$ by local QMLE, as described in (3.4), using data for units in $\mathcal{N}_i$ with treatment status equal to t

(2) Similarly, for each $i \in [n]$, estimate $\boldsymbol{\varrho}_i$ by local QMLE using data for units in $\mathcal{N}_i$

Step 3: Counterfactual Analysis

Input: $\{y_i\}_{i=1}^n$, $\{d_i(t)\}_{i=1}^n$, $\{\widehat{\zeta}_{i,t}\}_{i=1}^n$, $\{\widehat{\varrho}_i\}_{i=1}^n$

Output: $\{\widehat{\theta}_{t,t'}\}_{t,t'\in\mathcal{J}}$ and related quantities

(1) Obtain the estimator $\widehat{\theta}_{t,t'}$ of $\theta_{t,t'} = \mathbb{E}[y_i(t)|s_i = t']$ and its standard error $\widehat{\sigma}^2_{t,t'}$:

$$\widehat{\theta}_{t,t'} = \frac{1}{n}\sum_{i=1}^n \left[\frac{d_i(t')\widehat{\zeta}_{i,t}}{\widehat{\varrho}_{t'}} + \frac{\widehat{\varrho}_{i,t'}}{\widehat{\varrho}_{t'}}\frac{d_i(t)(y_i - \widehat{\zeta}_{i,t})}{\widehat{\varrho}_{i,t}}\right], \quad \widehat{\sigma}^2_{t,t'} = \frac{1}{n}\sum_{i=1}^n \left[\frac{d_i(t')(\widehat{\zeta}_{i,t} - \widehat{\theta}_{t,t'})^2}{\widehat{\varrho}^2_{t'}} + \frac{\widehat{\varrho}^2_{i,t'} d_i(t)(y_i - \widehat{\zeta}_{i,t})^2}{\widehat{\varrho}^2_{t'}\widehat{\varrho}^2_{i,t}}\right]$$

(2) Construct estimators of other quantities based on $\{\widehat{\theta}_{t,t'}\}$

i.e., the submatrix of $\boldsymbol{X}$ with row indices in $\mathcal{R}^\dagger$. Given a "distance" function $\rho : \mathbb{R}^{p^\dagger} \times \mathbb{R}^{p^\dagger} \mapsto \mathbb{R}$, search for a set of indices $\mathcal{N}_i$ for the K nearest neighbors of unit i (including i itself):

$$\mathcal{N}_i = \Big\{ j_k(i) : \sum_{\ell=1}^n \mathbb{1}\Big(\rho(\boldsymbol{X}^\dagger_{\cdot i}, \boldsymbol{X}^\dagger_{\cdot \ell}) \le \rho(\boldsymbol{X}^\dagger_{\cdot i}, \boldsymbol{X}^\dagger_{\cdot j_k(i)})\Big) \le K,\ 1 \le k \le K \Big\}. \tag{3.1}$$

We use the term "distance" in a loose sense so that ρ does not have to satisfy all the axioms for a distance function in math. Some usual choices include (i) the squared Euclidean distance $\rho(\boldsymbol{X}^{\dagger}_{\cdot i}, \boldsymbol{X}^{\dagger}_{\cdot j}) = \|\boldsymbol{X}^{\dagger}_{\cdot i} - \boldsymbol{X}^{\dagger}_{\cdot j}\|^2/p^{\dagger}$, (ii) the pseudo-max distance $\rho(\boldsymbol{X}^{\dagger}_{\cdot i}, \boldsymbol{X}^{\dagger}_{\cdot j}) = \max_{l \neq i,j} |(\boldsymbol{X}^{\dagger}_{\cdot i} - \boldsymbol{X}^{\dagger}_{\cdot j})'\boldsymbol{X}^{\dagger}_{\cdot l}|/p^{\dagger}$, and (iii) the distance of averages $\rho(\boldsymbol{X}^{\dagger}_{\cdot i}, \boldsymbol{X}^{\dagger}_{\cdot j}) = |\boldsymbol{1}'_{p^{\dagger}}(\boldsymbol{X}^{\dagger}_{\cdot i} - \boldsymbol{X}^{\dagger}_{\cdot j})|/p^{\dagger}$. Moreover, when the noisy measurements differ in scale or importance for revealing information on the latent variables, one can rescale or reweight different measurements when searching for nearest neighbors. Such transformations can be viewed as particular choices of the distance.

The number of nearest neighbors K is the key tuning parameter of the entire estimation procedure. A formal method for selecting K is beyond the scope of this paper, but we emphasize that taking advantage of doubly-robust score functions, our main inference results rely on mild restrictions on K, which reduces to $n/K^2 = o(1)$ (up to $\log n$ terms) in a practically relevant case. See more discussion in Remark 4.3.

Using the data of Acemoglu et al. (2016), we implement K-NN matching for each unit based on the pseudo-max distance of daily stock returns in the first 125 days. We make use of a data-dependent benchmark choice of $K = 99$, which is obtained based on leave-one-out cross validation (CV) for local constant nearest neighbors regression of the cumulative stock returns from date 0 to 1 on the pre-treatment return at date $t = -30$. This regression is not what we need for the causal analysis, but in principle, it gives a CV choice of K of an order $O(n^{4/5})$, satisfying the requirement for our inference theory (see Theorem 4.2). Note that due to the noise in the measurements, choosing a small K may not help reduce the resultant matching discrepancy (see discussions below Lemma 4.1). To have a sense of the performance of K-NN matching, we calculate for each unit the maximum distance of matched pairs divided by the standard deviation of the distance across all pairs, which can be viewed as a normalized matching discrepancy in terms of the observed returns. Table 1 reports some summary statistics for treated and control groups respectively. Matching performs relatively well for treated units, whereas some control units are matched with someone relatively far away. In later analysis, we will check the robustness of the results by varying the number of nearest neighbors or dropping control units with large discrepancy.

Table 1: K-NN Matching: Maximum Distance of Matched Pairs

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Treated	0.671	1.043	1.166	1.199	1.313	1.840
Control	0.602	0.881	1.064	1.228	1.345	6.638

Notes: For each unit, the maximum distance of matched pairs are normalized by dividing it by the standard deviation of the distance across all pairs.

Local Principal Component Analysis. This step makes use of the subsample labeled by ‡, i.e., the submatrix of $\boldsymbol{X}$ with row indices in $\mathcal{R}^\ddagger$. Given a set of nearest neighbors $\mathcal{N}_i$ from the previous step, define a $p^\ddagger \times K$ matrix $\boldsymbol{X}_{\langle i\rangle} = (\boldsymbol{X}^\ddagger_{\cdot j_1(i)}, \cdots, \boldsymbol{X}^\ddagger_{\cdot j_K(i)})$. The subscript $\langle i\rangle$ indicates that the data matrix is defined locally for unit i. For these nearest neighbors, the unknown function $\boldsymbol{\eta}$ can be locally approximated by a linear combination of some basis functions of latent variables. Then, $\boldsymbol{X}_{\langle i\rangle}$ admits a linear factor structure up to approximation errors:

$$\boldsymbol{X}_{\langle i\rangle} = \boldsymbol{F}_{\langle i\rangle}\boldsymbol{\Lambda}'_{\langle i\rangle} + \boldsymbol{\Xi}_{\langle i\rangle} + \boldsymbol{U}_{\langle i\rangle}, \tag{3.2}$$

where $\boldsymbol{U}_{\langle i\rangle} = (\boldsymbol{U}^\ddagger_{\cdot j_1(i)}, \cdots, \boldsymbol{U}^\ddagger_{\cdot j_K(i)})$, $\boldsymbol{F}_{\langle i\rangle} \in \mathbb{R}^{p^\ddagger \times \mathsf{d}_i}$ is the local factor matrix, $\boldsymbol{\Lambda}_{\langle i\rangle} \in \mathbb{R}^{K\times \mathsf{d}_i}$ is the local factor loading matrix, and $\boldsymbol{\Xi}_{\langle i\rangle} \in \mathbb{R}^{p^\ddagger \times K}$ is the corresponding approximation error. The user-specified parameter d_i governs the number of approximation terms. $\boldsymbol{F}_{\langle i\rangle}$ and $\boldsymbol{\Lambda}_{\langle i\rangle}$ can be identified up to a rotation and estimated by PCA (Bishop, 2006):

$$(\widehat{\boldsymbol{F}}_{\langle i\rangle}, \widehat{\boldsymbol{\Lambda}}_{\langle i\rangle}) = \underset{\tilde{\boldsymbol{F}}_{\langle i\rangle} \in \mathbb{R}^{p^\ddagger\times \mathsf{d}_i},\, \tilde{\boldsymbol{\Lambda}}_{\langle i\rangle} \in \mathbb{R}^{K\times \mathsf{d}_i}}{\arg\min}\ \mathrm{Tr}\Big[\Big(\boldsymbol{X}_{\langle i\rangle} - \tilde{\boldsymbol{F}}_{\langle i\rangle}\tilde{\boldsymbol{\Lambda}}'_{\langle i\rangle}\Big)\Big(\boldsymbol{X}_{\langle i\rangle} - \tilde{\boldsymbol{F}}_{\langle i\rangle}\tilde{\boldsymbol{\Lambda}}'_{\langle i\rangle})'\Big] \tag{3.3}$$

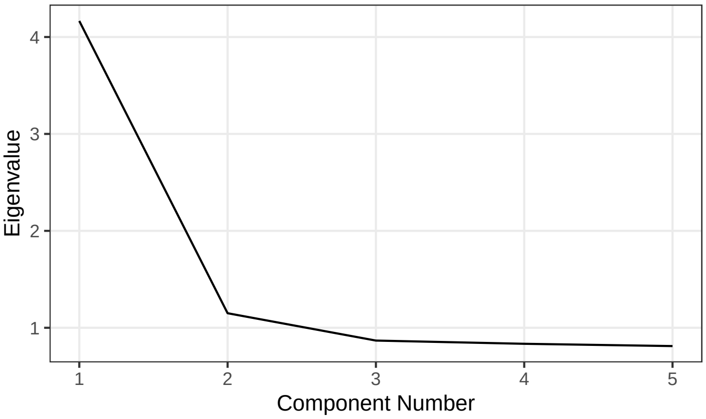
such that $\frac{1}{p^\ddagger}\tilde{\boldsymbol{F}}'_{\langle i\rangle}\tilde{\boldsymbol{F}}_{\langle i\rangle} = \boldsymbol{I}_{\mathsf{d}_i}$ and $\frac{1}{K}\tilde{\boldsymbol{\Lambda}}'_{\langle i\rangle}\tilde{\boldsymbol{\Lambda}}_{\langle i\rangle}$ is diagonal. Let $\widehat{\boldsymbol{\lambda}}_{\ell,\langle i\rangle}$ be the column in $\widehat{\boldsymbol{\Lambda}}_{\langle i\rangle}$ that corresponds to a generic unit ℓ.

The number of local principal components d_i plays a role similar to the degree of the basis in local polynomial regression and can be set as a fixed number (independent of n and p), as long as the extracted local factors has stronger signal strength relative to the noise $\boldsymbol{U}_{\langle i\rangle}$.

See more discussion below Assumption 4.

Remark 3.1 (Number of Latent Confounders r)**.** As will be shown later in Section 4, given the tuning parameters K and d_i, the number of latent confounders r implicitly affects the approximation quality of our proposed method. From the practical perspective, however, it is unnecessary to determine r in the above local PCA procedure. Assuming the latent functions η_l's are sufficiently smooth, one could, for example, extract all local factors associated with large eigenvalues (compared to the noise matrix), thus maximizing the approximation power of $\widehat{\boldsymbol{\Lambda}}_{\langle i\rangle}$ in the next step. Nevertheless, determining the number of latent confounders may be of independent interest. See Remark 4.4 of Feng (2023) for some discussion on how r can be determined via this local PCA procedure. $\lrcorner$

Figure 1: Local Eigenvalues for One Neighborhood



As an illustration, we implement local PCA for each unit using the data of Acemoglu et al. (2016). Recall that for each firm a set of nearest neighbors has been obtained using the daily stock returns in the first 125 days. PCA can be conducted for this subgroup of firms using their daily stock returns in the next 125 days. Figure 1 shows several leading eigenvalues corresponding to the neighborhood for a particular unit ("AMERICAN EXPRESS CO."), suggesting that extracting one or two local principal components is a reasonable choice. In the subsequent analysis, we employ a simple data-dependent rule to determine d_i and avoid extracting "too weak" factors: $\mathsf{d}_i = 2$ if $\upsilon_{2,\langle i\rangle}/\upsilon_{3,\langle i\rangle} \geq \log\log K$ and $\mathsf{d}_i = 1$ otherwise, for each $i \in [n]$.

3.2 Step 2: Local Quasi-Maximum Likelihood Estimation

For the outcome equation (2.2), consider a quasi-log-likelihood function $\mathcal{L}_{\mathsf{y}}(\zeta, y)$ such that $\frac{\partial}{\partial\zeta}\mathcal{L}_{\mathsf{y}}(\zeta, y) = \frac{y-\zeta}{V_{\mathsf{y}}(\zeta)}$. Then, a local quasi-maximum likelihood estimator of $\zeta_{i,t}$ is given by

$$\begin{gathered}\widehat{\zeta}_{i,t} = \psi_{\mathsf{y}}\Big(\widehat{\mu}_t(\boldsymbol{\alpha}_i) + \boldsymbol{z}_i'\widehat{\boldsymbol{\beta}}_{t,\langle i\rangle}\Big), \quad \widehat{\mu}_t(\boldsymbol{\alpha}_i) = \widehat{\boldsymbol{\lambda}}_{i,\langle i\rangle}'\widehat{\boldsymbol{b}}_{t,\langle i\rangle}, \quad \text{where} \\ \Big(\widehat{\boldsymbol{b}}_{t,\langle i\rangle}',\ \widehat{\boldsymbol{\beta}}_{t,\langle i\rangle}'\Big)' = \underset{(\boldsymbol{b}',\,\boldsymbol{\beta}')'\in\mathbb{R}^{\mathsf{d}_i+\mathsf{d}_z}}{\arg\max} \sum_{j\in\mathcal{N}_i} d_j(t)\mathcal{L}_{\mathsf{y}}\Big(\psi_{\mathsf{y}}(\widehat{\boldsymbol{\lambda}}_{j,\langle i\rangle}'\boldsymbol{b} + \boldsymbol{z}_j'\boldsymbol{\beta}),\ y_j\Big).\end{gathered} \tag{3.4}$$

For each unit i, the fitting is restricted to its local neighborhood $\mathcal{N}_i$, and $\widehat{\boldsymbol{\lambda}}_{j,\langle i\rangle}$ plays the role of generated regressors used to control for the latent confounders $\boldsymbol{\alpha}_j$ in this context. For continuous outcomes, it is common practice to implement local least squares regression.

Similarly, for the treatment equation (2.3), given a quasi-likelihood $\mathcal{L}_{\mathsf{p}}(\cdot,\cdot)$ associated with some functions $\{V_{\mathsf{p},t}(\cdot)\}_{t\in[J]}$ such that $\frac{\partial}{\partial\kappa_t}\mathcal{L}_{\mathsf{p}}(\boldsymbol{\psi}_{\mathsf{p}}(\boldsymbol{\kappa}), \boldsymbol{d}) = \frac{d(t)-\psi_{\mathsf{p},t}(\boldsymbol{\kappa})}{V_{\mathsf{p},t}(\boldsymbol{\kappa})}$ for $\boldsymbol{\kappa} = (\kappa_1, \cdots, \kappa_J)'$, implement a local quasi-maximum likelihood estimation, which gives the predicted conditional treatment probability

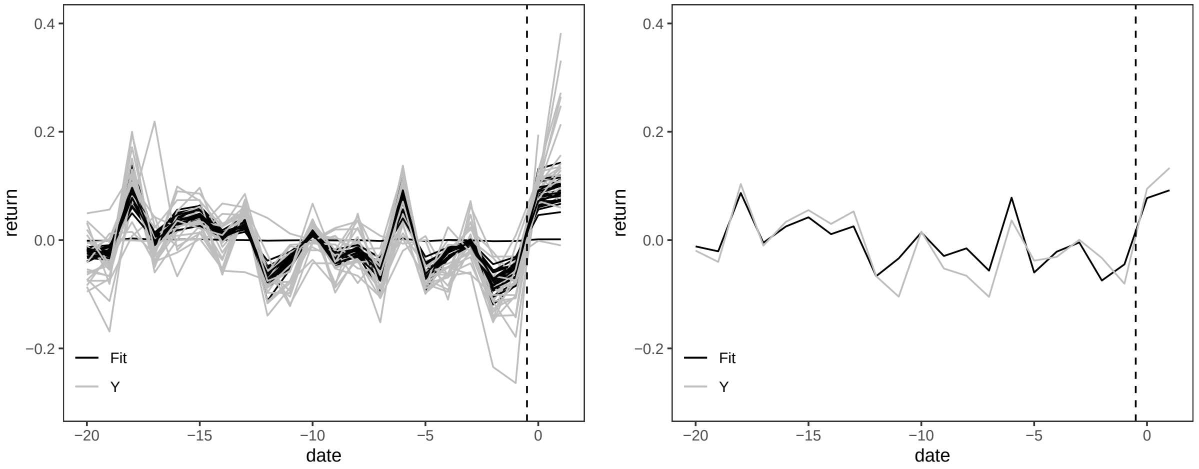
$$\widehat{\boldsymbol{\varrho}}_i = \boldsymbol{\psi}_{\mathsf{p}}(\widehat{\boldsymbol{e}}(\boldsymbol{\alpha}_i) + \widehat{\boldsymbol{\Gamma}}\boldsymbol{z}_i).$$

For discrete treatments, it is common practice to implement local least squares regression or local multinomial logit estimation.

One could also exploit other standard methods in the semiparametrics literature, e.g., profiled quasi-maximum likelihood, to estimate the parametric components $\boldsymbol{\beta}_t$ and $\boldsymbol{\Gamma}$, though it is computationally more burdensome. See Härdle et al. (2004) for implementation details.

For the purpose of illustration, we implement a local least squares regression of stock returns at date t on the local factor loadings extracted in the previous step, for each $t = -20, \cdots, 0, 1$, where $t = 0$ denotes the day when the treatment starts. Figure 2a shows the fitted values in black and the observed daily returns in grey for the 22 treated firms, and the result for American Express Co. is displayed in Figure 2b. Recall that the fitted values are the estimates of the conditional means of stock returns without treatment given the latent variables. Clearly, after day 0, many sequences of stock returns increase sharply compared to the corresponding fitted values.

Figure 2: Local Least Squares: Stock Returns



(a) All Treated Firms

(b) American Express Co.

3.3 Step 3: Counterfactual Analysis

The final step is to estimate the counterfactual means of potential outcomes, which forms the basis of estimators for other causal parameters. Specifically, consider $\theta_{t,t'} := \mathbb{E}[y_i(t)|t_i = t']$. Let $\varrho_t = \mathbb{P}(t_i = t)$ for any $t \in \mathcal{J}$. Under our unconfoundedness assumption (2.1),

$$\theta_{t,t'} = \mathbb{E}\left[\frac{d_i(t')\zeta_{i,t}}{\varrho_{t'}} + \frac{\varrho_{i,t'}}{\varrho_{t'}}\frac{d_i(t)(y_i - \zeta_{i,t})}{\varrho_{i,t}}\right].$$

An estimator of $\theta_{t,t'}$ is given by

$$\widehat{\theta}_{t,t'} := \frac{1}{n}\sum_{i=1}^{n}\left[\frac{d_i(t')\widehat{\zeta}_{i,t}}{\widehat{\varrho}_{t'}} + \frac{\widehat{\varrho}_{i,t'}}{\widehat{\varrho}_{t'}}\frac{d_i(t)(y_i - \widehat{\zeta}_{i,t})}{\widehat{\varrho}_{i,t}}\right], \tag{3.5}$$

where $\widehat{\zeta}_{i,t}$, $\widehat{\varrho}_{i,t}$ and $\widehat{\varrho}_{i,t'}$ are obtained in the second step, and $\widehat{\varrho}_{\ell} = \frac{1}{n}\sum_{i=1}^{n} d_i(\ell)$ for $\ell \in \mathcal{J}$. For the purpose of inference, a simple plug-in variance estimator for $\widehat{\theta}_{t,t'}$ is

$$\widehat{\sigma}^2_{t,t'} := \frac{1}{n}\sum_{i=1}^{n}\left[\frac{d_i(t')(\widehat{\zeta}_{i,t} - \widehat{\theta}_{t,t'})^2}{\widehat{\varrho}^2_{t'}}\right] + \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\widehat{\varrho}^2_{i,t'} d_i(t)(y_i - \widehat{\zeta}_{i,t})^2}{\widehat{\varrho}^2_{t'}\widehat{\varrho}^2_{i,t}}\right] \tag{3.6}$$

Under some regularity conditions, we show in Theorem 4.2 below that $\sqrt{n}\widehat{\sigma}^{-1}_{t,t'}(\widehat{\theta}_{t,t'} - \theta_{t,t'}) \rightsquigarrow \mathsf{N}(0,1)$. Confidence intervals and hypothesis testing procedures can be constructed accordingly.

Estimators of other parameters may be constructed in a similar way or based on $\{\widehat{\theta}_{t,t'}\}_{t,t'\in\mathcal{J}}$. For example, the average treatment effect on the treatment group $t' = \ell$ compared to the baseline treatment status $t = 0$ can be estimated by $\widehat{\theta}_{\ell,\ell} - \widehat{\theta}_{0,\ell}$ where $\widehat{\theta}_{\ell,\ell} = \sum_{i=1}^{n} d_i(\ell) y_i / \sum_{i=1}^{n} d_i(\ell)$.

As an illustration, we estimate the average treatment effect of Geithner connections on cumulative returns from day 0 to day 1 (CAR[0,1]) for firms with connections. Since the number of treated units is relatively small, the propensity score is estimated by taking a simple local average of treatment indicators within each local neighborhood. For the outcome equation, we implement a local least squares regression of cumulative stock returns of firms with no connections on the local factor loadings extracted previously. Different choices of K are considered, which correspond to $K = CK_0$ where $C = 0.5, 1, 2$ and $K_0 = 99$ is the previously described benchmark choice. The above procedure is applied to the full sample and a base sample. The latter, as defined in Acemoglu et al. (2016), excludes firms whose returns are highly correlated with Citigroup.

Results are reported in the first two columns of Table 2. We also include two results based on synthetic matching from Acemoglu et al. (2016) and one result based on a penalized synthetic control method from Abadie and L'Hour (2021). To make these results comparable, we follow these two papers and report the 95% acceptance regions for hypothesis testing of the average treatment effect (on the treated) being equal to zero (numbers in brackets in Table 2), but note that the underlying assumptions and inference methodology of the other two papers are different from ours. The estimated average cumulative abnormal return for the connected firms using the proposed method ranges from 0.053 to 0.105 and significantly differs from zero at the 0.05 level. Compared with the other two papers, the magnitude of the estimated effect is greater, and it is significant even when a full sample is utilized. We also check the robustness of the results by excluding firms in the control group with large normalized matching discrepancy (top 10% in Table 1). Results are similar and omitted to conserve space.

Table 2: Average Treatment Effect of Connections on the Treated

	No Covariates		**Add Covariates**	
	Full Sample	Base Sample	Full Sample	Base Sample
Local PCA, $K =$				
50	0.095	0.083	0.075	0.053
	[-0.054, 0.054]	[-0.049, 0.049]	[-0.054, 0.054]	[-0.052, 0.052]
99	0.103	0.094	0.089	0.073
	[-0.054, 0.054]	[-0.051, 0.051]	[-0.052, 0.052]	[-0.049, 0.049]
198	0.105	0.098	0.092	0.085
	[-0.055, 0.055]	[-0.053, 0.053]	[-0.055, 0.055]	[-0.052, 0.052]
Acemoglu et al. (2016)				
Estimate	0.005	0.060	-	-
AR for TE=0	[-0.029, 0.014]	[-0.068, 0.036]	-	-
Abadie and L'Hour (2021)				
Estimate	-	0.061	-	-
AR for TE=0	-	[-0.050, 0.061]	-	-

Notes: CAR[0,1] is the cumulative abnormal return from day 0 to day 1. The base sample excludes firms highly correlated with Citigroup. The numbers in brackets are the 95% acceptance regions for hypothesis testing of the effect of connections being equal to zero.

The analysis so far has controlled for latent variables only. Three additional covariates are available in the dataset of Acemoglu et al. (2016): firm size (log of total assets), profitability (return on equity), and leverage (total debt to total capital) as of 2008. They can be incorporated into the local least squares regression in Step 2 as additional regressors $\boldsymbol{z}_i$. Results are reported in the third and fourth columns of Table 2. The estimated effect is slightly smaller than that without additional covariates, but still significant at the 0.05 level.

4 Main Results

4.1 Assumptions

We begin with the unconfoundedness and overlap conditions commonly used in the causal inference literature. Note that the conditioning variables $\boldsymbol{\alpha}_i$ in this scenario are not directly observed.

Assumption 1 (Unconfoundedness and Overlap)**.** *$(y_i, t_i, \boldsymbol{z}_i, \boldsymbol{\alpha}_i)$ is i.i.d. over $i \in [n]$ and satisfies that (a) $y_i(t) \perp\!\!\!\perp d_i(t') \,|\, \boldsymbol{\alpha}_i, \boldsymbol{z}_i$, $\forall t, t' \in \mathcal{J}$; and (b) for all $t \in \mathcal{J}$, $\mathbb{P}(t_i = t|\boldsymbol{\alpha}_i, \boldsymbol{z}_i) \geq c_{\min} > 0$ for some constant $c_{\min}$ almost surely.*

The next assumption imposes mild regularity conditions on the treatment effects model and the latent structure of $\boldsymbol{x}_i$.

Assumption 2 (Regularities)**.** *Let $\bar{m} \geq 2$ and $\nu > 0$ be some constants. Equations (2.2), (2.3) and (2.5) hold with the following conditions satisfied:*

(a) For all $t \in \mathcal{J}$, $\mu_t(\cdot)$ and $e_t(\cdot)$ are $\bar{m}$-times continuously differentiable.

(b) $\boldsymbol{z}_i$ has a compact support and $\mathbb{E}[\tilde{\boldsymbol{z}}_i \tilde{\boldsymbol{z}}_i' | \boldsymbol{\alpha}_i] > 0$ a.s. for $\tilde{\boldsymbol{z}}_i = \boldsymbol{z}_i - \mathbb{E}[\boldsymbol{z}_i|\boldsymbol{\alpha}_i]$. Conditional on $\mathcal{F}$, $\{(\boldsymbol{\epsilon}_i, \boldsymbol{v}_i) : i \in [n]\}$ are independent across i with zero means and are independent of $\{\boldsymbol{x}_i : i \in [n]\}$. Also, $\max_{i\in[n]} \mathbb{E}[\|\boldsymbol{\epsilon}_i\|^{2+\nu}|\mathcal{F}] < \infty$ and $\max_{i\in[n]} \mathbb{E}[\|\boldsymbol{v}_i\|^{2+\nu}|\mathcal{F}] < \infty$ a.s. on $\mathcal{F}$.

(c) $\{\boldsymbol{\alpha}_i : i \in [n]\}$ has a compact convex support $\mathcal{A}$ with a density bounded and bounded away from zero.

(d) For all $l \in [p]$, $\eta_l(\cdot)$ is $\bar{m}$-times continuously differentiable with all partial derivatives of order no greater than $\bar{m}$ bounded by a universal constant.

(e) Conditional on $\mathcal{F}$, $\{u_{il} : i \in [n], l \in [p]\}$ are independent across i and over l, and $\max_{i\in[n],l\in[p]} \mathbb{E}[|u_{il}|^{2+\nu}|\mathcal{F}] < \infty$ a.s. on $\mathcal{F}$.

Parts (a), (b), and (c) concern the regularities of the treatment effects model characterized by Equations (2.2) and (2.3). The conditional means of potential outcomes and propensity scores are sufficiently smooth functions, and other standard conditions are imposed on the conditioning variables and errors. Regarding the latent structure of $\boldsymbol{x}_i$ described in Equation (2.5), part (d) ensures that all latent functions belong to a Hölder class of order $\bar{m}$, and part (e) are standard conditions on errors commonly used in factor analysis and graphon estimation. The constant $\bar{m}$ governs the smoothness of unknown functions, and ν controls the tails of error terms. They are assumed to be the same across Equations (2.2)–(2.5) to ease the presentation. Also, the independence of u_{il} across i and l is assumed for simplicity only and can be relaxed to accommodate weak dependence in one or two dimensions, albeit with more technical complexity.

The main task of the first step is to learn $\boldsymbol{\alpha}_i$ from $\boldsymbol{x}_i$. It is unnecessary to identify $\boldsymbol{\alpha}_i$ in an exact sense since only *predictions* based on $\boldsymbol{\alpha}_i$ matter for the causal analysis. Intuitively, the nearest neighbors and local factor loadings described in Section 3 suffice to restore unconfoundedness if they can reflect the local geometric relations among latent confounders $\boldsymbol{\alpha}_i$'s. The key conditions required are formalized in the next two assumptions. The first one, allowing for generic distance choices, ensures that the indirectly obtained nearest neighbors are truly close in terms of the unobserved confounders, with the corresponding matching discrepancy precisely quantified.

Assumption 3 (Indirect Matching)**.** *For some fixed positive constants ρ_0, $\underline{\varsigma}$, $\bar{\varsigma}$ and some positive sequence $a_n = o(1)$, the following conditions hold:*

(a) $\max_{1\leq i,j\leq n} |\rho(\boldsymbol{X}^{\dagger}_{\cdot i}, \boldsymbol{X}^{\dagger}_{\cdot j}) - \rho(\boldsymbol{H}^{\dagger}_{\cdot i}, \boldsymbol{H}^{\dagger}_{\cdot j}) - \rho_0| \lesssim_{\mathbb{P}} a_n$;

(b) $\max_{\substack{1\leq i,j\leq n\\ \boldsymbol{\alpha}_i\neq\boldsymbol{\alpha}_j}} \frac{\rho(\boldsymbol{H}^{\dagger}_{\cdot i}, \boldsymbol{H}^{\dagger}_{\cdot j})}{\|\boldsymbol{\alpha}_i-\boldsymbol{\alpha}_j\|^{\bar{\varsigma}}} \lesssim_{\mathbb{P}} 1$ *and* $\min_{\substack{1\leq i,j\leq n\\ \boldsymbol{\alpha}_i\neq\boldsymbol{\alpha}_j}} \frac{\rho(\boldsymbol{H}^{\dagger}_{\cdot i}, \boldsymbol{H}^{\dagger}_{\cdot j})}{\|\boldsymbol{\alpha}_i-\boldsymbol{\alpha}_j\|^{\underline{\varsigma}}} \gtrsim_{\mathbb{P}} 1.$

Assumption 3 is a high-level condition, accommodating a generic choice of the distance ρ. Part (a) implies that the distance of the noisy measurements can be translated into that of the noise-free mean structure, which is usually a mild requirement if the distance “averages” many independent (or weakly dependent) measurements. On the other hand, part (b) precisely links the distance of the noise-free mean vectors with that of the latent confounders. In general, the upper bound requirement is mild given the smoothness of latent functions, while the plausibility of the lower bound, the key requirement for informativeness of measurements, needs to be understood in context. (See Remark 4.1 below.) Feng (2023) provides further discussion of Assumption 3 and details the parameters $\underline{\varsigma}$, $\bar{\varsigma}$, ρ_0 and a_n for several specific distance choices. In particular, under different sufficient conditions in each case, we can show that (i) if ρ is the (squared) Euclidean distance, $\rho_0 = 2\sigma^2$, $a_n = (\frac{\log n}{p})^{1/4}$, and $\bar{\varsigma} = \underline{\varsigma} = 2$, where $\mathbb{E}[u_{il}^2|\mathscr{F}]$ is assumed to be a constant σ^2; (ii) if ρ is the pseudo-max distance, $\rho_0 = 0$, $a_n = (\frac{\log n}{p})^{1/2}$, and $\bar{\varsigma} = \underline{\varsigma} = 1$; and (iii) if ρ is the distance of averages, $\rho_0 = 0$, $a_n = (\frac{\log n}{p})^{1/2}$, and $\underline{\varsigma} = \bar{\varsigma} = 1$.

Remark 4.1 (Informativeness Requirement)**.** The lower bound condition in Assumption 3(b) is a fundamental requirement for informativeness of measurements. To get a sense of its plausibility, consider a linear factor model: $\eta_l(\alpha_i) = \varpi_l\alpha_i$ for $\varpi_l \in \mathbb{R}$. When the distance of averages is used, we typically need the probability limit of $\frac{1}{p^\dagger}\sum_{l\in\mathcal{R}^\dagger}\eta_l(\cdot)$ to be strictly monotonic. This may be violated, for example, when $\frac{1}{p^\dagger}\sum_{l\in\mathcal{R}^\dagger}\varpi_l \to_{\mathbb{P}} 0$. In other words, the averaged measurements could be completely uninformative about the latent confounders. However, as long as a non-negligible subset of measurements have nonzero ϖ_l’s and thus are (individually) informative, the Euclidean distance or pseudo-max distance may still be able to differentiate two units with different values of latent confounders. It is unnecessary to know the identities of informative measurements a priori, but if users do have some prior knowledge, a corresponding weighting scheme can be incorporated into the definition of the distance, which is completely accommodated in our framework. It should be emphasized that full verification of Assumption 3 also relies on other sufficient conditions for each distance function (see Appendix A of Feng, 2023 for more technical details), so the comparison here does not suggest a theoretically superior choice. Users should select an appropriate distance in context. $\lrcorner$

Next, we impose some regularity conditions on the local principal subspace approximation of the noise-free mean structure so that the information extracted through local PCA of the measurements matrix $\boldsymbol{X}_{\langle i\rangle}$ can aid the approximation of the nonparametric components $\mu_t(\cdot)$ and $e_t(\cdot)$ in the conditional means of potential outcomes $\zeta_{i,t}$'s and the propensity score $\boldsymbol{\varrho}_i$'s, the key building blocks for the final causal parameter identification. Let $\mathbb{E}^{\ddagger}$ denote the expectation operator conditional on $\boldsymbol{X}^{\dagger}$, and introduce a diagonal (scaling) matrix $\boldsymbol{\Upsilon}_{\langle i\rangle} = \mathrm{diag}\{\upsilon_{1,\langle i\rangle}, \cdots, \upsilon_{\mathsf{d}_i,\langle i\rangle}\}$. Without loss of generality, assume $\upsilon_{1,\langle i\rangle} \geq \upsilon_{2,\langle i\rangle} \geq \cdots \geq \upsilon_{\mathsf{d}_i,\langle i\rangle}$. Also, write $h_n = (K/n)^{\bar{\varsigma}/(\underline{\varsigma}\mathsf{d}_\alpha)}$, which denotes the (direct) matching discrepancy (see Lemma 4.1 below) and plays a role similar to "bandwidth" in kernel-based nonparametric estimation.

Assumption 4 (Local Approximation)**.** *For each $i \in [n]$, there exists some $\boldsymbol{\Lambda}_{\langle i\rangle} \in \mathbb{R}^{K\times \mathsf{d}_i}$ such that the following conditions hold:*

(a) There exists some diagonal matrix $\boldsymbol{\Upsilon}_{\langle i\rangle}$ such that $\max_{i\in[n]}\|\boldsymbol{\Lambda}_{\langle i\rangle}\boldsymbol{\Upsilon}^{-1}_{\langle i\rangle}\|_{\max} \lesssim_{\mathbb{P}} 1$ and

$$1 \lesssim_{\mathbb{P}} \min_{1\leq i\leq n} s_{\min}\Big(\frac{1}{K}\boldsymbol{\Upsilon}^{-1}_{\langle i\rangle}\boldsymbol{\Lambda}'_{\langle i\rangle}\boldsymbol{\Lambda}_{\langle i\rangle}\boldsymbol{\Upsilon}^{-1}_{\langle i\rangle}\Big) \leq \max_{1\leq i\leq n} s_{\max}\Big(\frac{1}{K}\boldsymbol{\Upsilon}^{-1}_{\langle i\rangle}\boldsymbol{\Lambda}'_{\langle i\rangle}\boldsymbol{\Lambda}_{\langle i\rangle}\boldsymbol{\Upsilon}^{-1}_{\langle i\rangle}\Big) \lesssim_{\mathbb{P}} 1.$$

Either $\upsilon_{j,\langle i\rangle}/\upsilon_{j+1,\langle i\rangle} \lesssim 1$ or $\upsilon_{j,\langle i\rangle}/\upsilon_{j+1,\langle i\rangle} \to \infty$ holds for $j \in [\mathsf{d}_i - 1]$;

(b) $\boldsymbol{H}_{\langle i\rangle} = \boldsymbol{F}_{\langle i\rangle}\boldsymbol{\Lambda}'_{\langle i\rangle} + \boldsymbol{\Xi}_{\langle i\rangle}$ where $\boldsymbol{F}_{\langle i\rangle} = \mathbb{E}^{\ddagger}[\boldsymbol{H}_{\langle i\rangle}\boldsymbol{\Lambda}_{\langle i\rangle}]\mathbb{E}^{\ddagger}[\boldsymbol{\Lambda}'_{\langle i\rangle}\boldsymbol{\Lambda}_{\langle i\rangle}]^{-1}$, $\max_{i\in[n]}\|\boldsymbol{\Xi}_{\langle i\rangle}\|_{\max} \lesssim_{\mathbb{P}} h_n^m = \bar{o}(\upsilon_{\mathsf{d}_i,\langle i\rangle})$ for some $m \leq \bar{m}$, $\delta_n^{-1}/\upsilon_{\mathsf{d}_i,\langle i\rangle} = \bar{o}(1)$, and $1 \lesssim_{\mathbb{P}} \min_{i\in[n]} s_{\min}\Big(\frac{1}{p^{\ddagger}}\boldsymbol{F}'_{\langle i\rangle}\boldsymbol{F}_{\langle i\rangle}\Big) \leq \max_{i\in[n]} s_{\max}\Big(\frac{1}{p^{\ddagger}}\boldsymbol{F}'_{\langle i\rangle}\boldsymbol{F}_{\langle i\rangle}\Big) \lesssim_{\mathbb{P}} 1$.

(c) There exists $\mathfrak{b}_{t,\langle i\rangle}$ and $\mathfrak{c}_{t,\langle i\rangle}$ such that $\max_{j\in\mathcal{N}_i}|\mu_t(\boldsymbol{\alpha}_j) - \mathfrak{b}'_{t,\langle i\rangle}\boldsymbol{\lambda}_{j,\langle i\rangle}| = \bar{O}(\mathfrak{r}_{\mathsf{y},\langle i\rangle})$ and $\max_{j\in\mathcal{N}_i}|e_t(\boldsymbol{\alpha}_j) - \mathfrak{c}'_{t,\langle i\rangle}\boldsymbol{\lambda}_{j,\langle i\rangle}| = \bar{O}(\mathfrak{r}_{\mathsf{p},\langle i\rangle})$ for all $t \in \mathcal{J}$ and some $\mathfrak{r}_{\mathsf{y},\langle i\rangle} = \bar{o}(1)$ and $\mathfrak{r}_{\mathsf{p},\langle i\rangle} = \bar{o}(1)$.

Assumption 4 is a high-level condition, delineating the key requirements for informativeness of local factor loadings. We make several remarks on each part.

First, the matrix $\boldsymbol{\Lambda}_{\langle i\rangle}$, termed local factor loadings, can be viewed as an approximation basis that characterizes the relation of different units in terms of the latent confounders. Thus, Assumption 4(a) is a mild condition that specifies the possibly heterogeneous magnitude of different components of the basis and requires the basis matrix be non-degenerate. In local PCA of the measurements matrix $\boldsymbol{X}_{\langle i \rangle}$, this condition implies that the strength of factors for each local neighborhood may be heterogeneous, which differs from the usual linear factor analysis assuming *strong* factors.

Second, Assumption 4(b) quantifies the precision of the L_2-type approximation for the noise-free measurement matrix $\boldsymbol{H}_{\langle i \rangle}$ that can be achieved by the basis $\boldsymbol{\Lambda}_{\langle i \rangle}$. The matrix $\boldsymbol{F}_{\langle i \rangle}$, termed local factors, plays the role of the coefficients on the basis in this approximation, and $\boldsymbol{\Xi}_{\langle i \rangle}$ is the corresponding L_2-approximation error. As in usual factor analysis, $\boldsymbol{F}_{\langle i \rangle}$ is assumed to be non-degenerate. We allow d_i to differ across local neighborhoods, making this non-degeneracy condition accommodate the cases where the nonlinear surface generated by $\boldsymbol{\eta}(\cdot)$ have heterogeneous degrees or patterns of nonlinearity in different regions. The rate restriction $\delta_n^{-1} = \bar{o}(\upsilon_{\mathsf{d}_i,\langle i \rangle})$ ensures that even the "weakest" factors one desire to extract has stronger signal strength than the noise and thus are consistently estimable. See Feng (2023) for examples and detailed discussions about these requirements.

Finally, Assumption 4(c) requires the same basis $\boldsymbol{\Lambda}_{\langle i \rangle}$ also has ability to approximate the unknown functions $\mu_t(\cdot)$ and $e_t(\cdot)$, which are necessary for recovering the conditional means of potential outcomes and propensity scores in the final causal analysis. Importantly, as discussed before, the latent surface generated by the latent function $\boldsymbol{\eta}(\cdot)$ may have different degrees of local nonlinearity, making the ability of the local basis $\boldsymbol{\Lambda}_{\langle i \rangle}$ to approximate $\mu_t(\boldsymbol{\alpha}_i)$ and $e_t(\boldsymbol{\alpha}_i)$ heterogeneous across different units. This possibility is characterized by the dependence of the error bounds $\mathfrak{r}_{\mathsf{y},\langle i \rangle}$ and $\mathfrak{r}_{\mathsf{p},\langle i \rangle}$ on the neighborhood index $\langle i \rangle$. Our main results only need the approximation errors to be sufficiently small in an *average* sense rather than uniformly over all units. Thus, weak approximation power of $\boldsymbol{\Lambda}_{\langle i \rangle}$ is permitted in some (but not too many) neighborhoods.

Remark 4.2 (Verification of Assumption 4)**.** From a practical perspective, users do not need specify the basis $\boldsymbol{\Lambda}_{\langle i \rangle}$ explicitly, and the PCA procedure, as shown later, consistently estimate a rotated version of $\mathbf{\Lambda}_{\langle i\rangle}$ under the imposed conditions. From a theoretical perspective, however, it is convenient to verify Assumption 4 if $\mathbf{\Lambda}_{\langle i\rangle}$ can be explicitly specified. An important special case, though probably restrictive in some scenarios, is that the derivatives of latent functions η_l's up to a certain order are not too collinear so that $\mathbf{\Lambda}_{\langle i\rangle}$ can be understood as the usual monomial basis. Specifically, assume that for some $\underline{c} > 0$ and $2 \leq m \leq \bar{m}$,

$$\lim_{n,p^\ddagger\to\infty} \mathbb{P}\left\{\min_{i\in[n]} s_{\min}\left(\frac{1}{p^\ddagger}\sum_{l\in\mathcal{R}^\ddagger}(\mathscr{D}^{[m-1]}\boldsymbol{\eta}_l(\boldsymbol{\alpha}_i))(\mathscr{D}^{[m-1]}\boldsymbol{\eta}_l(\boldsymbol{\alpha}_i))'\right) \geq \underline{c}\right\} = 1, \tag{4.1}$$

where $\mathscr{D}^{[m-1]}\boldsymbol{\eta}_l(\cdot)$ is a column vector that stores all partial derivatives of $\eta_l(\cdot)$ up to order $m-1$. In this case Feng (2023) shows that Assumptions 4(a) and 4(b) hold for $\mathbf{\Lambda}_{\langle i\rangle} = (\boldsymbol{\lambda}(\boldsymbol{\alpha}_{j_1(i)}), \cdots, \boldsymbol{\lambda}(\boldsymbol{\alpha}_{j_K(i)}))'$ with $\boldsymbol{\alpha} \in \mathcal{A} \mapsto \boldsymbol{\lambda}(\boldsymbol{\alpha}) := (\lambda_1(\boldsymbol{\alpha}), \cdots, \lambda_{\mathsf{d}_i}(\boldsymbol{\alpha}))'$ a d_α-variate monomial basis of degree no greater than $m-1$ centered at $\boldsymbol{\alpha}_i$ (including the constant). Accordingly, Assumption 4(c) is immediate given the approximation power of the (local) monomial basis.

Note that in (4.1) the degree of nonlinearity is assumed for simplicity to be the same across different units (the "uniformity" over i in this condition), and consequently, we can set a universal $\mathsf{d}_i = \binom{m-1+\mathsf{d}_\alpha}{\mathsf{d}_\alpha}$, the number of the d_α-variate monomial basis functions. In general, this condition can be relaxed. As discussed above, we can let the number of local factors extracted be different across local neighborhoods so that the non-degeneracy condition of local factors is satisfied, while this may affect the ability of the basis to locally approximate μ_t and e_t in later analysis. $\lrcorner$

Finally, we impose a set of regularity conditions on the loss and link functions used by the local quasi-maximum likelihood estimation. Note that they are *unnecessary* in the special case of local least squares. Let $\bar{\mathcal{F}} = \mathcal{F} \cup \{\boldsymbol{z}_i\}_{i=1}^n$, and $\psi_{\mathsf{y}}^{(1)}(\cdot)$ and $\psi_{\mathsf{p},t}^{(1)}$ denote their first derivatives. Define $\mathcal{L}_{\ell,\mathsf{y}}(\kappa, y) = \frac{\partial^\ell}{\partial\kappa^\ell}\mathcal{L}_{\mathsf{y}}(\psi_{\mathsf{y}}(\kappa), y)$ for $\ell = 1, 2$, $\boldsymbol{\mathcal{L}}_{1,\mathsf{p}}(\boldsymbol{\kappa}, \boldsymbol{d}) = \frac{\partial}{\partial\boldsymbol{\kappa}}\mathcal{L}_{\mathsf{p}}(\boldsymbol{\psi}_{\mathsf{p}}(\boldsymbol{\kappa}), \boldsymbol{d})$, and $\boldsymbol{\mathcal{L}}_{2,\mathsf{p}}(\boldsymbol{\kappa}, \boldsymbol{d}) = \frac{\partial^2}{\partial\boldsymbol{\kappa}\partial\boldsymbol{\kappa}'}\mathcal{L}_{\mathsf{p}}(\boldsymbol{\psi}_{\mathsf{p}}(\boldsymbol{\kappa}), \boldsymbol{d})$.

Assumption 5 (Local QMLE)**.**

(a) For some fixed constant $\Delta > 0$*,*

$$\mathbb{E}\left[\max_{1\le i\le n}\frac{1}{K}\sum_{j\in\mathcal{N}_i}\sup_{|\tilde{\kappa}-\kappa|\le\Delta}\frac{|\mathcal{L}_{2,\mathsf{y}}(\tilde{\kappa},y_j)-\mathcal{L}_{2,\mathsf{y}}(\kappa,y_j)|}{|\tilde{\kappa}-\kappa|}\middle|\bar{\mathcal{F}},\{\boldsymbol{x}_i\}_{i=1}^n\right]\lesssim_{\mathbb{P}} 1,$$

and the same condition also holds for every element of $\boldsymbol{\mathcal{L}}_{2,\mathsf{p}}(\boldsymbol{\kappa},\boldsymbol{d})$*;*

(b) $\mathcal{L}_{2,\mathsf{y}}(\kappa,y)<0$ *for* $\kappa\in\mathbb{R}$ *and* $y\in\mathcal{Y}$*, and* $\boldsymbol{\mathcal{L}}_{2,\mathsf{p}}(\boldsymbol{\kappa},\boldsymbol{d})$ *is negative definite for* $\boldsymbol{\kappa}\in\mathbb{R}^J$ *and* $\boldsymbol{d}\in\{0,1\}^{J+1}$*;*

(c) For all $t\in\mathcal{J}$*,* $\psi_{\mathsf{y}}(\cdot)$*,* $\psi_{\mathsf{p},t}(\cdot)$*,* $V_{\mathsf{y}}(\cdot)$ *and* $V_{\mathsf{p},t}(\cdot)$ *are twice continuously differentiable,* $\psi_{\mathsf{y}}^{(1)}(\mu_t(\boldsymbol{\alpha})+\boldsymbol{z}'\boldsymbol{\beta}_t)$ *and* $\psi_{\mathsf{p},t}^{(1)}(\boldsymbol{e}(\boldsymbol{\alpha})+\boldsymbol{\Gamma}\boldsymbol{z})$ *are nonzero,* $V_{\mathsf{y}}(\psi_{\mathsf{y}}(\mu_t(\boldsymbol{\alpha})+\boldsymbol{z}'\boldsymbol{\beta}_t))>0$*, and* $V_{\mathsf{p},t}(\boldsymbol{e}(\boldsymbol{\alpha})+\boldsymbol{\Gamma}\boldsymbol{z})>0$ *over the support of* $\boldsymbol{\alpha}$ *and* $\boldsymbol{z}$*.*

4.2 Theoretical Results

Throughout the analysis below, we write $\delta_n=(K^{1/2}\wedge p^{1/2})/\sqrt{\log(n\vee p)}$. Recall that K is the number of nearest neighbors, p is the dimension of $\boldsymbol{x}_i$, and d_i is the number of leading local principal components extracted in the neighborhood of unit i. The asymptotic analysis is conducted assuming both K and p diverge as $n\to\infty$.

The following lemma shows that the nearest neighbors and extracted local factor loadings are informative about the unobserved confounders. The proof is available in Feng (2023).

Lemma 4.1. *Suppose that Assumptions 2(c), 2(d), 2(e) and 3 hold. If* $\frac{\log(n/K)}{K}=o(1)$ *and* $\frac{K\log n}{n}=o(1)$*, then,*

$$\max_{i\in[n]}\max_{k\in[K]}\|\boldsymbol{\alpha}_i-\boldsymbol{\alpha}_{j_k(i)}\|\lesssim_{\mathbb{P}}(K/n)^{\bar{\varsigma}/(\underline{\varsigma}\mathsf{d}_\alpha)}+a_n^{1/\underline{\varsigma}}.$$

If, in addition, Assumptions 4(a) and 4(b) hold and $(np)^{\frac{2}{\nu}}\delta_n^{-2}\lesssim 1$*, then for* $\ell\in[\mathsf{d}_i]$*, there* *exists* $\boldsymbol{R}_{\langle i\rangle}$ *such that*

$$\max_{i\in[n]}\|\widehat{\boldsymbol{\Lambda}}_{\cdot\ell,\langle i\rangle}-\boldsymbol{\Lambda}_{\langle i\rangle}\boldsymbol{R}_{\cdot\ell,\langle i\rangle}\|_{\max}\lesssim_{\mathbb{P}}\delta_n^{-1}+h_n^m,$$

where $1\lesssim_{\mathbb{P}} s_{\min}(\boldsymbol{R}_{\langle i\rangle})\leq s_{\max}(\boldsymbol{R}_{\langle i\rangle})\lesssim_{\mathbb{P}}1$.

The first result in this lemma precisely quantifies the matching discrepancy of nearest neighbors in terms of the unobserved confounders. The first term, $(K/n)^{\bar{\varsigma}/(\underline{\varsigma}\mathsf{d}_\alpha)}$, is the usual discrepancy that would have existed even if the matching could be done directly on $\boldsymbol{\alpha}_i$. By contrast, the second term $a_n^{1/\underline{\varsigma}}$ arises from the fact that the matching can only be conducted in terms of the noisy measurements $\boldsymbol{x}_i$ rather than $\boldsymbol{\alpha}_i$. Typically, $a_n\to 0$ when many measurements are available ($p\to\infty$), and since we assume $K/n\to 0$, the (maximum) matching discrepancy diminishes to 0 in large samples. In other words, the indirectly obtained nearest neighbors are close to the target unit i in terms of the unobservables.

The second result of Lemma 4.1 shows that the extracted factor loading $\widehat{\boldsymbol{\Lambda}}_{\langle i\rangle}$ is consistent for the underlying loading matrix $\boldsymbol{\Lambda}_{\langle i\rangle}$, up to a rotation matrix $\boldsymbol{R}_{\langle i\rangle}$, and the estimation error shrinks when both K and p grow large and K/n gets small. The loading $\boldsymbol{\Lambda}_{\langle i\rangle}$ characterizes the local relations of nearest neighbors in terms of the latent confounders and thus can be used to restore unconfoundedness in causal inference.

The following is our first main result, which shows that the conditional means of potential outcomes and the propensity scores can be consistently estimated by the proposed local quasi-maximum likelihood method.

Theorem 4.1 (Local QMLE)**.** *Suppose that Assumptions 2–5 hold. If* $(np)^{\frac{2}{\nu}}\delta_n^{-2}\lesssim 1$, *then for each* $t\in\mathcal{J}$,

$$|\widehat{\zeta}_{i,t}-\zeta_{i,t}|=\bar{O}_{\mathbb{P}}(\delta_n^{-1}+h_n^m+\mathfrak{r}_{\mathsf{y},\langle i\rangle})\quad\textit{and}\quad|\widehat{\varrho}_{i,t}-\varrho_{i,t}|=\bar{O}_{\mathbb{P}}(\delta_n^{-1}+h_n^m+\mathfrak{r}_{\mathsf{p},\langle i\rangle}).$$

Detailed asymptotic expansions are given by Equation (SA-2.1) *in the SA.*

$\widehat{\zeta}_{i,t}$ and $\widehat{\varrho}_{i,t}$ are nonparametric estimators of conditional mean functions in the sense that we do not specify the functional forms of μ_t and e_t, the components related to the unobserved confounders. Consequently, we obtain in Theorem 4.1 a nonparametric convergence rates: δ_n^{-1} is the variance term, and $\mathfrak{r}_{\mathsf{y},\langle i\rangle}$ and $\mathfrak{r}_{\mathsf{p},\langle i\rangle}$ reflect the smoothing bias. The major difference compared to usual nonparametric results is that the convergence rate above also depends on the number of measurements p and the approximation error h_n^m from the first step, which determine the quality of the information about $\boldsymbol{\alpha}_i$ used in the estimation. Also, the convergence result above is uniform over all the data points indexed by i, which respects the fact that $\boldsymbol{\alpha}_i$ is not directly observed and we obtain information on it for the n units in the dataset. In this sense, it slightly differs from the classical nonparametric or semiparametric estimation where uniformity over the whole support of conditioning variables is derived (or assumed directly).

We also emphasize that the proposed local QMLE is a general strategy that can be used to nonparametrically estimate quantities other than the two specific conditional mean functions in the potential outcomes framework. As mentioned before, the previous local PCA step provides all necessary ingredients for a (kernel-based) nonparametric analysis: the nearest neighbors allow us to define a local region for each unit, as a compactly supported kernel function does, and the estimated factor loadings, similar to a polynomial basis, play the role of the (local) approximation basis. Therefore, the two outputs, $\mathcal{N}_i$ and $\widehat{\mathbf{\Lambda}}_{\langle i\rangle}$, are readily used as inputs in other semiparametric or nonparametric estimation methods.

Now, we are ready to apply previous results to inference on the counterfactual means of potential outcomes. The following theorem, our second main result, establishes the asymptotic normality of the proposed estimator.

Theorem 4.2 (Causal Inference)**.** *Suppose that Assumptions 1–5 hold. If* $(np)^{\frac{2}{\nu}}\delta_n^{-2} \lesssim 1$ *and* $\sqrt{n}(\delta_n^{-2} + h_n^{2m} + \mathbb{E}_n[\mathfrak{r}^2_{\mathsf{y},\langle i\rangle}]\log n + \mathbb{E}_n[\mathfrak{r}^2_{\mathsf{p},\langle i\rangle}]) = o(1)$, *then*

(a) $\sqrt{n}(\widehat{\theta}_{t,t'} - \theta_{t,t'}) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \varphi_{i,t,t'} + o_{\mathbb{P}}(1)$ *where* $\varphi_{i,t,t'} = \frac{d_i(t')(\zeta_{i,t}-\theta_{t,t'})}{\varrho_{t'}} + \frac{\varrho_{i,t'}}{\varrho_{t'}}\frac{d_i(t)(y_i-\zeta_{i,t})}{\varrho_{i,t}}$;

(b) $\sqrt{n}(\widehat{\theta}_{t,t'} - \theta_{t,t'})/\widehat{\sigma}_{t,t'} \rightsquigarrow \mathsf{N}(0,1)$.

Theorem 4.2 takes advantage of the doubly-robust score function to relax the conditions on the convergence rates of $\widehat{\zeta}_{i,t}$ and $\widehat{\varrho}_{i,t}$. Specifically, the second rate restriction essentially requires the squares of the two estimation errors (and thus their product as well) be of an order smaller than $n^{-1/2}$, which is consistent with the results in the double/debiased machine learning literature. Importantly, the smoothing bias in estimating conditional mean functions of potential outcomes and propensity scores, characterized by $\mathfrak{r}_{\mathsf{y},\langle i\rangle}$ and $\mathfrak{r}_{\mathsf{p},\langle i\rangle}$ respectively, need *not* be small uniformly over all units. Instead, we only require the *average* squared bias be small, thus accommodating scenarios where the nearest neighbors and local factor loadings from the first step are not highly informative about the latent confounders and offer poor approximation power for some units.

As described before, K is the main tuning parameter to be selected in the whole estimation procedure. To get a sense of the rate restrictions imposed, consider a simple but practically relevant case where $K \lesssim p$, ν is sufficiently large, and all local factors "stronger" than the noise have been extracted (so that $h_n^m \lesssim \delta_n^{-1}$). Then, we roughly need $n/K^2 = o(1)$, up to $\log n$ terms, in addition to the previously discussed condition on the average squared smoothing bias. Note that if (4.1) also holds, the discussion in Remark 4.2 implies that $\mathfrak{r}_{\mathsf{y},\langle i\rangle}$ and $\mathfrak{r}_{\mathsf{p},\langle i\rangle}$ are $O(h_n^m)$ uniformly over $\langle i\rangle$, in which case all restrictions imposed in Theorem 4.2 reduces to $n/K^2 = o(1)$, up to $\log n$ terms.

Remark 4.3 (K Selection)**.** A formal procedure for selecting K is beyond the scope of this paper and left for future research, but the previous discussion implies that many feasible choices of K satisfy the above mild restrictions. For instance, one can consider the nearest-neighbors-based local $(m-1)$th-order polynomial regression of y_i on any r out of the p noisy measurements. Under standard regularity conditions, the optimal K in the MSE sense for this regression is $O(n^{\frac{2m}{2m+r}})$ if m is even, and is $O(n^{\frac{2m+2}{2m+2+r}})$ if m is odd, which grows faster than $n^{1/2}$ if $m > r/2$ or $m > r/2-1$. Then, usual plug-in or cross-validation methods in the nonparametric regression literature can be employed to select K, though such choices may not be optimal for the final inference purpose. $\lrcorner$

Note that this paper focuses on large-K asymptotics, which is analogous to a large bandwidth in kernel estimation or a small number of approximation terms in series estimation. If K is small relative to the sample size, a non-negligible undersmoothing bias may arise in the distributional approximation, and bias-robust inference may be needed. See Cattaneo and Jansson (2018); Cattaneo, Jansson and Ma (2019); Matsushita and Otsu (2021) for more discussions of undersmoothing bias and possible solutions.

4.3 Numerical Results

We conducted a Monte Carlo investigation of the finite sample performance of the proposed method. We consider a binary treatment design $\mathcal{J} = \{0, 1\}$. The potential outcomes are $y_i(0) = \alpha + \alpha^2 + \epsilon_{i,0}$ and $y_i(1) = 2\alpha + \alpha^2 + 1 + \epsilon_{i,1}$. The treatment is $s_i = \mathbb{1}(v_i \leq \varrho_i)$ where $\varrho_i = \exp((\alpha - 0.5) + (\alpha - 0.5)^2)/(1 + \exp((\alpha - 0.5) + (\alpha - 0.5)^2))$. The observed covariates are generated based on $x_{il} = \eta_l(\alpha_i) + u_{il}$. To check the performance of our local approximation strategy in different scenarios, we consider the following three specifications of η_l:

- Model 1: $\eta_l(\alpha_i) = (\alpha_i - \varpi_l)^2$
- Model 2: $\eta_l(\alpha_i) = \frac{1}{0.1\sqrt{2\pi}} \exp(-100(\alpha_i - \varpi_l)^2)$
- Model 3: $\eta_l(\alpha_i) = \exp(-10|\alpha_i - \varpi_l|)$

Model 1 is highly smooth and has relatively low degree of nonlinearity, while the other two are more nonlinear, making the smoothing bias more pronounced (given the number of nearest neighbors K and the number of the extracted local factors d_i). Moreover, we let $\epsilon_{i,0}, \epsilon_{i,1} \sim \mathsf{N}(0, 1)$, $\alpha_i, \varpi_l, v_i \sim \mathsf{U}(0, 1)$, $u_{il} \sim \mathsf{N}(0, .5^2)$ and is i.i.d over i and l. The sequences $\{\epsilon_{i,0}\}$, $\{\epsilon_{i,1}\}$, $\{\alpha_i\}$, $\{\varpi_l\}$, $\{v_i\}$ and $\{u_{il}\}$ are independent.

We consider 2 000 simulated datasets with $n = p = 1\,000$ each. For each simulated dataset, a point estimate of the counterfactual mean $\theta_{0,1} = \mathbb{E}[y_i(0)|s_i = 1]$ is obtained. We report bias (BIAS), standard deviation (SD), root mean squared error (RMSE), coverage rate (CR) of the nominal 95% confidence interval and its average length (AL) in Table 3.

The results in the first four columns are obtained based on our proposed method, using the same strategy as described in our empirical illustration. Specifically, we take the first half ($p^{\dagger} = 500$) of measurements for nearest neighbors matching and the second half ($p^{\ddagger} = 500$) for extracting the local factor loadings. The number of nearest neighbors is taken to be $K = Cn^{4/5}$ for $C = 0.5$, 1, 1.5 respectively. This rate is the MSE-optimal for the (infeasible) cross-sectional local constant/linear nearest neighbors regression of y_i on α_i and satisfies the restriction for our inference theory. We also consider a data-dependent choice $\widehat{K}_{\mathsf{CV}}$, which is obtained by 20-fold cross validation for the local nearest neighbors regression of y_i on one noisy measurement x_{ip}. Analogously to usual nonparametric regression analysis, we take K as the main tuning parameter and prefer to take a relatively small number of local factors. Thus, a simple rule is employed to determine d_i: for each unit $i \in [n]$, set $\mathsf{d}_i = 2$ if $\upsilon_{2,\langle i \rangle}/\upsilon_{3,\langle i \rangle} \geq \log\log K$ and $\mathsf{d}_i = 1$ otherwise.

For comparison, we also consider another approximation strategy: the conditional means of potential outcomes and propensity scores are estimated by taking the local averages of nearest neighbors, which is analogous to local constant smoothing in nonparametric regression analysis. Such ideas have been proposed in the panel and network data literature for different purposes (e.g. Zhang et al., 2017; Bonhomme et al., 2022). Results are reported in Columns 5–8 ("local average").

The proposed method performs relatively well throughout the three specifications. In Model 1 the coverage rate is close to the nominal level and is robust to different choices of K. In the other two models with high degrees of nonlinearity, our method still has satisfactory performance, but as expected, using a large K may lower the coverage rate, showcasing the importance of "localization" by K-NN matching in the first step. By contrast, the local average method suffers from more severe undercoverage. Even in Model 1 where nonlinearity is mild, the nominal level is not achieved. Also, the performance of local average varies considerably across different choices of K.

Table 3: Simulation Results, $n = p = 1\,000$, $2\,000$ replications

	Local PCA, $K =$				**Local average,** $K =$			
	125	251	376	$\widehat{K}_{\texttt{CV}}$	125	251	376	$\widehat{K}_{\texttt{CV}}$
Model 1								
BIAS	−0.007	−0.006	−0.005	−0.006	−0.014	−0.021	−0.035	−0.025
SD	0.057	0.056	0.055	0.056	0.054	0.053	0.052	0.053
RMSE	0.057	0.056	0.056	0.056	0.056	0.057	0.062	0.059
CR	0.953	0.959	0.956	0.957	0.931	0.916	0.873	0.899
AL	0.228	0.226	0.226	0.226	0.205	0.199	0.194	0.198
Model 2								
BIAS	−0.002	−0.014	−0.027	−0.009	−0.001	−0.015	−0.040	−0.010
SD	0.054	0.054	0.055	0.055	0.054	0.054	0.053	0.055
RMSE	0.054	0.056	0.061	0.055	0.054	0.056	0.067	0.056
CR	0.953	0.950	0.908	0.947	0.948	0.925	0.864	0.933
AL	0.220	0.220	0.213	0.219	0.213	0.206	0.201	0.209
Model 3								
BIAS	−0.021	−0.028	−0.037	−0.030	−0.014	−0.029	−0.050	−0.035
SD	0.055	0.054	0.054	0.055	0.054	0.053	0.052	0.054
RMSE	0.059	0.061	0.066	0.063	0.056	0.060	0.072	0.064
CR	0.934	0.918	0.883	0.905	0.933	0.897	0.818	0.869
AL	0.217	0.213	0.210	0.212	0.207	0.201	0.197	0.200

Notes: SD = standard deviation of point estimator, RMSE = root MSE of point estimator, CR = coverage rate of 95% nominal confidence intervals, AL = average interval length of 95% nominal confidence intervals. $\widehat{K}_{\texttt{CV}}$= cross-validation choice of K.

5 Extension: Uniform Inference

In many applications, the outcome of interest is a certain transformation of the original potential outcome via a function $g(\cdot) \in \mathcal{G}$, and uniform inference over the function class $\mathcal{G}$ is desired. In general, the goal can be achieved in two steps: (i) strengthen the asymptotic expansion in Theorem 4.2(a) to be uniform, that is, the remainder needs to be negligible uniformly over $g \in \mathcal{G}$; and (ii) show that the influence function as a process indexed by $\mathcal{G}$ weakly converges to a limiting process. The general treatment of such issues can be found in, e.g., Barrett and Donald (2003); Chernozhukov et al. (2013); Donald and Hsu (2014).

We will focus on counterfactual distributions, the analysis of which relies on a particular function class $\mathcal{G} = \{y \mapsto \mathbb{1}(y \leq \tau) : \tau \in \mathcal{Y}\}$. Each $g(\cdot) \in \mathcal{G}$ corresponds to a particular value $\tau \in \mathcal{Y}$. Therefore, we write $y_{i,\tau}(t) = \mathbb{1}(y_i(t) \leq \tau)$ and $y_{i,\tau} = \mathbb{1}(y_i \leq \tau)$. Accordingly,

Equation (2.2) becomes

$$y_{i,\tau}(t) = \zeta_{i,t,\tau} + \epsilon_{i,t,\tau}, \quad \zeta_{i,t,\tau} = \mu_{t,\tau}(\boldsymbol{\alpha}_i) + \boldsymbol{z}_i'\boldsymbol{\beta}_{t,\tau},$$

where $\zeta_{i,t,\tau} = \mathbb{P}(y_i(t) \le \tau | \boldsymbol{z}_i, \boldsymbol{\alpha}_i)$. For each τ, the second step of the estimation procedure in Section 3 is implemented to obtain an estimator $\widehat{\zeta}_{i,t,\tau}$ of $\zeta_{i,t,\tau}$. The parameter of interest is $\theta_{t,t'}(\tau) = \mathbb{E}[\mathbb{1}(y_i(t) \le \tau) | t_i = t']$, the counterfactual distribution function of the potential outcome $y_i(t)$ for the group with a treatment status t'. From the perspective of uniform inference, $\theta_{t,t'}(\cdot)$ is a parameter in $\ell^\infty(\mathcal{Y})$, a function space of bounded functions on $\mathcal{Y}$ equipped with sup-norm. To establish the limiting distribution of the proposed estimator, we slightly strengthen Assumptions 2(a) and 5(a) by imposing Assumption 6 below.

Assumption 6 (Local QMLE, Uniform Inference)**.**

(a) *For some fixed constant $\Delta > 0$,*

$$\mathbb{E}\left[\sup_{\tau \in \mathcal{Y}} \max_{1 \le i \le n} \frac{1}{K} \sum_{j=1}^{n} \sup_{|\tilde{\kappa} - \kappa| \le \Delta} \frac{|(\mathcal{L}_{2,\mathsf{y}}(\tilde{\kappa}, y_{j,\tau}) - \mathcal{L}_{2,\mathsf{y}}(\kappa, y_{j,\tau}))| \mathbb{1}(j \in \mathcal{N}_i)}{|\tilde{\kappa} - \kappa|} \middle| \bar{\mathcal{F}} \right] \lesssim_{\mathbb{P}} 1;$$

(b) *For all $\tau \in \mathcal{Y}$, $\mu_{t,\tau}(\cdot)$ is $\bar{m}$-times continuously differentiable with all partial derivatives of order no greater than $\bar{m}$ bounded by a universal constant, and $\mu_{t,\tau}(\cdot)$ is Lipschitz with respect to τ uniformly over $\mathcal{A}$.*

The following theorem shows that the (scaled) counterfactual distribution process weakly converges to a limiting Gaussian process indexed by $\tau \in \mathcal{Y}$, which forms the basis of uniform inference. See Van Der Vaart and Wellner (1996) for underlying technical details.

Theorem 5.1 (Uniform Inference)**.** *Under Assumptions 1–6, if $(np)^{\frac{2}{\nu}} \delta_n^{-2} \lesssim 1$ and $\sqrt{n}(\delta_n^{-2} + h_n^{2m} + \mathbb{E}_n[\mathfrak{r}^2_{\mathsf{y},\langle i \rangle}] \log n + \mathbb{E}_n[\mathfrak{r}^2_{\mathsf{p},\langle i \rangle}]) = o(1)$, then*

$$\sqrt{n}\Big(\theta_{t,t'}(\cdot) - \theta_{t,t'}(\cdot)\Big) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi_{i,t,t'}(\cdot) + o_{\mathbb{P}}(1) \rightsquigarrow \mathsf{Z}_{t,t'}(\cdot) \quad \textit{in} \;\; \ell^\infty(\mathcal{Y}),$$

where $\varphi_{i,t,t'}(\cdot) = \frac{d_i(t')(\zeta_{i,t,\cdot} - \theta_{t,t'}(\cdot))}{\varrho_{t'}} + \frac{\varrho_{i,t'}}{\varrho_{t'}} \frac{d_i(t)(y_{i,\cdot} - \zeta_{i,t,\cdot})}{\varrho_{i,t}}$ *and* $\mathsf{Z}_{t,t'}(\cdot)$ *is a zero-mean Gaussian process with covariance kernel* $\mathbb{E}[\varphi_{i,t,t'}(\tau_1)\varphi_{i,t,t'}(\tau_2)]$ *for* $\tau_1, \tau_2 \in \mathcal{Y}$*.*

Under proper regularity conditions, the weak convergence above can be applied to construct inference procedures for other quantities such as quantile treatment effects by the functional delta method, as shown in the following corollary.

Corollary 5.1.1 (Functional Delta Method)**.** *Let the conditions in Theorem 5.1 hold. Consider the parameter* θ *as an element of a parameter space* $D\theta \subseteq \ell^\infty(\mathcal{Y})$*, a function space of bounded functions on* $\mathcal{Y}$*, with* $D\theta$ *containing the true value* $\theta_{t,t'}$*. Let a functional* $\Psi(\theta)$ *mapping* $D\theta$ *to* $\ell^\infty(\mathcal{Q})$ *be Hadamard differentiable in* θ *at* $\theta_{t,t'}$ *with derivative* Ψ'_θ*. Then,* $|\sqrt{n}(\Psi(\widehat{\theta}_{t,t'})(\cdot) - \Psi(\theta_{t,t})(\cdot)) - n^{-1/2}\sum_{i=1}^n \Psi'_\theta(\varphi_{i,t,t'})(\cdot)| = o_{\mathbb{P}}(1)$ *and* $\sqrt{n}(\Psi(\widehat{\theta}_{t,t'})(\cdot) - \Psi(\theta_{t,t'})(\cdot)) \rightsquigarrow \Psi'_\theta(\mathsf{Z}_{t,t'})(\cdot)$ *that is a Gaussian process in* $\ell^\infty(\mathcal{Q})$ *with mean zero and covariance kernel defined by the limit of the second moment of* $\Psi'_\theta(\varphi_{i,t,t'})$*.*

The limiting Gaussian process can be approximated based on a practically feasible multiplier bootstrap procedure widely used in the literature. To be specific, take an i.i.d. sequence of random variables $\{\omega_i\}_{i\in[n]}$ independent of the data with mean zero and variance one. Define a uniformly consistent estimator of $\varphi_{i,t,t'}(\cdot)$:

$$\widehat{\varphi}_{i,t,t'}(\cdot) = \frac{d_i(t')(\widehat{\zeta}_{i,t,\cdot} - \widehat{\theta}_{t,t'}(\cdot))}{\widehat{\varrho}_{t'}} + \frac{\widehat{\varrho}_{i,t'}}{\widehat{\varrho}_{t'}} \frac{d_i(t)(y_{i,\cdot} - \widehat{\zeta}_{i,t,\cdot})}{\widehat{\varrho}_{i,t}}.$$

The following corollary shows that conditional on the data, $\frac{1}{\sqrt{n}}\sum_{i=1}^n \omega_i \widehat{\varphi}_{i,t,t'}(\cdot)$ weakly converges to the same limiting process $\mathsf{Z}_{t,t'}(\cdot)$ as in Theorem 5.1. In practice, one only needs to simulate this feasible approximation process by taking random draws of $\{\omega_i\}_{i\in[n]}$.

Corollary 5.1.2 (Multiplier Bootstrap)**.** *Let the conditions of Theorem 5.1 hold. Then, conditional on the data,* $n^{-1/2}\sum_{i=1}^n \omega_i \widehat{\varphi}_{i,t,t'}(\cdot) \rightsquigarrow \mathsf{Z}_{t,t'}(\cdot)$ *that is the Gaussian process defined in Theorem 5.1 with probability approaching one.*

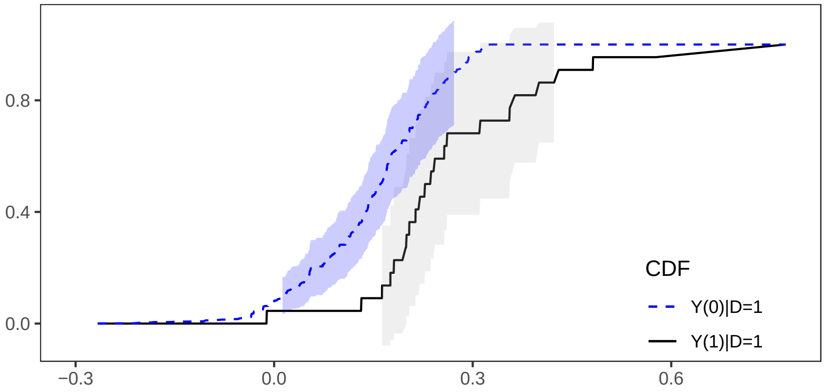
To showcase the uniform inference procedure, we use the data of Acemoglu et al. (2016) to check the (first-order) stochastic dominance (SD) of $\theta_{1,1}(\cdot)$ over $\theta_{0,1}(\cdot)$, where $\theta_{1,1}(\cdot)$ and

$\theta_{0,1}(\cdot)$ respectively denote the cumulative distribution functions (CDFs) of potential stock returns of firms connected to Geithner if they were connected and not connected with him. The main ideas are outlined here. By definition of SD, the null hypothesis is $\theta_{1,1}(\tau) \leq \theta_{0,1}(\tau)$ for all $\tau \in \mathcal{Y}$. An intuitive test statistic is $\sqrt{n}\sup_{\tau\in\mathcal{Y}}(\theta_{1,1}(\tau) - \theta_{0,1}(\tau))$. The null hypothesis is rejected if the test statistic is greater than a certain critical value. Given the asymptotic expansions of $\widehat{\theta}_{1,1}(\cdot)$ and $\widehat{\theta}_{0,1}(\cdot)$, the critical value can be obtained by simulating the supremum of the approximation process, i.e., $\sup_{\tau\in\mathcal{Y}}(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\widehat{\varphi}_{i,1,1}(\tau) - \widehat{\varphi}_{i,0,1}(\tau)))$. In practice, the supremum over the whole support is simply replaced by maximum over a set of user-specified evaluation points.

For each $\tau \in \mathcal{Y}$, implement the estimation procedure as described in Section 3. Varying the values of τ, we obtain two estimated distribution functions for firms with connections, as shown in Figure 3. The treated outcome $Y(1)$ is the potential cumulative return with connections to Geithner and the untreated outcome $Y(0)$ refers to that without connections. To better understand the estimation uncertainty, 95% confidence bands for the two estimated CDFs are plotted, which are based on simulating the maximum absolute value of the corresponding (studentized) approximation processes. In each case, the value of τ is restricted to range from 0.1-quantile to 0.9-quantile of the estimated distribution.

It turns out that the estimated CDF for the treated outcome is well below that for the untreated outcome. Formally, we take all observed values of the cumulative returns as the evaluation points, and then simulate the maximum of the approximation process by taking 500 draws of random weights $\{\omega_i\}_{i=1}^{n}$. In this simple example, the test statistic equals 0, which is well below the critical value 7.23 for a confidence level of 0.95 obtained through simulation. Thus, SD of $\theta_{1,1}(\cdot)$ over $\theta_{0,1}(\cdot)$ cannot be rejected. It implies that the positive effects of political connections in this example are felt over the entire distribution of the stock returns of financial firms connected with Geithner, which is a stronger conclusion than that based simply on the mean in Section 3.

Figure 3: Estimated CDFs of Potential Outcomes for the Treated



6 Conclusion

This paper has developed a causal inference method for treatment effects models with some confounders not directly observed. Relevant information on these latent confounders is extracted from a large set of noisy measurements that admits an unknown, possibly nonlinear factor structure. Such information is then used to match comparable units in the subsequent counterfactual analysis. Large-sample properties of the proposed estimators are established. The results cover a large class of causal parameters, including average treatment effects and counterfactual distributions. The method is illustrated with an empirical application studying the effect of political connections on stock returns of financial firms.

References

Abadie, A. (2021), "Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects," *Journal of Economic Literature*, 59, 391–425.

Abadie, A., and Cattaneo, M. D. (2018), "Econometric Methods for Program Evaluation," *Annual Review of Economics*, 10, 465–503.

Abadie, A., and Imbens, G. W. (2006), "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica*, 74, 235–267.

Abadie, A., and L'Hour, J. (2021), "A Penalized Synthetic Control Estimator for Disaggregated Data," *Journal of the American Statistical Association*, 116, 1817–1834.

Acemoglu, D., Johnson, S., Kermani, A., Kwak, J., and Mitton, T. (2016), "The Value of Connections in Turbulent Times: Evidence from the United States," *Journal of Financial Economics*, 121, 368–391.

Altonji, J. G., and Mansfield, R. K. (2018), "Estimating Group Effects Using Averages of Observables to Control for Sorting on Unobservables: School and Neighborhood Effects," *American Economic Review*, 108, 2902–46.

Arias-Castro, E., Lerman, G., and Zhang, T. (2017), "Spectral Clustering Based on Local PCA," *Journal of Machine Learning Research*, 18, 253–309.

Athey, S., and Imbens, G. W. (2022), "Design-based Analysis in Difference-in-Differences Settings with Staggered Adoption," *Journal of Econometrics*, 226, 62–79.

Bai, J. (2003), "Inferential Theory for Factor Models of Large Dimensions," *Econometrica*, 71, 135–171.

——— (2009), "Panel Data Models with Interactive Fixed Effects," *Econometrica*, 77, 1229–1279.

Bai, J., and Ng, S. (2006), "Confidence Intervals for Diffusion Index Forecasts and Inference for Factor-Augmented Regressions," *Econometrica*, 74, 1133–1150.

Bai, J., and Wang, P. (2016), "Econometric Analysis of Large Factor Models," *Annual Review of Economics*, 8, 53–80.

Barrett, G. F., and Donald, S. G. (2003), "Consistent Tests for Stochastic Dominance," *Econometrica*, 71, 71–104.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014), "Inference on Treatment Effects after Selection among High-Dimensional Controls," *The Review of Economic Studies*, 81, 608–650.

Bishop, C. M. (2006), "Continuous Latent Variables," in *Pattern Recognition and Machine Learning*, chapter 12, New York: Springer, pp. 559–604.

Bonhomme, S., Lamadon, T., and Manresa, E. (2022), "Discretizing Unobserved Heterogeneity," *Econometrica*, 90, 625–643.

Cattaneo, M. D. (2010), "Efficient Semiparametric Estimation of Multi-valued Treatment Effects under Ignorability," *Journal of Econometrics*, 155, 138–154.

Cattaneo, M. D., and Jansson, M. (2018), "Kernel-Based Semiparametric Estimators: Small Bandwidth Asymptotics and Bootstrap Consistency," *Econometrica*, 86, 955–995.

Cattaneo, M. D., Jansson, M., and Ma, X. (2019), "Two-Step Estimation and Inference with Possibly Many Included Covariates," *The Review of Economic Studies*, 86, 1095–1122.

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, C1–C68.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., and Robins, J. M. (2022), "Locally Robust Semiparametric Estimation," *Econometrica*, 90, 1501–1535.

Chernozhukov, V., Fernández-Val, I., and Melly, B. (2013), "Inference on Counterfactual Distributions," *Econometrica*, 81, 2205–2268.

Cunha, F., Heckman, J. J., and Schennach, S. M. (2010), "Estimating the Technology of Cognitive and Noncognitive Skill Formation," *Econometrica*, 78, 883–931.

Donald, S. G., and Hsu, Y.-C. (2014), "Estimation and Inference for Distribution Functions and Quantile Functions in Treatment Effect Models," *Journal of Econometrics*, 178, 383–397.

Fan, J., and Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications*, London: Chapman & Hall.

Farrell, M. H. (2015), "Robust Inference on Average Treatment Effects with Possibly More Covariates than Observations," *Journal of Econometrics*, 189, 1–23.

Feng, Y. (2023), "Optimal Estimation of Large-Dimensional Nonlinear Factor Models," *arXiv preprint arXiv:2311.07243.*

Härdle, W. K., Müller, M., Sperlich, S., and Werwatz, A. (2004), *Nonparametric and Semi-*

parametric Models, Berlin: Springer.

Heckman, J. J., Humphries, J. E., and Veramendi, G. (2018), "Returns to Education: The Causal Effects of Education on Earnings, Health, and Smoking," *Journal of Political Economy*, 126, S197–S246.

Hu, Y., and Schennach, S. M. (2008), "Instrumental Variable Treatment of Nonclassical Measurement Error Models," *Econometrica*, 76, 195–216.

Matsushita, Y., and Otsu, T. (2021), "Jackknife Empirical Likelihood: Small Bandwidth, Sparse Network and High-dimensional Asymptotics," *Biometrika*, 108, 661–674.

Miao, W., Geng, Z., and Tchetgen Tchetgen, E. J. (2018), "Identifying Causal Effects with Proxy Variables of an Unmeasured Confounder," *Biometrika*, 105, 987–993.

Nagasawa, K. (2022), "Treatment Effect Estimation with Noisy Conditioning Variables," *arXiv preprint arXiv:1811.00667*.

Robins, J. M., and Rotnitzky, A. (1995), "Semiparametric Efficiency in Multivariate Regression Models with Missing Data," *Journal of the American Statistical Association*, 90, 122–129.

Schennach, S. M. (2016), "Recent Advances in the Measurement Error Literature," *Annual Review of Economics*, 8, 341–377.

Stock, J. H., and Watson, M. W. (2002), "Forecasting Using Principal Components from a Large Number of Predictors," *Journal of the American Statistical Association*, 97, 1167–1179.

Van Der Vaart, A. W., and Wellner, J. A. (1996), *Weak Convergence and Empirical Processes*, Springer Series in Statistics, New York: Springer.

Wang, W., and Fan, J. (2017), "Asymptotics of Empirical Eigenstructure for High Dimensional Spiked Covariance," *Annals of Statistics*, 45, 1342–1374.

Zhang, Y., Levina, E., and Zhu, J. (2017), "Estimating Network Edge Probabilities by Neighbourhood Smoothing," *Biometrika*, 104, 771–783.

Zhang, Z., and Zha, H. (2004), "Principal Manifolds and Nonlinear Dimensionality Reduction via Tangent Space Alignment," *SIAM Journal on Scientific Computing*, 26, 313–338.