Causal Inference in Possibly Nonlinear Factor Models *

Yingjie Feng†

Abstract

This paper develops a general causal inference method for treatment effects models with noisily measured confounders. The key feature is that a large number of noisy proxies are available and 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, and their large-sample properties are established. We require relatively mild conditions on local principal subspace approximation, allow a mixture of informative and uninformative proxies with unknown identities, and permit measurement error to be possibly correlated with the outcome or treatment. We also obtain uniformly consistent estimators of the conditional average treatment effect given the latent confounders. 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 proxies, nonlinear factor model, heterogeneous treatment effects, doubly-robust estimator, high-dimensional data

JEL: C14, C21, C23, C38

*Feng gratefully acknowledges financial support from the National Natural Science Foundation of China (NSFC) through grants 72203122 and 72522001.

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

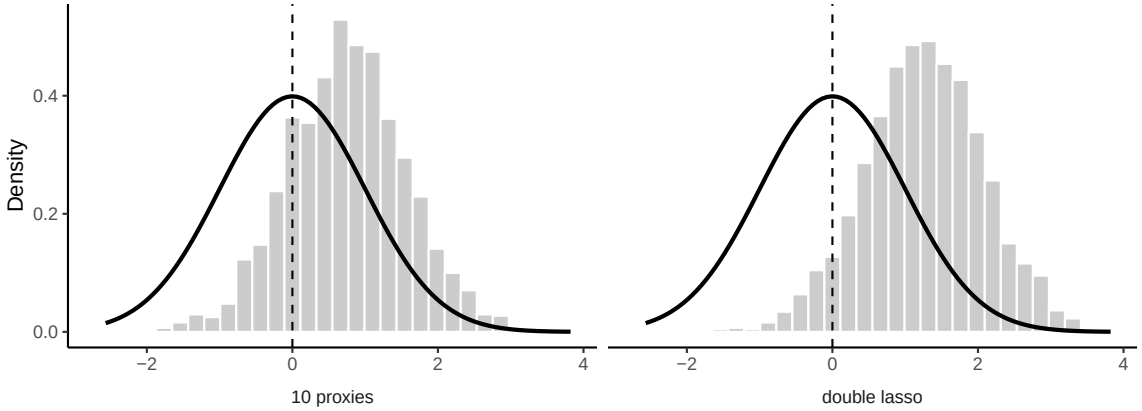
1 Introduction

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, which make it difficult to uncover the true causal relation between the outcome and the treatment. Nowadays, with the advent of big data, a researcher may have access to a large number of *noisy* proxies for these latent confounders, possibly collected over many time periods (large panel data), across many products or product categories (recommendation systems), or from interactions with many other units (large network data). For instance, in finance applications one may be concerned about the confounding effect of a firm's unobserved entrepreneurial ability or management quality, which may be imperfectly captured by a firm's historical performance such as stock returns (Holmström and Tirole, 1993; Pan et al., 2015). It is tempting to ask in this context whether and how high-dimensional proxies can mitigate confounding in causal inference.

Unfortunately, due to the presence of measurement error, usual selection-on-observables approaches that naively control for a number of selected noisy proxies are generally invalid. This remains true even if modern machine learning techniques such as post double Lasso (Belloni et al., 2014) are employed for control variables selection. Figure 1 shows a simple Monte Carlo experiment with 250 noisy proxies. The left panel shows the sampling distribution of the t-statistic for the conventional doubly robust estimator of the average treatment effect on the treated (ATT) when we control for the first 10 noisy proxies, and the right panel shows that for the ATT estimator using double Lasso for variable selection. Both deviate substantially from the standard normal. This occurs because sparsity-based techniques select a relatively small number of noisy proxies, leading to a non-vanishing bias caused by measurement error. A more detailed explanation is available in Section 2.1 below.

On the other hand, the unclear, possibly nonlinear relationship between noisy proxies and latent confounders makes this problem even more challenging. Traditional solutions

Figure 1: ATT Estimators Based on Selected Proxies



Notes: This figure shows the sampling distribution of t-statistics for ATT estimators based on selected proxies. The outcome and treatment are generated as described in Section 4.3. There are 250 noisy proxies $x_{il} = \alpha_i + u_{il}$, $l = 1, \cdots, 250$, with α_i a scalar latent confounder and u_{il} a standard normal error. The dashed vertical line marks zero. The solid curve is the standard normal density.

to measurement error in the literature often rely on strong functional form assumptions such as linearity (see, e.g., Schennach, 2016, for a review). However, in many applications confounders such as "ability" or "quality" are unobserved and often loosely defined; it is difficult or even impossible to specify how each proxy (e.g., the stock return in each period) relates to the latent confounders. In addition, when proxies are discrete, censored, truncated, or otherwise observed with limitations, nonlinear models arise naturally; one example is the graphon model, which characterizes link formation in network studies (Gao et al., 2015).

This paper proposes a novel approach for treatment effects analysis that addresses these issues in settings with high-dimensional proxies. Our key assumption is that the noisy proxies are linked with the latent confounders via an unknown, possibly nonlinear factor model. The proposed method builds on a local principal component analysis (PCA) method developed in Feng (2024). We first find the K-nearest neighbors (K-NN) for each unit based on observed proxies (with K diverging as the sample size grows). If different values of latent confounders can induce non-negligible differences in many (but possibly not all) observed proxies, then the K nearest neighbors, appropriately measured by the noisy proxies, are also 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 PCA.

As in linear factor models (Bai, 2003), the values of the latent confounders cannot be exactly recovered without additional normalizations. Nevertheless, the nearest neighbors and estimated local factor loadings from local PCA provide enough information to restore unconfoundedness in our treatment effects analysis. Intuitively, the nearest neighbors tell us, for instance, which control units are "close" to a particular treated unit in terms of the latent confounders, and the local factor loadings further characterize their distance. Such information suffices for usual causal analysis where *predictions* based on latent confounders are needed. In particular, we propose a general 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 functions of the latent. 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 a small fixed number.

In contrast with standard nonparametric regression, the conditioning variables in this context are *indirectly* obtained from proxies, 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 (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 intermediate nonparametric estimators. On the other hand, we do *not* require sample splitting across units that separates nuisance function estimation from final inference, which is similar to Cattaneo (2010); Farrell (2015); Farrell et al. (2021), but differs from the common practice in double machine learning. This is made possible by carefully investigating the specific structure and the asymptotic expansion of our nonparametric ingredients (nearest-neighbor-based local smoothing estimators).

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. In addition to the robustness with respect to the relationship between noisy proxies and latent confounders discussed before, our method has other 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 high-dimensional noisy proxies. It only requires some but *not all* proxies to be informative about latent confounders, and it is *unnecessary* to know their identities a priori (see Remark 4.2 below). Second, our theory builds on a generic choice of the "distance" to define nearest neighbors, accommodating and extending previous matching strategies 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 provides more flexible approximation to the nonlinear factor structure and delivers more robust inference results, compared to local constant smoothing based solely on matching techniques. Third, thanks to the availability of high-dimensional data, we can obtain a uniformly consistent estimator of the conditional average treatment effect (CATE) given latent confounders, and also relax

some restrictive assumptions on the independence between measurement error and outcomes (or treatment) that are common in the prior literature. 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 high-dimensional proxies for 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 (SA hereafter) contains several extensions of our theory, discussions of key assumptions used in the paper, all theoretical proofs, and additional numerical and technical results. Replications of the simulation study and empirical illustration are available at `https://github.com/yingjieum/replication-Feng_2026`.

1.1 Related Literature

This paper contributes to the large literature on causal inference and program evaluation (see Abadie and Cattaneo, 2018 for a review). Our theoretical framework is closely related to nonlinear models with measurement errors. Much effort has been devoted to the identification of such models (Schennach, 2016). For example, factor models can be utilized to construct repeated measurements of unobserved variables, which allows for 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). A related strand of work uses a fixed number of noisy proxies to construct control

functions (Altonji and Mansfield, 2018; Nagasawa, 2022) or confounding bridge functions (Miao et al., 2018; Singh, 2020), enabling identification of treatment effects provided that the proxies are sufficiently informative and satisfy certain conditional independence restrictions. By comparison, this paper considers the alternative setting with many noisy proxies and illustrates the gains from high-dimensional data that are typically unavailable in prior studies. In particular, identification of causal effects of interest does not require recovering the exact values (or distributions) of latent confounders; we permit a mixture of informative and uninformative proxies with unknown identities, and allow potential correlation between the measurement error and the outcome (or treatment); we also obtain uniformly consistent estimators of the conditional average treatment effect given the latent confounders.

The idea of local PCA, the technical building block of the inference method in this paper, has precedents in the machine learning literature (e.g., Zhang and Zha, 2004; Arias-Castro et al., 2017). These studies focus on the reconstruction of a global latent nonlinear structure underlying the data for the purpose of classification, clustering or data visualization, but do not provide formal estimation and inference theory. Recently, building on and extending results on large-dimensional factor models (Bai and Wang, 2016; Wang and Fan, 2017), Feng (2024) has established the estimation theory for local PCA based on general neighborhood formulation in the nonlinear factor model, and applied it to optimal matrix estimation problems. Unlike these studies, in this paper recovering the nonlinear factor structure underlying the noisy proxies is not of direct interest; rather, we exploit it to provide formal estimation and inference theory for a large class of causal parameters. We also address a challenge due to heterogeneous approximation errors, which is more specific to this causal inference problem (see Remark 4.3 below). The latent heterogeneity may also be addressed via other machine learning techniques. For instance, 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 informativeness of proxies. By contrast, this paper relies on a more general, weaker informativeness requirement (see Assumption 4 and

Remark 4.2 for details) and can achieve more flexible approximation of smooth functions of latent features via local PCA. The intermediate result (Theorem 4.1) also characterizes the uniform convergence of nonparametric estimators of individual-specific features and may be of independent interest.

2 Basic Framework

In this section we consider a binary treatment effects model with latent confounders that suffices to communicate the key ideas of the proposed method. A more general framework with several extensions, including multi-valued treatment effects, general likelihood functions, and covariates-adjusted nonlinear factor models, is presented in the SA.

Suppose that the observed data is a random sample $\{(y_i, t_i, \boldsymbol{x}_i, \boldsymbol{z}_i)\}_{i=1}^n$, where $y_i \in \mathbb{R}$ is the outcome of interest, $t_i \in \{0, 1\}$ is a binary treatment, and $\boldsymbol{x}_i \in \mathbb{R}^p$ and $\boldsymbol{z}_i \in \mathbb{R}^{r_z}$ are vectors of covariates. $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ play different roles: $\boldsymbol{x}_i$ is a vector of noisy proxies for some *unobserved* confounders $\boldsymbol{\alpha}_i \in \mathbb{R}^{r_\alpha}$, 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 $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ share some variables in common. The asymptotic theory in this paper is developed assuming the sample size n and the number of noisy proxies p simultaneously increase to infinity, whereas the number of confounders $r_\alpha + r_z$ is fixed.

We follow the standard potential outcomes framework. Let $y_i(1)$ and $y_i(0)$ denote unit i's potential outcome under treatment and that under no treatment respectively, and the observed outcome can be written as $y_i = t_i y_i(1) + (1 - t_i) y_i(0)$. For the moment, we focus on the analysis of the ATT parameter $\mathbb{E}[y_i(1) - y_i(0)|t_i = 1]$, and defer extensions to other causal parameters, including quantile and distributional treatment effects, to Section 5.

For the treated group, the treated outcome $y_i(1)$ is observed while the untreated one $y_i(0)$ is not. Thus, we focus on the key challenge of identifying $\mathbb{E}[y_i(0)|t_i = 1]$. This missing data problem is common in causal inference and is often overcome by imposing an unconfounded-

ness condition—the treatment assignment becomes independent of potential outcomes after conditioning on a set of covariates. In this paper we assume that

$$(y_i(1), y_i(0)) \perp\!\!\!\perp t_i \,|\, \boldsymbol{\alpha}_i, \boldsymbol{z}_i. \tag{2.1}$$

Thus, $\boldsymbol{\alpha}_i$ and $\boldsymbol{z}_i$ play the role of confounders, but recall that $\boldsymbol{\alpha}_i$ is *not* directly observed in our setup, making (2.1) different from the standard selection on observables assumption.

We assume a semi-linear regression model to characterize the untreated outcome:

$$y_i(0) = \mu(\boldsymbol{\alpha}_i) + \boldsymbol{z}_i'\boldsymbol{\beta} + \epsilon_i, \qquad \mathbb{E}[\epsilon_i|\boldsymbol{\alpha}_i, \boldsymbol{z}_i] = 0, \tag{2.2}$$

where $\mu(\cdot)$ is some unknown function whose functional form is left unspecified. (2.2) implies that $\zeta_i := \mathbb{E}[y_i(0)|\boldsymbol{\alpha}_i, \boldsymbol{z}_i] = \mu(\boldsymbol{\alpha}_i) + \boldsymbol{z}_i'\boldsymbol{\beta}$. On the other hand, we assume a logistic semi-linear model for the treatment:

$$t_i = \psi(e(\boldsymbol{\alpha}_i) + \boldsymbol{z}_i'\boldsymbol{\gamma}) + v_i, \qquad \mathbb{E}[v_i|\boldsymbol{\alpha}_i, \boldsymbol{z}_i] = 0, \tag{2.3}$$

where $\psi(a) = 1/(1 + \exp(-a))$ is the standard logistic function and $e(\cdot)$ is an unknown function left unspecified. (2.3) implies that $\varrho_i := \mathbb{P}(t_i = 1|\boldsymbol{\alpha}_i, \boldsymbol{z}_i) = \psi(e(\boldsymbol{\alpha}_i) + \boldsymbol{z}_i'\boldsymbol{\gamma})$, which would be the usual propensity score if $\boldsymbol{\alpha}_i$ were observed. Note that $\boldsymbol{z}_i$ and $\boldsymbol{\alpha}_i$ enter the two equations simultaneously, which is consistent with the unconfoundedness assumption (2.1).

Equations (2.2) and (2.3) can be viewed as special cases of generalized semi-linear models with an identity link and a logistic link respectively. Such choices of link functions are common in practice when the outcome is continuous and the treatment is binary, which also match our empirical application. Extensions to other links are allowed in the general theory presented in the SA. Notice that despite specific links used, our model remains nonparametric with respect to the dependence of the outcome and treatment on latent $\boldsymbol{\alpha}_i$, since $\mu(\cdot)$ and $e(\cdot)$ are unspecified. This is appealing because $\boldsymbol{\alpha}_i$ is unobserved (and sometimes difficult to

precisely define); allowing it to enter the model nonparametrically offers greater flexibility and more robustness in confounding adjustment. On the other hand, incorporating observed $\boldsymbol{z}_i$ (e.g., various types of fixed effects) linearly is a usual strategy in practice to maintain tractability of the full model.

2.1 Structure of High-Dimensional Proxies

Since $\boldsymbol{\alpha}_i$ is not observed, regression techniques cannot be applied to estimate $\mathbb{E}[y_i(0)|\boldsymbol{\alpha}_i, \boldsymbol{z}_i]$ or $\mathbb{P}(t_i = 1|\boldsymbol{\alpha}_i, \boldsymbol{z}_i)$ directly, making many classical methods based on (2.1), such as regression adjustment and inverse probability weighting, infeasible. However, we have a vector of noisy proxies $\boldsymbol{x}_i = (x_{i1}, \cdots, x_{ip})'$ linked with the unobserved $\boldsymbol{\alpha}_i$ through a nonlinear factor model:

$$\boldsymbol{x}_i = \boldsymbol{\eta}(\boldsymbol{\alpha}_i) + \boldsymbol{u}_i, \quad \mathbb{E}[\boldsymbol{u}_i|\boldsymbol{\alpha}_i] = 0, \tag{2.4}$$

where $\boldsymbol{\eta}(\cdot) = (\eta_1(\cdot), \cdots, \eta_p(\cdot))' : \mathbb{R}^{r_\alpha} \mapsto \mathbb{R}^p$ is a vector of latent functions possibly heterogeneous across proxies, and $\boldsymbol{u}_i = (u_{i1}, \cdots, u_{ip})'$ is a vector of idiosyncratic errors. Define $n \times p$ 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, (2.4) can be written in matrix form $\boldsymbol{X} = \boldsymbol{H} + \boldsymbol{u}$. Compared to linear factor models, the latent mean matrix $\boldsymbol{H}$ could be full-rank due to the potential nonlinearity of $\boldsymbol{\eta}$.

Conceptually, the latent function $\boldsymbol{\eta}$ can be either fixed or random, depending on the specific context. For instance, when $\boldsymbol{x}_i$ is a vector of proxies for one's latent ability such as test scores, it may be appropriate to assume each η_l is a fixed "production function" that maps one's ability to a specific output; when p proxies are collected over time, $\{\eta_l(\cdot)\}_{l=1}^p$ may be viewed as a functional time series with each $\eta_l(\cdot)$ corresponding to a random function at time l (see, e.g., Ferraty and Vieu, 2006, for more details about functional data). In the literature on panel and network data (e.g., Bai, 2003; Gao et al., 2015), it is also common to assume a more specific setup: $\eta_l(\boldsymbol{\alpha}_i) = \boldsymbol{\varpi}_l'\boldsymbol{\alpha}_i$ ("interactive fixed-effect model") or $\eta_l(\boldsymbol{\alpha}_i) = \eta(\boldsymbol{\alpha}_i, \boldsymbol{\varpi}_l)$ ("graphon model"), for some fixed bivariate function $\eta(\cdot, \cdot)$ and latent variables $\boldsymbol{\varpi}_l$. The

heterogeneity (and potential randomness) of η_l across l then is driven by $\boldsymbol{\varpi}_l$. For ease of presentation, we treat $\boldsymbol{\eta}(\cdot)$ as fixed (conditioned upon) in this paper and keep agnostic about its specific structure and the potential dependence over l.

The high-dimensional $\boldsymbol{x}_i$ contains information on $\boldsymbol{\alpha}_i$ if $\boldsymbol{\eta}(\cdot)$ is nondegenerate in some proper sense formalized later. Conceptually, we may find some low-dimensional statistic $\mathcal{T}_i(\boldsymbol{X})$, as a function of observables, such that $\mathbb{E}[y_i(0)|\boldsymbol{\alpha}_i, \boldsymbol{z}_i]$ and $\mathbb{P}[t_i = 1|\boldsymbol{\alpha}_i, \boldsymbol{z}_i]$, the building blocks of many treatment effects estimators, are approximately the same as $\mathbb{E}[y_i(0)|\mathcal{T}_i(\boldsymbol{X}), \boldsymbol{z}_i]$ and $\mathbb{P}[t_i = 1|\mathcal{T}_i(\boldsymbol{X}), \boldsymbol{z}_i]$, when p is large. In other words, the challenge of selection on the unobserved $\boldsymbol{\alpha}_i$ in this case is resolved by selection on a proper transformation of observables, as will be formalized in our theory.

Remark 2.1 (Conditioning on $\boldsymbol{x}_i$)**.** When many noisy proxies are available, it is tempting to ask whether $\mathbb{E}[y_i(0)|\boldsymbol{x}_i]$ is (approximately) the same as $\mathbb{E}[y_i(0)|\boldsymbol{\alpha}_i]$, justifying the use of estimation methods that directly control for $\boldsymbol{x}_i$. Unfortunately, this alternative approach suffers from several issues and is usually invalid. To gain some intuition, consider a Gaussian error model:

$$y_i(0) = \alpha_i + \epsilon_i, \quad x_{il} = \alpha_i + u_{il}, \;\; l = 1, \cdots, p,$$

where $(\alpha_i, \epsilon_i, \boldsymbol{u}_i')' \in \mathbb{R}^{2+p}$ are multivariate normal, with zero means and all components uncorrelated with each other except that $\sigma_{p\epsilon} := \mathbb{C}\text{ov}(u_{ip}, \epsilon_i)$ is possibly nonzero. In this special case, it can be shown that

$$\mathbb{E}[y_i(0)|\boldsymbol{x}_i] = \frac{p + \sigma_{p\epsilon}}{p+1}\mathbb{E}[y_i(0)|\alpha_i] + \frac{(1 - \sigma_{p\epsilon})}{p+1}\sum_{l=1}^{p} u_{il} + \sigma_{p\epsilon}u_{ip}.$$

This calculation has several implications. First, the two estimands $\mathbb{E}[y_i(0)|\boldsymbol{x}_i]$ and $\mathbb{E}[y_i(0)|\alpha_i]$ are generally different for any finite p, and the last term $\sigma_{p\epsilon}u_{ip}$ does *not* vanish even when p is large. It implies that methods that directly control for $\boldsymbol{x}_i$ are invalid for estimating the conditional means of potential outcomes (or the CATE parameter $\mathbb{E}[y_i(1) - y_i(0)|\boldsymbol{\alpha}_i]$), even if there is only one single error u_{ip} is correlated with the outcome. Such correlation is

practically relevant. For instance, in panel data settings, noisy proxies may be past outcomes, and the most recent ones may be strongly correlated with the outcome in the current period of interest.

Second, when $\sigma_{p\epsilon} = 0$, though $\mathbb{E}[y_i(0)|\boldsymbol{x}_i] - \mathbb{E}[y_i(0)|\alpha_i]$ converges in probability to zero as p grows, the coefficients on conditioning variables $\boldsymbol{x}_i$ are small but *non-sparse*, which implies that many sparsity-based high-dimensional regression techniques such as Lasso are invalid for estimating quantities such as $\mathbb{E}[y_i(0)|\alpha_i]$. In other words, conditioning on a few selected coordinates of the high-dimensional $\boldsymbol{x}_i$ is unable to remove confounding due to α_i.

Third, due to the specific linear, homogeneous structure for x_{il}'s, α_i can be approximately recovered simply by averaging all proxies in this example. However, when $\eta_l(\cdot)$'s are unknown and heterogeneous, it is no longer straightforward what "proper" transformation is able to remove confounding. Our proposed method below can be viewed as a general strategy to effectively extract information on α_i in possibly nonlinear settings, while allowing for potential correlation between measurement errors and outcomes (or treatment assignment). See more discussions below Assumption 3 and in Section SA-1 of the SA. ⌟

2.2 Notation

For any scalars a and b, $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. Define $\delta_{Kp} = (K^{1/2} \wedge p^{1/2})/\sqrt{\log(n \vee p)}$. For a finite set $\mathcal{S}$, $|\mathcal{S}|$ denotes its cardinality. For an integer $m > 0$, $[m] = \{1, 2, \cdots, m\}$.

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 $\boldsymbol{A}$, $\|\boldsymbol{A}\|_{\max} = \max_{i\in[m],j\in[n]} |a_{ij}|$ is the entrywise sup-norm of $\boldsymbol{A}$. Moreover, $\boldsymbol{A}_{i\cdot}$ denotes the i-th row of $\boldsymbol{A}$.

For sequences of numbers or random variables, $a_n \lesssim b_n$ or $a_n = O(b_n)$ denotes $\limsup_n |a_n/b_n|$ is finite, and $a_n \lesssim_{\mathbb{P}} b_n$ or $a_n = O_{\mathbb{P}}(b_n)$ 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 $a_n/b_n \to_{\mathbb{P}} 0$, 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 strictly positive sequence $\{r_{i,n}\}_{i\in[n]}$, we write $a_{i,n} = \bar{o}_{\mathbb{P}}(r_{i,n})$ if $\max_{i\in[n]} |a_{i,n}/r_{i,n}| \to_{\mathbb{P}} 0$.

3 Outline of Estimation Procedure

This section describes our proposed estimation procedure. First, given the structure (2.4), relevant information on $\boldsymbol{\alpha}_i$ is extracted from $\boldsymbol{x}_i$ via a local PCA method. Second, we estimate $\zeta_i = \mathbb{E}[y_i(0)|\boldsymbol{\alpha}_i, \boldsymbol{z}_i]$ and $\varrho_i = \mathbb{P}(t_i = 1|\boldsymbol{\alpha}_i, \boldsymbol{z}_i)$ via a local quasi-maximum likelihood method, where the extracted information from the first step plays the role of kernel functions and generated regressors. Finally, the estimator of $\mathbb{E}[y_i(0)|t_i = 1]$ are constructed based on the doubly-robust score function. See Algorithm 1 for a short summary.

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. 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 below:

- Potential outcomes $y_i(1)$ and $y_i(0)$: the cumulative stock returns of firm i from day 0 to day 1 that would be observed with and without Geithner connections;
- Noisy proxies $\boldsymbol{x}_i$: 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;
- ATT parameter $\mathbb{E}[y_i(1) - y_i(0)|t_i = 1]$: the average cumulative abnormal returns of connected firms from day 0 to day 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 proxies $\boldsymbol{x}_i$ only include the daily stock returns for 250 days that ends 30 days prior to the Geithner announcement ($p = 250$). In this application, one concern is that even after controlling for the observed covariates $\boldsymbol{z}_i$, "connected and non-connected firms could be different in other ways" (Acemoglu et al., 2016, pp.19). For instance, two groups of firms may be different in some unobserved characteristics $\boldsymbol{\alpha}_i$ such as entrepreneurial ability or management quality, which could be implicitly captured by stock returns as discussed before.

3.1 Step 1: Latent Variables Extraction

The goal of this step is to extract information on latent confounders $\boldsymbol{\alpha}_i$ by employing the structure (2.4). The key ideas are outlined below.

Proxy-wise Data Splitting. We recommend users separate the K-NN matching and PCA steps via proxy-wise data splitting. Specifically, split the proxy ("column") index set $\mathcal{C} = [p]$ of $\boldsymbol{X}$ into two non-overlapping subsets: $\mathcal{C} = \mathcal{C}^\dagger \cup \mathcal{C}^\ddagger$ with $p^\dagger = |\mathcal{C}^\dagger|$, $p^\ddagger = |\mathcal{C}^\ddagger|$ 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 column indices in $\mathcal{C}^\dagger$ and $\mathcal{C}^\ddagger$ respectively. $\boldsymbol{H}^\dagger$, $\boldsymbol{H}^\ddagger$, $\boldsymbol{U}^\dagger$ and $\boldsymbol{U}^\ddagger$ are defined similarly.

Note that the splitting is conducted across proxies. It is used to obtain desired convergence properties of the local PCA estimator, which differs from sample splitting across units that separates nuisance function estimation and final inference in the double machine learning literature.

Remark 3.1 (Implementation: data splitting)**.** 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). For independent errors across l as specified in Assumption 3 below, this goal can be achieved by, e.g., randomly splitting the column index set. When proxies are collected over time, we can, e.g., take the first half of time periods for K-NN and the second half for PCA, so the time series structure is retained. ⌟

Algorithm 1 (Estimation and inference for $\mathbb{E}[y_i(0)|t_i = 1]$)

Step 1: Latent Variables Extraction

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

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

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

For each $i \in [n]$,

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

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

(2) use $\boldsymbol{X}^\ddagger_{\mathcal{N}_i}$ to obtain the local factor loading $\widehat{\boldsymbol{\Lambda}}_{\mathcal{N}_i}$ by local PCA:

$$(\widehat{\boldsymbol{F}}_{\mathcal{N}_i}, \widehat{\boldsymbol{\Lambda}}_{\mathcal{N}_i}) = \underset{\tilde{\boldsymbol{F}}_{\mathcal{N}_i}\in\mathbb{R}^{p^\ddagger\times \mathsf{d}_i}, \tilde{\boldsymbol{\Lambda}}_{\mathcal{N}_i}\in\mathbb{R}^{K\times \mathsf{d}_i}}{\arg\min} \operatorname{Tr}\Big[\Big(\boldsymbol{X}^\ddagger_{\mathcal{N}_i} - \tilde{\boldsymbol{\Lambda}}_{\mathcal{N}_i}\tilde{\boldsymbol{F}}'_{\mathcal{N}_i}\Big)\Big(\boldsymbol{X}^\ddagger_{\mathcal{N}_i} - \tilde{\boldsymbol{\Lambda}}_{\mathcal{N}_i}\tilde{\boldsymbol{F}}'_{\mathcal{N}_i}\Big)'\Big]$$

Step 2: Local Quasi-Maximum Likelihood Estimation

Input: $\{y_i, t_i, \boldsymbol{z}_i, \widehat{\boldsymbol{\Lambda}}_{\mathcal{N}_i}, \mathcal{N}_i\}_{i=1}^n$

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

For each $i \in [n]$,

(1) estimate $\zeta_i = \mathbb{E}[y_i(0)|\boldsymbol{\alpha}_i, \boldsymbol{z}_i]$ by local least squares as in (3.4), using data for units in $\mathcal{N}_i$ with $t_i = 0$

(2) estimate $\varrho_i = \mathbb{P}(t_i = 1|\boldsymbol{\alpha}_i, \boldsymbol{z}_i)$ by local logistic regression as in (3.5) using data for units in $\mathcal{N}_i$

Step 3: Counterfactual Analysis

Input: $\{y_i, t_i, \widehat{\zeta}_i, \widehat{\varrho}_i\}_{i=1}^n$

Output: estimate $\widehat{\theta}$ of $\theta = \mathbb{E}[y_i(0)|t_i = 1]$ and its standard error $\widehat{\sigma}$

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^n \left[\frac{t_i\widehat{\zeta}_i}{\widehat{\varrho}} + \frac{\widehat{\varrho}_i}{\widehat{\varrho}}\frac{(1-t_i)(y_i - \widehat{\zeta}_i)}{1-\widehat{\varrho}_i}\right], \quad \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \left[\frac{t_i(\widehat{\zeta}_i - \widehat{\theta})^2}{\widehat{\varrho}^2} + \frac{\widehat{\varrho}_i^2(1-t_i)(y_i - \widehat{\zeta}_i)^2}{\widehat{\varrho}^2(1-\widehat{\varrho}_i)^2}\right]$$

***K*-NN Matching**. This step makes use of the subsample labeled by †, i.e., the submatrix of $\boldsymbol{X}$ with column indices in $\mathcal{C}^\dagger$. 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:

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

We use the term "distance" in a loose sense here, so ρ does not have to satisfy all the axioms for a distance function in math.

Remark 3.2 (Implementation: choice of distance)**.** Some usual choices of distance in the literature include (i) Euclidean distance $\rho(\boldsymbol{X}_{i\cdot}^\dagger, \boldsymbol{X}_{j\cdot}^\dagger) = \|\boldsymbol{X}_{i\cdot}^\dagger - \boldsymbol{X}_{j\cdot}^\dagger\|^2/p^\dagger$; (ii) pseudo-max distance $\rho(\boldsymbol{X}_{i\cdot}^\dagger, \boldsymbol{X}_{j\cdot}^\dagger) = \max_{\ell\neq i,j} |(\boldsymbol{X}_{i\cdot}^\dagger - \boldsymbol{X}_{j\cdot}^\dagger)'\boldsymbol{X}_{\ell\cdot}^\dagger|/p^\dagger$; and (iii) distance of averages $\rho(\boldsymbol{X}_{i\cdot}^\dagger, \boldsymbol{X}_{j\cdot}^\dagger) = |\sum_{l\in\mathcal{C}^\dagger}(x_{il} - x_{jl})|/p^\dagger$. Indirect matching based on each of these choices can be justified under certain conditions on the data generating process, and generally, no one strictly dominates in theory. However, the Euclidean distance-based matching requires an unrealistic conditionally homoskedasticity assumption, and matching using the distance of averages needs a strong injectivity condition on the averaged proxies. By contrast, the pseudo-max distance can be justified under practically more plausible conditions and thus is our more recommended choice among the three. See Remarks 4.1 and 4.2 and Section SA-2 of the SA for details.

Moreover, when proxies differ in scale or importance for revealing information on the latent variables, one can rescale or reweight different proxies when searching for nearest neighbors. Such transformations can be viewed as particular choices of the distance $\rho(\cdot,\cdot)$. For instance, given a vector of weights w_l, $l \in \mathcal{C}^\dagger$, one can define $\rho(\boldsymbol{X}_{i\cdot}^\dagger, \boldsymbol{X}_{j\cdot}^\dagger) = \sum_{l\in\mathcal{C}^\dagger}(w_l x_{il} - w_l x_{jl})^2$, which is the squared Euclidean distance of the reweighted proxies. ⌟

Remark 3.3 (Implementation: choice of K)**.** Our simulation studies in Section 4.3 show strong robustness of our method with respect to the choice of K. Theoretically, by taking advantage of doubly-robust score functions, our main results rely on mild restrictions on K, which reduces to $n/K^2 = o(1)$ (up to $\log n$ terms) in practically relevant cases (see discussions below Theorem 4.2). Thus, many feasible choices of K is valid in theory. 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 proxies. 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 $\sqrt{n}$ if $m > r/2$ or $m > r/2-1$. 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. ⌟

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. Given the discussion in Remark 3.3, we set $K = 99$ as a benchmark choice, which is obtained based on leave-one-out cross validation for local constant nearest-neighbor regression of cumulative stock returns from day 0 to 1 on the pre-treatment return on day -30. To have a sense of the performance of K-NN matching, we report some summary statistics of matching discrepancies in Table 1. Specifically, 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. Matching performs relatively well for treated units, whereas some control units are matched with someone relatively far away. In later analysis, we check the robustness of our 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: K-NN matching with $K = 99$ is conducted. For each unit, we normalize the maximum matched-pair distance by dividing it by the standard deviation of distances across all pairs.

Local PCA. This step makes use of the subsample labeled by ‡, i.e., the submatrix of $\boldsymbol{X}$ with column indices in $\mathcal{C}^{\ddagger}$. Given the nearest neighbors of i found in the first step, define $K \times p^{\ddagger}$ submatrix $\boldsymbol{X}^{\ddagger}_{\mathcal{N}_i}$ of $\boldsymbol{X}^{\ddagger}$, with each row corresponding to a neighbor's proxies that

belong to $\mathcal{C}^\ddagger$. Similarly, $\boldsymbol{H}^\ddagger_{\mathcal{N}_i}$ and $\boldsymbol{U}^\ddagger_{\mathcal{N}_i}$ denote submatrices of $\boldsymbol{H}^\ddagger$ and $\boldsymbol{U}^\ddagger$ respectively, formed by rows corresponding to K nearest neighbors of unit i. Then, we write

$$\boldsymbol{X}^\ddagger_{\mathcal{N}_i} = \boldsymbol{\Lambda}_{\mathcal{N}_i}\boldsymbol{F}'_{\mathcal{N}_i} + \boldsymbol{\Xi}_{\mathcal{N}_i} + \boldsymbol{U}^\ddagger_{\mathcal{N}_i}, \tag{3.2}$$

where the local factor matrix $\boldsymbol{F}_{\mathcal{N}_i} \in \mathbb{R}^{p^\ddagger \times \mathsf{d}_i}$ and the loading matrix $\boldsymbol{\Lambda}_{\mathcal{N}_i} \in \mathbb{R}^{K \times \mathsf{d}_i}$ are defined to be the solution to the noise-free local PCA problem:

$$(\boldsymbol{F}_{\mathcal{N}_i}, \boldsymbol{\Lambda}_{\mathcal{N}_i}) := \underset{\tilde{\boldsymbol{F}}_{\mathcal{N}_i} \in \mathbb{R}^{p^\ddagger \times \mathsf{d}_i},\, \tilde{\boldsymbol{\Lambda}}_{\mathcal{N}_i} \in \mathbb{R}^{K \times \mathsf{d}_i}}{\arg\min} \operatorname{Tr}\left[\left(\boldsymbol{H}^\ddagger_{\mathcal{N}_i} - \tilde{\boldsymbol{\Lambda}}_{\mathcal{N}_i}\tilde{\boldsymbol{F}}'_{\mathcal{N}_i}\right)\left(\boldsymbol{H}^\ddagger_{\mathcal{N}_i} - \tilde{\boldsymbol{\Lambda}}_{\mathcal{N}_i}\tilde{\boldsymbol{F}}'_{\mathcal{N}_i})'\right]\right.$$

such that $\tilde{\boldsymbol{F}}'_{\mathcal{N}_i}\tilde{\boldsymbol{F}}_{\mathcal{N}_i}/p^\ddagger = \boldsymbol{I}_{\mathsf{d}_i}$ and $\tilde{\boldsymbol{\Lambda}}'_{\mathcal{N}_i}\tilde{\boldsymbol{\Lambda}}_{\mathcal{N}_i}/K$ is diagonal. d_i denotes the number of local factors extracted, and $\boldsymbol{\Xi}_{\mathcal{N}_i} \in \mathbb{R}^{K \times p^\ddagger}$ is the corresponding approximation error for $\boldsymbol{H}^\ddagger_{\mathcal{N}_i}$. In practice, we estimate $\boldsymbol{F}_{\mathcal{N}_i}$ and $\boldsymbol{\Lambda}_{\mathcal{N}_i}$ by PCA:

$$(\widehat{\boldsymbol{F}}_{\mathcal{N}_i}, \widehat{\boldsymbol{\Lambda}}_{\mathcal{N}_i}) = \underset{\tilde{\boldsymbol{F}}_{\mathcal{N}_i} \in \mathbb{R}^{p^\ddagger \times \mathsf{d}_i},\, \tilde{\boldsymbol{\Lambda}}_{\mathcal{N}_i} \in \mathbb{R}^{K \times \mathsf{d}_i}}{\arg\min} \operatorname{Tr}\left[\left(\boldsymbol{X}^\ddagger_{\mathcal{N}_i} - \tilde{\boldsymbol{\Lambda}}_{\mathcal{N}_i}\tilde{\boldsymbol{F}}'_{\mathcal{N}_i}\right)\left(\boldsymbol{X}^\ddagger_{\mathcal{N}_i} - \tilde{\boldsymbol{\Lambda}}_{\mathcal{N}_i}\tilde{\boldsymbol{F}}'_{\mathcal{N}_i})'\right]\right. \tag{3.3}$$

such that $\tilde{\boldsymbol{F}}'_{\mathcal{N}_i}\tilde{\boldsymbol{F}}_{\mathcal{N}_i}/p^\ddagger = \boldsymbol{I}_{\mathsf{d}_i}$ and $\tilde{\boldsymbol{\Lambda}}'_{\mathcal{N}_i}\tilde{\boldsymbol{\Lambda}}_{\mathcal{N}_i}/K$ is diagonal.

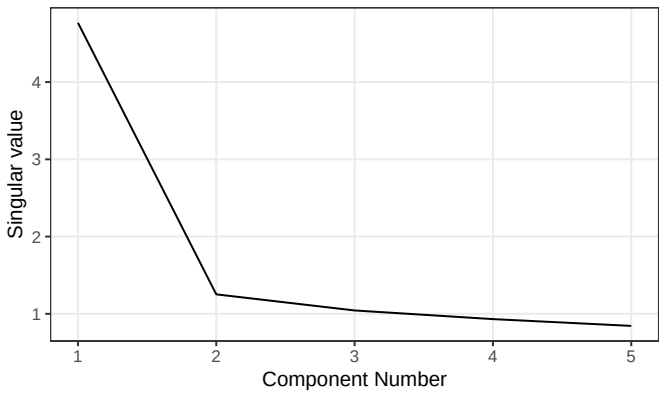
Remark 3.4 (Implementation: choice of d_i)**.** In this paper we treat d_i as a fixed tuning parameter that is possibly heterogeneous across neighborhoods, which governs the number of terms in the local approximation. In principle, d_i is set such that the extracted local factors have stronger "signal strength" relative to the noise $\boldsymbol{U}^\ddagger_{\mathcal{N}_i}$ (see discussions below Assumption 5). Note that although the number of latent confounders r_α implicitly affects the approximation quality of the proposed method, it is *unnecessary* in practice to determine r_α in the above procedure. Also, d_i need *not* be greater than or equal to r_α (e.g., $\mathsf{d}_i = 1$ is conceptually analogous to local constant smoothing in nonparametrics).

In practice, we recommend users select d_i in a data-dependent fashion. For instance, in each local neighborhood we extract all local factors from $\boldsymbol{X}^\ddagger_{\mathcal{N}_i}$ with the associated singular values greater than $C(\sqrt{K} + \sqrt{p^\ddagger})$ for some large constant $C > 0$. Since the order of noise

matrix $\boldsymbol{U}_{\mathcal{N}_i}^{\ddagger}$ is typically $O_{\mathbb{P}}(\sqrt{K}+\sqrt{p^{\ddagger}})$, this strategy ensures we can almost attain the "best" approximation allowed by the data. Refined strategies based on similar ideas are available in the literature on PCA (e.g., Johnstone, 2001; Dobriban and Owen, 2019). ⌟

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 then is conducted for this subgroup of firms using their daily stock returns in the next 125 days. Figure 2 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 empirical analysis, we employ the strategy described in Remark 3.4 to choose d_i in a data-dependent fashion.

Figure 2: Local Eigenvalues for One Neighborhood



3.2 Step 2: Local Quasi-Maximum Likelihood Estimation

The goal of this step is to estimate $\zeta_i = \mathbb{E}[y_i(0)|\boldsymbol{\alpha}_i, \boldsymbol{z}_i]$ and $\varrho_i = \mathbb{P}(t_i = 1|\boldsymbol{\alpha}_i, \boldsymbol{z}_i)$. Section SA-3.1 of the SA considers the more general quasi-maximum likelihood estimation (QMLE), but to be consistent with our basic setup in (2.2) and (2.3), here we focus on Gaussian likelihood for the outcome equation and Logit Likelihood for the treatment equation.

A local least squares estimator of ζ_i is given by

$$\begin{aligned} \widehat{\zeta}_i &= \widehat{\mu}(\boldsymbol{\alpha}_i) + \boldsymbol{z}_i'\widehat{\boldsymbol{\beta}}_{\mathcal{N}_i}, \quad \widehat{\mu}(\boldsymbol{\alpha}_i) = \widehat{\boldsymbol{\lambda}}_{i,\mathcal{N}_i}'\widehat{\boldsymbol{b}}_{\mathcal{N}_i}, \quad \text{where} \\ (\widehat{\boldsymbol{b}}_{\mathcal{N}_i}',\ \widehat{\boldsymbol{\beta}}_{\mathcal{N}_i}')' &= \underset{(\boldsymbol{b}_1',\,\boldsymbol{b}_2')'\in\mathbb{R}^{\mathtt{d}_i+r_z}}{\arg\min} \sum_{j\in\mathcal{N}_i}(1-t_i)(y_j - \widehat{\boldsymbol{\lambda}}_{j,\mathcal{N}_i}'\boldsymbol{b}_1 - \boldsymbol{z}_j'\boldsymbol{b}_2)^2, \end{aligned} \tag{3.4}$$

where $\widehat{\boldsymbol{\lambda}}_{j,\,\mathcal{N}_i}$ is the estimated factor loading vector in $\widehat{\boldsymbol{\Lambda}}_{\mathcal{N}_i}$ corresponding to unit j. For the treatment equation (2.3), the estimator of $\varrho_i = \mathbb{P}(t_i = 1|\boldsymbol{\alpha}_i, \boldsymbol{z}_i)$ is given by

$$\begin{aligned} \widehat{\varrho}_i &= \psi(\widehat{e}(\boldsymbol{\alpha}_i) + \boldsymbol{z}_i'\widehat{\boldsymbol{\gamma}}_{\mathcal{N}_i}), \quad \widehat{e}(\boldsymbol{\alpha}_i) = \widehat{\boldsymbol{\lambda}}_{i,\mathcal{N}_i}'\widetilde{\boldsymbol{b}}_{\mathcal{N}_i}, \quad \text{where} \\ (\widetilde{\boldsymbol{b}}_{\mathcal{N}_i}',\ \widehat{\boldsymbol{\gamma}}_{\mathcal{N}_i}')' &= \underset{(\boldsymbol{b}_1',\,\boldsymbol{b}_2')'\in\mathbb{R}^{\mathtt{d}_i+r_z}}{\arg\max} \sum_{j\in\mathcal{N}_i} \log[\psi_j(\boldsymbol{b}_1,\boldsymbol{b}_2)^{t_j}(1-\psi_j(\boldsymbol{b}_1,\boldsymbol{b}_2))^{1-t_j}], \end{aligned} \tag{3.5}$$

where $\psi_j(\boldsymbol{b}_1,\boldsymbol{b}_2) := \psi(\widehat{\boldsymbol{\lambda}}_{j,\mathcal{N}_i}'\boldsymbol{b}_1 + \boldsymbol{z}_j'\boldsymbol{b}_2)$.

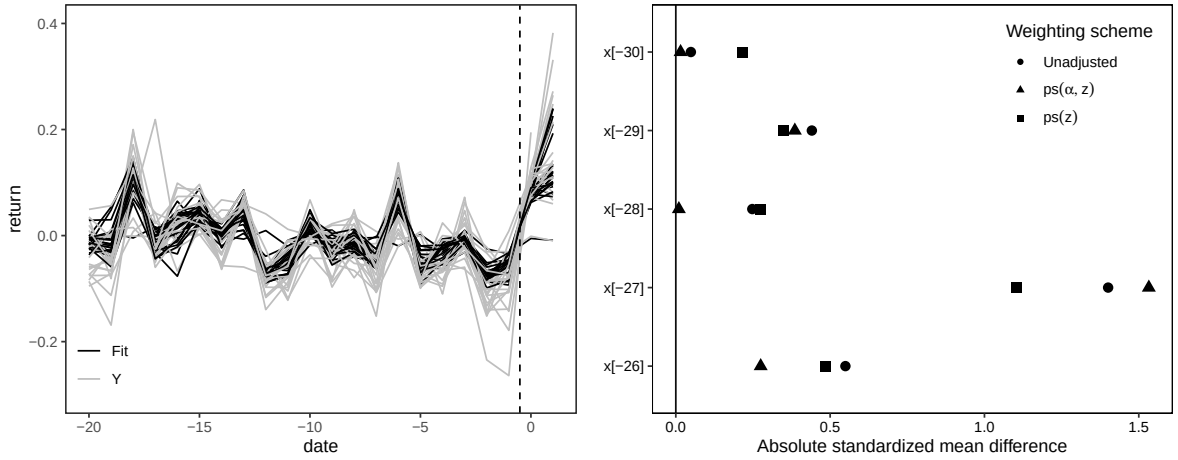
Remark 3.5 (Implementation: MLE)**.** For local least squares we have close-form solutions, while for other general quasi-MLE methods, the maximization is done numerically in practice. Our outputs from local PCA are readily used as inputs to any existing fast MLE algorithms, and also note that only fitted values $\widehat{\zeta}_i$ and $\widehat{\varrho}_i$ are necessary for causal inference.

In (3.4) and (3.5), we simply add $\boldsymbol{z}_i$ as additional controls in each local estimation problem, which does not fully exploit the model structure to estimate the low-dimensional parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. Alternatively, one can use other methods in the semiparametrics literature, e.g., profiled quasi-maximum likelihood, to estimate these parametric components, though it is computationally more burdensome. See Härdle et al. (2004) for such methods. ⌟

An an illustration, for each treated firm, we find its nearest neighbors and obtain the local factor loadings. Then, for returns in each day from -20 through $+1$, we use the data of the treated firm's untreated neighbors to perform a local least squares regression on the extracted factor loadings. Figure 3a shows the fitted values in black and the observed daily returns in grey for the treated firms. Recall that the fitted values are the estimates of the conditional means of potential returns *without* treatment given the latent variables. We find

that compared to the fitted values, closer to event day 0, some treated firms' observed returns decrease, while after day 0, most sequences of returns increase sharply. Given the propensity score estimates obtained in this step, we can also check the balance in pre-treatment returns that are not used for local PCA. The result is reported in Figure 3b, which compares our approach with propensity score estimates based on $\boldsymbol{z}_i$ only. We remind readers that balance diagnostics in this context should be interpreted cautiously, since the pre-treatment outcomes are noisy proxies rather than the true confounders, and measurement error can distort proxy-based balance measures.

Figure 3: Local MLE



(a) Conditional means of untreated outcome (b) Balance in pre-treatment returns

Notes: The left panel shows the fitted values for the conditional mean of untreated returns for treated firms, along with the observed return sequences. The right panel shows the balance in returns from day −30 through day −26, where $\mathsf{ps}(\alpha,\mathsf{z})$ and $\mathsf{ps}(\mathsf{z})$ denote propensity scores estimates from our approach and from a specification using $\boldsymbol{z}_i$ only.

3.3 Step 3: Counterfactual Analysis

The final step is to estimate $\theta = \mathbb{E}[y_i(0)|t_i = 1]$. Let $\varrho = \mathbb{P}(t_i = 1)$. Under our unconfoundedness assumption (2.1),

$$\theta = \mathbb{E}\left[\frac{t_i\zeta_i}{\varrho} + \frac{\varrho_i}{\varrho}\frac{(1-t_i)(y_i-\zeta_i)}{1-\varrho_i}\right].$$

Therefore, an estimator of θ is given by

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{t_i\widehat{\zeta}_i}{\widehat{\varrho}} + \frac{\widehat{\varrho}_i}{\widehat{\varrho}}\frac{(1-t_i)(y_i-\widehat{\zeta}_i)}{1-\widehat{\varrho}_i}\right], \tag{3.6}$$

where $\widehat{\zeta}_i$ and $\widehat{\varrho}_i$ are obtained in the second step, and $\widehat{\varrho} = \sum_{i=1}^{n} t_i/n$. For the purpose of inference, a simple plug-in variance estimator for $\widehat{\theta}$ is

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{t_i(\widehat{\zeta}_i-\widehat{\theta})^2}{\widehat{\varrho}^2}\right] + \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\widehat{\varrho}_i^2(1-t_i)(y_i-\widehat{\zeta}_i)^2}{\widehat{\varrho}^2(1-\widehat{\varrho}_i)^2}\right]. \tag{3.7}$$

Under some regularity conditions, we show in Theorem 4.2 below that $\sqrt{n}\widehat{\sigma}^{-1}(\widehat{\theta}-\theta) \rightsquigarrow \mathsf{N}(0,1)$. Estimators of other causal parameters can be constructed accordingly. For instance, an estimator of ATT is simply $(\sum_{i=1}^{n} t_i)^{-1}\sum_{i=1}^{n} t_i y_i - \widehat{\theta}$.

Remark 3.6 (Implementation: small denominators)**.** It is well known in the literature that inverse probability weighting estimators may suffer from the "small denominators" issue (propensity scores too close to 0 or 1). In such cases some principled trimming methods may be needed to ensure more robust performance of the estimator. See, e.g., Crump et al. (2009); Busso et al. (2014); Ma and Wang (2020) for discussions and possible solutions. ⌟

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. 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; for treatment equation, we perform a local logit estimation. We consider several choices of $K = C\widehat{K}_{\mathsf{CV}}$ with $C = 0.5, 1, 2$, where $\widehat{K}_{\mathsf{CV}}$ is the cross-validation benchmark choice described before. We apply our procedure to the full sample and a base sample as defined in Acemoglu et al. (2016) (the base sample excludes firms whose returns are highly correlated with Citigroup).

Results without covariates are reported in the first two columns of Table 2, while the third and fourth columns give results that control for the three observed firm characteristics. For

comparison, we also include the estimates based on double lasso. The estimated average cumulative abnormal return for the connected firms using the proposed method ranges from 0.049 to 0.107, and in most cases the effects are significant at the 0.05 level. With covariates included, double lasso produces larger point estimates than ours and larger standard errors. By contrast, Acemoglu et al. (2016) uses a synthetic control method without covariate adjustment and finds an insignificant estimate of 0.005 for the full sample and a significant effect of 0.06 for the base sample. 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	0.096	0.077	0.060	0.057
	(0.028)	(0.029)	(0.024)	(0.03)
99	0.105	0.087	0.049	0.074
	(0.028)	(0.026)	(0.027)	(0.032)
198	0.107	0.093	0.064	0.080
	(0.028)	(0.026)	(0.029)	(0.03)
Double lasso				
	0.099	0.104	0.092	0.109
	(0.034)	(0.04)	(0.033)	(0.042)

Notes: The numbers in parentheses are standard errors. The base sample excludes firms highly correlated with Citigroup.

4 Main Results

4.1 Assumptions

We begin with the unconfoundedness and overlap conditions commonly used in the causal inference literature. Recall that the confounders $\boldsymbol{\alpha}_i$ are not directly observed.

Assumption 1 (Unconfoundedness and Overlap)**.** $(y_i, t_i, \boldsymbol{\alpha}_i, \boldsymbol{z}_i)$ *is i.i.d. over* $i \in [n]$ *and satisfies that (a)* $(y_i(0), y_i(1)) \perp\!\!\!\perp t_i \,|\, \boldsymbol{\alpha}_i, \boldsymbol{z}_i$*; and (b)* $0 < c_{\min} \leq \mathbb{P}(t_i = 1 | \boldsymbol{\alpha}_i, \boldsymbol{z}_i) \leq c_{\max} < 1$ *for some constants* $c_{\min}$ *and* $c_{\max}$ *almost surely.*

The conditional independence between the outcome and the treatment is a key feature of our problem and will be always maintained. The data are assumed to be i.i.d. over i for simplicity; in the empirical application it means that we have a random sample of firms whose stock returns are independent of each other. With additional technicalities it is possible to extend the results to allow for dependence within small clusters (so firms within the same industry can be correlated).

We impose some mild regularities conditions on covariates, errors and latent functions.

Assumption 2 (Regularities)**.** $\boldsymbol{\alpha}_i$ *has a compact convex support with a density bounded and bounded away from zero;* $\boldsymbol{z}_i$ *has a compact support, and* $\mathbb{E}[\tilde{\boldsymbol{z}}_i \tilde{\boldsymbol{z}}_i' | \boldsymbol{\alpha}_i]$ *has all eigenvalues bounded away from zero a.s., with* $\tilde{\boldsymbol{z}}_i = \boldsymbol{z}_i - \mathbb{E}[\boldsymbol{z}_i | \boldsymbol{\alpha}_i]$*;* $\max_{i \in [n], l \in [p]} \mathbb{E}[|u_{il}|^{2+\nu} | \boldsymbol{\alpha}_i] < \infty$*, and* $\max_{i \in [n]} \mathbb{E}[|\epsilon_i|^{2+\nu} | \boldsymbol{\alpha}_i, \boldsymbol{z}_i] < \infty$*, for some* $\nu > 2$*; for all* $l \in [p]$*,* $\eta_l(\cdot)$ *is bounded by a universal constant.*

Next, we impose an assumption restricting the errors in our model.

Assumption 3 (Independence of Errors)**.** u_{il} *is independent across* $i \in [n]$ *and over* $l \in [p]$*, conditional on* $\{\boldsymbol{\alpha}_i\}_{i \in [n]}$*;* (ϵ_i, v_i) *is independent of* $\boldsymbol{u}_i$*, conditional on* $\{(\boldsymbol{\alpha}_i, \boldsymbol{z}_i)\}_{i \in [n]}$*.*

As discussed before, Assumption 3 is common in the literature, implying that proxies are conditionally independent and do not directly affect the outcome or treatment. We assume it here for simplicity only. In this high-dimensional setting, the independence of u_{il} across l can be relaxed to, e.g., martingale difference or β-mixing conditions to accommodate proxies collected over time, albeit with more technical complexity. Note that proxies are still allowed to be correlated given this assumption, since the dependence structure of η_l's is left unspecified. On the other hand, the conditional independence between the measurement error and the outcome (or treatment) can also be relaxed. In the SA, we introduce a quantity

$\mathfrak{m}_p$ to explicitly characterize the bias due to the correlation of $\boldsymbol{u}_i$ with ϵ_i or v_i, reflected through the correlation between (ϵ_i, v_i) and our transformations of proxies, $\mathcal{N}_i$ and $\widehat{\boldsymbol{\Lambda}}_i$. Our inference result is valid as long as $\sqrt{n}\mathfrak{m}_p^2 = o(1)$. Typically, $\mathfrak{m}_p \to 0$ as p grows. Since it is technically more involved, we defer detailed discussion to Section SA-2.1 in the SA.

In our context, 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, we only need the nearest neighbors and local factor loadings described in Section 3 provide enough approximation power for the unknown function $\mu(\cdot)$ and $e(\cdot)$ when estimating $\mathbb{E}[y_i(0)|\boldsymbol{\alpha}_i, \boldsymbol{z}_i]$ and $\mathbb{P}(t_i = 1|\boldsymbol{\alpha}_i, \boldsymbol{z}_i)$. This is formalized in the next two assumptions.

Assumption 4 (Indirect Matching)**.** *For some fixed constants $\rho_0 \in \mathbb{R}$, $\underline{\varsigma} > 0$, $\bar{\varsigma} > 0$ and some positive sequence $a_p = o(1)$, the following conditions hold:*

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

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

Assumption 4 is a high-level condition, accommodating a generic choice of the distance ρ. In the following we make several remarks on each part, and defer more details, including sufficient conditions for Assumption 4, to Section SA-2.2 of the SA.

Part (a) implies that the distance of the noisy proxies can be translated into that of the noise-free mean structure, which is usually a mild requirement if the distance can "average" many proxies that are weakly dependent (conditional on $\boldsymbol{\alpha}_i$'s).

Remark 4.1 (Noise reduction)**.** For the Euclidean distance, since $\mathbb{E}[\boldsymbol{u}_i|\boldsymbol{\alpha}_i] = 0$ and u_{il} is independent over l, we roughly have $\|\boldsymbol{X}_{i\cdot}^\dagger - \boldsymbol{X}_{j\cdot}^\dagger\|^2/p^\dagger \approx \|\boldsymbol{H}_{i\cdot}^\dagger - \boldsymbol{H}_{j\cdot}^\dagger\|^2/p^\dagger + \|\boldsymbol{U}_{i\cdot}^\dagger\|^2/p^\dagger + \|\boldsymbol{U}_{j\cdot}^\dagger\|^2/p^\dagger$. Then, part (a) holds if $\|\boldsymbol{U}_{i\cdot}^\dagger\|^2$ are approximately the same across all units, which amounts to a (usually unrealistic) homoskedasticity condition—$\mathbb{E}[u_{il}^2|\boldsymbol{\alpha}_i]$ is a constant σ^2. Then, we can take $\rho_0 = 2\sigma^2$ and $a_p = (\frac{\log n}{p})^{1/2}$.

By contrast, for the pseudo-max distance, since the distance of any two units are measured using a third unit (and units are independent of each other), we have $(\boldsymbol{X}_{i\cdot}^\dagger - \boldsymbol{X}_{j\cdot}^\dagger)'\boldsymbol{X}_{\ell\cdot}^\dagger/p^\dagger \approx$

$(\boldsymbol{H}_{i\cdot}^{\dagger} - \boldsymbol{H}_{j\cdot}^{\dagger})'\boldsymbol{H}_{\ell\cdot}^{\dagger}/p^{\dagger}$. Homoskedasticity of errors is unnecessary, and we can take $\rho_0 = 0$ and $a_p = (\frac{\log n}{p})^{1/2}$.

For the distance of averages, $\sum_{l\in\mathcal{C}^{\dagger}}[x_{il} - x_{jl}]/p^{\dagger} \approx \sum_{l\in\mathcal{C}^{\dagger}}[\eta_l(\boldsymbol{\alpha}_i) - \eta_l(\boldsymbol{\alpha}_j)]/p^{\dagger}$ since $\mathbb{E}[\boldsymbol{u}_i|\boldsymbol{\alpha}_i] = 0$. Thus, we can take $\rho_0 = 0$ and $a_p = (\frac{\log n}{p})^{1/2}$. ⌟

On the other hand, part (b) precisely links the distance of the noise-free mean structures with that of the latent confounders. The upper bound requirement is mild. For example, if the latent functions η_l's are Lipschitz continuous, we can take $\bar{\varsigma} = 2$ for (squared) Euclidean distance and $\bar{\varsigma} = 1$ for pseudo-max distance and distance of averages. By contrast, the plausibility of the lower bound, the key requirement for informativeness of proxies, needs to be understood in context. See the following remark.

Remark 4.2 (Informativeness)**.** When the distance of averages is used, we typically need (the limit of) $\frac{1}{p^{\dagger}}\sum_{l\in\mathcal{C}^{\dagger}}\eta_l(\cdot)$ to be strictly monotonic. Then, we can take $\underline{\varsigma} = 1$. This is possible only for a scalar α_i. When there is more than one confounder, one may need to assume that each component of $\boldsymbol{\alpha}_i$ is separately measured by a set of proxies. The averaged information across these sets then can be combined to define the final nearest-neighbor set.

In addition, simple averaging may also lose useful information about confounders. For instance, consider a linear factor model: $\eta_l(\alpha_i) = \varpi_l\alpha_i$ for some $\varpi_l \in \mathbb{R}$. Strict monotonicity of the averaged η_l's is violated if $\frac{1}{p^{\dagger}}\sum_{l\in\mathcal{C}^{\dagger}}\varpi_l \approx 0$. That is, the averaged proxies can be approximately uninformative about the latent, even if a non-negligible subset of proxies have nonzero ϖ_l's and thus are individually informative. By contrast, the Euclidean distance or pseudo-max distance can avoid this drawback. Generally, as long as $\boldsymbol{\alpha}$ can generate sufficient variation in many (but not necessarily all) proxies and thus in the entire sequence $\{\eta_l(\boldsymbol{\alpha})\}_{l\in\mathcal{C}^{\dagger}}$, the lower bound requirement can be satisfied for Euclidean distance ($\underline{\varsigma} = 2$) and pseudo-max distance ($\underline{\varsigma} = 1$). It is unnecessary to know the identities of informative proxies 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. ⌟

Next, we impose a high-level condition on the local approximation (3.2). It ensures that the local PCA output from the noisy proxy matrix $\boldsymbol{X}^{\ddagger}_{\mathcal{N}_i}$ can aid the approximation of the unknown functions $\mu(\cdot)$ and $e(\cdot)$ in the outcome and treatment equations. Recall that by construction, $\boldsymbol{\Upsilon}_{\mathcal{N}_i} := \boldsymbol{\Lambda}'_{\mathcal{N}_i}\boldsymbol{\Lambda}_{\mathcal{N}_i}/K$ is a diagonal matrix of top d_i eigenvalues of $(\boldsymbol{H}^{\ddagger}_{\mathcal{N}_i})'\boldsymbol{H}^{\ddagger}_{\mathcal{N}_i}/pK$. Without loss of generality, write $\boldsymbol{\Upsilon}_{\mathcal{N}_i} = \mathrm{diag}\{\upsilon^2_{1,\mathcal{N}_i}, \cdots, \upsilon^2_{\mathsf{d}_i,\mathcal{N}_i}\}$ with $\upsilon_{1,\mathcal{N}_i} \geq \upsilon_{2,\mathcal{N}_i} \geq \cdots \geq \upsilon_{\mathsf{d}_i,\mathcal{N}_i} > 0$. $\boldsymbol{\lambda}_{j,\mathcal{N}_i}$ denotes the loading vector in $\boldsymbol{\Lambda}_{\mathcal{N}_i}$ corresponding to unit j. Recall $\delta_{Kp} = (K^{1/2} \wedge p^{1/2})/\sqrt{\log(n \vee p)}$.

Assumption 5 (Local Approximation)**.** *For each $i \in [n]$, the following conditions hold:*

(a) $\max_{i\in[n]} \|\boldsymbol{\Lambda}_{\mathcal{N}_i}\boldsymbol{\Upsilon}^{-1}_{\mathcal{N}_i}\|_{\max} \lesssim_{\mathbb{P}} 1$*; either* $\upsilon_{j,\mathcal{N}_i}/\upsilon_{j+1,\mathcal{N}_i} \lesssim 1$ *or* $\upsilon_{j,\mathcal{N}_i}/\upsilon_{j+1,\mathcal{N}_i} \to \infty$ *for* $j \in [\mathsf{d}_i - 1]$ *and* $\upsilon_{\mathsf{d}_i,\mathcal{N}_i}/\delta^{-1}_{Kp} \to \infty$*, with probability approaching one.*

(b) $\mathfrak{r}_{\eta,\mathcal{N}_i} := \|\boldsymbol{\Xi}_{\mathcal{N}_i}\|_{\max} = \bar{o}_{\mathbb{P}}(\upsilon_{\mathsf{d}_i,\mathcal{N}_i})$*, and there exist some* $\mathfrak{b}_{\mathcal{N}_i}, \mathfrak{c}_{\mathcal{N}_i} \in \mathbb{R}^{\mathsf{d}_i}$ *such that* $\mathfrak{r}_{\mu,\mathcal{N}_i} := \max_{j\in\mathcal{N}_i} |\mu(\boldsymbol{\alpha}_j) - \mathfrak{b}'_{\mathcal{N}_i}\boldsymbol{\lambda}_{j,\mathcal{N}_i}| = \bar{o}_{\mathbb{P}}(1)$ *and* $\mathfrak{r}_{e,\mathcal{N}_i} := \max_{j\in\mathcal{N}_i} |e(\boldsymbol{\alpha}_j) - \mathfrak{c}'_{\mathcal{N}_i}\boldsymbol{\lambda}_{j,\mathcal{N}_i}| = \bar{o}_{\mathbb{P}}(1)$.

Assumption 5 is a high-level condition, delineating the key requirements for informativeness of the local factor loadings $\boldsymbol{\Lambda}_{\mathcal{N}_i}$. We make several remarks on each part, and more detailed discussions, including sufficient conditions for Assumption 5, are available in Section SA-2.3 of the SA.

First, $\boldsymbol{\Lambda}_{\mathcal{N}_i}$ can be viewed as an approximation basis that characterizes the relation of different units in terms of $\boldsymbol{\alpha}_i$. Thus, Assumption 5(a) is a mild condition that specifies the possibly heterogeneous magnitude of different components of the basis. The requirement that $\upsilon_{\mathsf{d}_i,\mathcal{N}_i}$ is much larger than δ^{-1}_{Kp} is standard in the factor model literature (Wang and Fan, 2017), implying that the "weakest" factors to be extracted has stronger signal strength compared to the noise (and thus are consistently estimable).

Second, Assumption 5(b) quantifies the approximation accuracy for the three nonparametric components in our model, $\boldsymbol{\eta}(\cdot)$, $\mu(\cdot)$ and $e(\cdot)$, through the corresponding error bounds $\mathfrak{r}_{\eta,\mathcal{N}_i}$, $\mathfrak{r}_{\mu,\mathcal{N}_i}$, and $\mathfrak{r}_{e,\mathcal{N}_i}$. The informativeness of $\boldsymbol{\Lambda}_{\mathcal{N}_i}$ in this context is made clear by the assumption that the same basis $\boldsymbol{\Lambda}_{\mathcal{N}_i}$ extracted from the noise-free structure $\boldsymbol{H}^{\ddagger}_{\mathcal{N}_i}$ also has

ability to approximate $\mu(\cdot)$ and $e(\cdot)$. Usually, such conditions are verified by imposing some smoothness restrictions on $\eta_l(\cdot)$'s, $\mu(\cdot)$ and $e(\cdot)$. See examples below.

Example 4.1 (Linear factor model)**.** To fix ideas, consider a linear model $\eta_l(\boldsymbol{\alpha}_i) = (\boldsymbol{C}\boldsymbol{\alpha}_i)'\boldsymbol{\varpi}_l$ for some $\boldsymbol{\varpi}_l \in \mathbb{R}^{r_\alpha}$. We introduce a diagonal matrix $\boldsymbol{C} = \text{diag}\{c_1, \cdots, c_{r_\alpha}\}$ to explicitly characterize the "signal strength" of factors. Assume that $\sum_{l\in\mathcal{C}^\ddagger} \boldsymbol{\varpi}_l\boldsymbol{\varpi}_l'/p^\ddagger$ and $\sum_{j\in\mathcal{N}_i} \boldsymbol{\alpha}_j\boldsymbol{\alpha}_j'/K$ are nondegenerate (eigenvalues are bounded away from zero). If $\delta_{Kp}^{-1} = o(\min_j |c_j|)$ and $\mathsf{d}_i = r_\alpha$, Assumption 5(a) holds with $\mathfrak{r}_{\eta,\mathcal{N}_i} = 0$ (exact approximation). Note that in this example, the extracted d_i factor loadings may miss a local constant term, but we can always include an intercept in $\boldsymbol{z}_i$ in the quasi-MLE estimation of $\mu(\cdot)$ and $e(\cdot)$. Therefore, if $\mu(\cdot)$ and $e(\cdot)$ are twice continuously differentiable, we can achieve local linear approximations of $\mu(\cdot)$ and $e(\cdot)$. As will be formalized below, typically, the implicit bandwidth of K-NN matching is $O((K/n)^{1/r_\alpha})$. Thus, Assumption 5(b) holds with $\mathfrak{r}_{\mu,\mathcal{N}_i}$ and $\mathfrak{r}_{e,\mathcal{N}_i}$ of order $O((K/n)^{2/r_\alpha})$. ⌟

Example 4.2 (Nonlinear models)**.** Suppose that $\max_{j\in\mathcal{N}_i} \|\boldsymbol{\alpha}_j - \boldsymbol{\alpha}_i\| \lesssim_{\mathbb{P}} h_n$ for some "bandwidth" h_n, and $\mu(\cdot)$, $e(\cdot)$ and $\eta_l(\cdot)$'s are twice continuously differentiable with bounded derivatives. Let $\nabla\eta_l(\cdot)$ denote the gradient vector of $\eta_l(\cdot)$. Then, $|\eta_l(\boldsymbol{\alpha}_j) - \eta_l(\boldsymbol{\alpha}_i) - \nabla\boldsymbol{\eta}_l(\boldsymbol{\alpha}_i)'(\boldsymbol{\alpha}_j - \boldsymbol{\alpha}_i)| \lesssim_{\mathbb{P}} h_n^2$ for all $l \in [p]$. If $\frac{1}{p^\ddagger}\sum_{l\in\mathcal{C}^\ddagger}(\eta_l(\boldsymbol{\alpha}_i), \nabla\boldsymbol{\eta}_l(\boldsymbol{\alpha}_i)')'(\eta_l(\boldsymbol{\alpha}_i), \nabla\boldsymbol{\eta}_l(\boldsymbol{\alpha}_i)')$ is nondegenerate (i.e., the minimum eigenvalue is bounded away from zero), then we can let $\mathsf{d}_i = 1 + r_\alpha$, and Assumption 5(a) holds with $\upsilon_{1,\mathcal{N}_i} \asymp 1$, $\upsilon_{j,\mathcal{N}_i}, \asymp h_n$ for $j = 2, \cdots, 1 + r_\alpha$. Accordingly, Assumption 5(b) holds with $\mathfrak{r}_{\eta,\mathcal{N}_i}, \mathfrak{r}_{\mu,\mathcal{N}_i}$, and $\mathfrak{r}_{e,\mathcal{N}_i}$ of order $O(h_n^2)$. The nondegeneracy condition on partial derivatives of η_l's is often mild and has an intuitive interpretation: the tangent plane of the "surface" generated by $\boldsymbol{\eta}(\cdot)$ does not collapse in any directions at the point $\boldsymbol{\alpha}_i$ (so we do have r_α latent confounders). This idea extends to higher-order approximations. If each η_l is $(m+1)$-times continuously differentiable and the local factors formed by the partial derivatives of η_l up to order m are nondegenerate, then the approximation errors $\mathfrak{r}_{\eta,\mathcal{N}_i}$, $\mathfrak{r}_{\mu,\mathcal{N}_i}$, and $\mathfrak{r}_{e,\mathcal{N}_i}$ are $O(h_n^{m+1})$. See discussions in Section SA-2.3 of the SA. ⌟

Remark 4.3 (Heterogeneous approximation power)**.** We allow the number of factors extracted d_i to vary across local neighborhoods. It accommodates the possibility that the

nonlinear surface induced by $\boldsymbol{\eta}(\cdot)$ may exhibit varying degrees of nonlinearity at different points. Relative to the noise, some approximation terms may yield signals too weak to be recovered from the proxies in specific regions. Accordingly, the ability of the local basis $\boldsymbol{\Lambda}_{\mathcal{N}_i}$ to approximate $\mu(\boldsymbol{\alpha}_i)$ or $e(\boldsymbol{\alpha}_i)$ may also be heterogeneous across units. For instance, if $\boldsymbol{\eta}(\boldsymbol{\alpha})$ is almost linear at $\boldsymbol{\alpha} = \boldsymbol{\alpha}_i$, then the basis $\boldsymbol{\Lambda}_{\mathcal{N}_i}$ extracted from the proxies only capture the information about local constant and local linear basis, and thus we cannot achieve higher-order approximation of $\mu(\boldsymbol{\alpha})$ at $\boldsymbol{\alpha} = \boldsymbol{\alpha}_i$. This feature is more specific to this causal inference problem. If the goal is only to estimate the mean structure $\boldsymbol{H}$, the recommended strategy in Remark 3.4 (extracting all principal components whose singular values exceed the noise edge) may allow us to approximately achieve $\mathfrak{r}_{\eta,\mathcal{N}_i} = O_{\mathbb{P}}(\delta_{Kp}^{-1})$ across all units.

Two features of our method can alleviate the concern about such heterogeneity: (i) when p is large, a mixture of less nonlinear proxies and highly nonlinear ones may still permit the characterization of high-order basis; (ii) our main inference results only need the approximation errors to be sufficiently small in an *average* sense rather than uniformly over all units, so weak approximation power of $\boldsymbol{\Lambda}_{\mathcal{N}_i}$ is permitted in some (or even many) neighborhoods. See more discussion in Section SA-2.3 of the SA. ⌟

4.2 Theoretical Results

Recall that K is the number of nearest neighbors, p is the number of proxies, 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 indirect matching and local PCA are used to extract information on the latent variables $\boldsymbol{\alpha}_i$. Specifically, under Assumptions 2–4, if $\frac{\log(n/K)}{K} = o(1)$ and $\frac{K \log n}{n} = o(1)$, it can be shown that

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

It formalizes the idea that matching on the noisy proxies approximately match the latent

confounders as well. In particular, the implied matching discrepancy of nearest neighbors (the "bandwidth" for subsequent local estimation) shrinks as $K/n \to 0$ and $a_p \to 0$. The first term, $(K/n)^{\bar{\varsigma}/(\underline{\varsigma} r_\alpha)}$, is the usual discrepancy that would have existed even if the matching could be done directly on $\boldsymbol{\alpha}_i$. When $\underline{\varsigma} = \bar{\varsigma}$ (applicable to pseudo-max distance or distance of averages), the first term reduces to $(K/n)^{1/r_\alpha}$, which is the implicit "bandwidth" in K-NN matching. By contrast, the second term $a_p^{1/\underline{\varsigma}}$ arises from the fact that the matching can only be conducted on the noisy proxies rather than the latent confounders. As described in Remark 4.1, typically, $a_p \to 0$ as $p \to \infty$ (more proxies).

If Assumption 5 also holds and $(np)^{\frac{2}{\nu}} \delta_{Kp}^{-2} \lesssim 1$, it can be further shown that

$$\max_{i \in [n]} \|\widehat{\boldsymbol{\Lambda}}_{\mathcal{N}_i} - \boldsymbol{\Lambda}_{\mathcal{N}_i} \boldsymbol{R}_{\mathcal{N}_i}\|_{\max} \lesssim_{\mathbb{P}} \delta_{Kp}^{-1} + \mathfrak{r}_{\eta, \mathcal{N}_i},$$

for some rotation matrix $\boldsymbol{R}_{\mathcal{N}_i}$. It formalizes the idea that the extracted factor loading $\widehat{\boldsymbol{\Lambda}}_{\mathcal{N}_i}$ from the noisy proxies is "consistent" for the loading matrix $\boldsymbol{\Lambda}_{\mathcal{N}_i}$ that has approximation power for functions of the latent confounders. The error bound shrinks as both K and p grow large and if the approximation error $\mathfrak{r}_{\eta, \mathcal{N}_i}$ diminishes.

Built on these properties of local PCA, we obtain 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 QMLE method.

Theorem 4.1 (Local QMLE)**.** *Suppose that Assumptions 2–5 hold. If* $(np)^{\frac{2}{\nu}} \delta_{Kp}^{-2} \lesssim 1$*, then*

$$\max_{i \in [n]} |\widehat{\zeta}_i - \zeta_i| \lesssim_{\mathbb{P}} \delta_{Kp}^{-1} + \mathfrak{r}_{\eta, \mathcal{N}_i} + \mathfrak{r}_{\mu, \mathcal{N}_i} \quad \textit{and} \quad \max_{i \in [n]} |\widehat{\varrho}_i - \varrho_i| \lesssim_{\mathbb{P}} \delta_{Kp}^{-1} + \mathfrak{r}_{\eta, \mathcal{N}_i} + \mathfrak{r}_{e, \mathcal{N}_i}.$$

Detailed asymptotic expansions are given by Equation (SA-7.1) *in the SA.*

Noting that the functional forms of $\mu(\cdot)$ and $e(\cdot)$ are unspecified, the results in Theorem 4.1 match the usual nonparametric convergence rates: δ_{Kp}^{-1} is the variance term, and $\mathfrak{r}_{\mu, \mathcal{N}_i}$ and $\mathfrak{r}_{e, \mathcal{N}_i}$ reflect the smoothing bias. The major difference here is that the convergence rate

also depends on the number of proxies p and the approximation error $\mathfrak{r}_{\eta,\mathcal{N}_i}$, reflecting the fact that the information on $\boldsymbol{\alpha}_i$ is indirectly obtained via local PCA. Also, the convergence 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. This slightly differs from the classical nonparametric or semiparametric estimation where uniformity over the whole support of conditioning variables is derived (or assumed directly).

Remark 4.4 (CATE estimator)**.** In this setup a conditional average treatment effect given the confounders is defined as $\mathbb{E}[y_i(1) - y_i(0)|\boldsymbol{\alpha}_i, \boldsymbol{z}_i]$. Theorem 4.1 shows that $\widehat{\zeta}_i$ is a uniformly consistent estimator of $\mathbb{E}[y_i(0)|\boldsymbol{\alpha}_i, \boldsymbol{z}_i]$. The same method can be applied to estimate $\mathbb{E}[y_i(1)|\boldsymbol{\alpha}_i, \boldsymbol{z}_i]$. This allows us to construct a valid estimator of CATE with guaranteed uniform convergence rates and to explore treatment effect heterogeneity across different dimensions. ⌟

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* $(np)^{\frac{2}{\nu}}\delta_{Kp}^{-2} \lesssim 1$ *and* $\sqrt{n \log n}(\delta_{Kp}^{-2} + \sum_{i\in[n]}[\mathfrak{r}_{\eta,\mathcal{N}_i}^2 + \mathfrak{r}_{\mu,\mathcal{N}_i}^2 + \mathfrak{r}_{e,\mathcal{N}_i}^2]/n) = o_{\mathbb{P}}(1)$*, then*

(a) $\sqrt{n}(\widehat{\theta} - \theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \varphi_i + o_{\mathbb{P}}(1)$ *where* $\varphi_i = \frac{t_i(\zeta_i - \theta)}{\varrho} + \frac{\varrho_i}{\varrho}\frac{(1-t_i)(y_i - \zeta_i)}{1-\varrho_i}$*;*

(b) $\sqrt{n}(\widehat{\theta} - \theta)/\widehat{\sigma} \rightsquigarrow \mathsf{N}(0, 1)$.

Theorem 4.2 takes advantage of the doubly-robust score function and the properties of our first step estimator to relax conditions on the convergence rates of $\widehat{\zeta}_i$ and $\widehat{\varrho}_i$, while not requiring sample splitting across units. 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 results in the double/debiased machine learning literature. Importantly, the approximation errors in our model, characterized by

$\mathfrak{r}_{\eta,\mathcal{N}_i}$, $\mathfrak{r}_{\mu,\mathcal{N}_i}$ and $\mathfrak{r}_{e,\mathcal{N}_i}$, need *not* be small uniformly over all units. Instead, we only require that the *average* squared errors be small, allowing for cases in which the nearest neighbors and local factor loadings from the first step may not be highly informative about the latent confounders and yield poor approximation for some units.

To get a better sense of the rate restrictions imposed, consider a simple but practically relevant case with $K \lesssim p$, ν sufficiently large (e.g., sub-Gaussian errors), and all the unknown functions η_l's, μ and e sufficiently smooth. Then, $\sqrt{n \log n}\delta_{Kp}^{-2} = o(1)$ reduces to $n/K^2 = o(1)$, up to $\log n$ terms. On the other hand, in the case of local linear approximation in Example 4.2, $\mathfrak{r}_{\eta,\mathcal{N}_i}$, $\mathfrak{r}_{\mu,\mathcal{N}_i}$ and $\mathfrak{r}_{e,\mathcal{N}_i}$ are typically $O((K/n)^{2/r_\alpha})$, and thus we roughly need $\sqrt{n}(K/n)^{4/r_\alpha} = o(1)$. If we use the recommended strategy in Remark 3.4 and extract all approximation terms with signal strength exceeding the noise level, then in the best case $\mathfrak{r}_{\eta,\mathcal{N}_i}, \mathfrak{r}_{\mu,\mathcal{N}_i}$ and $\mathfrak{r}_{e,\mathcal{N}_i}$ can be as small as $O(\delta_{Kp}^{-1})$. Then, all restrictions imposed in Theorem 4.2 reduces to $n/K^2 = o(1)$, up to $\log n$ terms.

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 Monte Carlo experiments to check the finite sample performance of our proposed method. The SA contains complete results, and here we give a short summary. We consider (i) three specifications for $\boldsymbol{\eta}(\cdot)$ with one latent confounder ($r_\alpha = 1$) and three with two latent confounders ($r_\alpha = 2$), with varying degree of nonlinearity and smoothness; (ii) different sample size combinations $(n, p) = (500, 125), (500, 250), (500, 500), (500, 750)$; (iii) three specifications for the measurement errors u_{il}'s: i.i.d., conditionally heteroskedastic,

and autocorrelated across l (and correlated with ϵ_i); and (iv) different choices of the tuning parameter K. We also compare our method with three alternative ones in this setting: (i) the usual doubly robust estimator that controls for two x_{il}'s directly, which mimics the practice that "naively" regards the noisy proxies as the true confounders ("DR"); (ii) the doubly robust estimator that employs double-lasso to select controls among the high-dimensional $\boldsymbol{x}_i$ ("DR-LASSO"); and (iii) the doubly robust estimator that uses local average smoothing to estimate conditional means of outcomes and propensity scores (i.e., taking local averages of nearest neighbors without the PCA step). The idea of local average smoothing has been proposed in the panel and network data literature for different purposes (e.g., Zhang et al., 2017; Bonhomme et al., 2022).

To conserve space, we only summarize results for one setup. As shown by our theory, the key feature that determines the performance of our method is the signal strength in the nonlinear factor model, reflected by the effective rank of the noise-free matrix $\boldsymbol{H}$. Therefore, we consider the Gaussian kernel $\eta_l(\alpha_i) = 2\exp(-(\alpha_i - \varpi_l)^2/\sigma)$, for which we can conveniently change the smoothness of $\boldsymbol{\eta}(\cdot)$ by varying σ. Typically, decreasing σ makes $\boldsymbol{H}$ more complex with more non-negligible singular values ("high-rank"). The measurement error u_{il}, given α_i and ϖ_l, is independent over i and l and follows zero-mean normal distribution with the conditional variance depending on α_i and ϖ_l. Also, $\mathbb{E}[y_i(0)|\alpha_i] = \alpha_i + \alpha_i^2$ and $\mathbb{E}[y_i(1)|\alpha_i] = 2\alpha_i + \alpha_i^2 + 1$, and the binary treatment is $t_i = \mathbb{1}(v_i \leq \varrho_i)$ with $\varrho_i = \psi((\alpha_i - 0.5) + (\alpha_i - 0.5)^2)$. $\alpha_i, \varpi_l, v_i \sim \mathsf{U}[0,1]$, and the error terms in $y_i(0)$ and $y_i(1)$ are standard normal. We consider 2 000 simulated datasets with $n = 500, p = 250$, which matches the sample sizes in our empirical application. For each simulated dataset, we obtain a point estimate of ATT, and report 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 five columns are obtained based on our proposed method, using the same strategy as described in our empirical illustration. Specifically, we take the first half ($p^\dagger = 125$) of proxies for matching and the second half ($p^\ddagger = 125$) for extracting the

local factor loadings. We let $K = Cn^{4/5}$ for $C = 0.5,\ 1,\ 1.5,\ 2$ respectively, which has the MSE-optimal rate 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 proxy x_{ip}. We choose the largest d_i such that the extracted factors are associated with singular values greater than $C_n(\sqrt{n}+\sqrt{K})$ for a large constant $C_n > 0$. For comparison, we report results based on DR and DR-DL in the last two columns.

Our method performs relatively well throughout the three specifications, and is robust to different choices of K in a wide range. By contrast, DR and DR-DL have unstable performance across different specifications, do not achieve the nominal coverage rate and usually have wider confidence intervals. The results in the SA further shows that (i) the local average smoothing has more unstable performance across K's and suffers from more severe undercoverage in more complex models (e.g., with multiple confounders); and (ii) larger p improves the performance of our method, while for small p, the proposed method may undercover but still have better performance compared to DR and DR-DL.

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, i.e., 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

Table 3: Simulation Results (heteroskedastic error, $n = 500$, $p = 250$)

	Local PCA, $K =$					**DR**	**DR-DL**
	73	145	217	289	$\widehat{K}_{\mathtt{CV}}$		
low-rank							
RMSE	0.100	0.098	0.098	0.097	0.099	0.173	0.187
CR	0.943	0.942	0.949	0.950	0.944	0.867	0.838
AL	0.378	0.375	0.374	0.373	0.375	0.509	0.509
mid-rank							
RMSE	0.096	0.094	0.093	0.093	0.094	0.156	0.147
CR	0.947	0.951	0.951	0.951	0.951	0.916	0.915
AL	0.375	0.372	0.372	0.373	0.373	0.501	0.481
high-rank							
RMSE	0.100	0.098	0.099	0.102	0.099	0.177	0.196
CR	0.938	0.940	0.944	0.940	0.940	0.861	0.823
AL	0.376	0.375	0.377	0.381	0.376	0.511	0.517

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

function class $\mathcal{G} = \{y \mapsto \mathbb{1}(y \leq \tau) : \tau \in \mathcal{Y}\}$ for some compact set $\mathcal{Y} \subset \mathbb{R}$. Each $g(\cdot) \in \mathcal{G}$ corresponds to a particular value $\tau \in \mathcal{Y}$. Therefore, we write $y_{i,\tau}(t) = \mathbb{1}(y_i(t) \leq \tau)$, $t = 0, 1$, and $y_{i,\tau} = \mathbb{1}(y_i \leq \tau)$. One may be interested in $\mathbb{E}[\mathbb{1}(y_i(t) \leq \tau)|t_i = \tilde{t}]$, the counterfactual distribution function of the potential outcome $y_i(t)$ for the group with a treatment status $\tilde{t}$, which is a parameter in $\ell^\infty(\mathcal{Y})$, a function space of bounded functions on $\mathcal{Y}$ equipped with sup-norm. For now, we focus on $\theta(\tau) = \mathbb{E}[\mathbb{1}(y_i(0) \leq \tau)|t_i = 1]$ in our binary treatment framework, and defer the general case to the SA. Now, Equation (2.2) becomes

$$y_{i,\tau}(0) = \zeta_{i,\tau} + \epsilon_{i,\tau}, \quad \zeta_{i,\tau} := \mathbb{P}(y_i(0) \leq \tau|\boldsymbol{\alpha}_i, \boldsymbol{z}_i) = \mu_\tau(\boldsymbol{\alpha}_i) + \boldsymbol{z}_i'\boldsymbol{\beta}_\tau.$$

For each τ, the second step of the estimation procedure in Section 3 is implemented to obtain an estimator $\widehat{\zeta}_{i,\tau}$ of $\zeta_{i,\tau}$.

To establish the limiting distribution of the proposed estimator, we impose an additional assumption.

Assumption 6 (Uniform Inference)**.** *$\mu_\tau(\boldsymbol{\alpha})$ is Lipschitz in τ uniformly over $\boldsymbol{\alpha}$; for all $\tau \in \mathcal{Y}$,*

there exists some $\mathfrak{b}_{\tau,\mathcal{N}_i} \in \mathbb{R}^{\mathsf{d}_i}$ *such that* $\bar{\mathfrak{r}}_{\mu,\mathcal{N}_i} := \max_{\tau\in\mathcal{Y}} \max_{j\in\mathcal{N}_i} |\mu_\tau(\boldsymbol{\alpha}_j) - \mathfrak{b}'_{\tau,\mathcal{N}_i}\boldsymbol{\lambda}_{j,\mathcal{N}_i}| = \bar{o}_{\mathbb{P}}(1)$.

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* $(np)^{\frac{2}{\nu}}\delta_{Kp}^{-2} \lesssim 1$ *and* $\sqrt{n\log n}(\delta_{Kp}^{-2} + \sum_{i\in[n]}[\mathfrak{r}^2_{\eta,\mathcal{N}_i} + \mathfrak{r}^2_{\mu,\mathcal{N}_i} + \mathfrak{r}^2_{e,\mathcal{N}_i}]/n) = o_{\mathbb{P}}(1)$*, then*

$$\sqrt{n}\Big(\theta(\cdot) - \theta(\cdot)\Big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi_i(\cdot) + o_{\mathbb{P}}(1) \rightsquigarrow \mathsf{Z}(\cdot) \quad \textit{in} \quad \ell^\infty(\mathcal{Y}),$$

where $\varphi_i(\cdot) = \frac{t_i(\zeta_{i,\cdot} - \theta(\cdot))}{\varrho} + \frac{\varrho_i}{\varrho}\frac{(1-t_i)(y_{i,\cdot} - \zeta_{i,\cdot})}{1-\varrho_i}$ *and* $\mathsf{Z}(\cdot)$ *is a zero-mean Gaussian process with covariance kernel* $\mathbb{E}[\varphi_i(\tau_1)\varphi_i(\tau_2)]$ *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* $\tilde{\theta}$ *as an element of a parameter space* $D\tilde{\theta} \subseteq \ell^\infty(\mathcal{Y})$*, a function space of bounded functions on* $\mathcal{Y}$*, with* $D\tilde{\theta}$ *containing the true value* θ*. Let a functional* $\Psi(\tilde{\theta})$ *mapping* $D\tilde{\theta}$ *to* $\ell^\infty(\mathcal{Q})$ *be Hadamard differentiable in* $\tilde{\theta}$ *at* θ *with derivative* $\dot{\Psi}_\theta$*. Then,* $|\sqrt{n}(\Psi(\widehat{\theta})(\cdot) - \Psi(\theta)(\cdot)) - n^{-1/2}\sum_{i=1}^{n}\dot{\Psi}_\theta(\varphi_i)(\cdot)| = o_{\mathbb{P}}(1)$ *and* $\sqrt{n}(\Psi(\widehat{\theta})(\cdot) - \Psi(\theta)(\cdot)) \rightsquigarrow \dot{\Psi}_\theta(\mathsf{Z})(\cdot)$ *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* $\dot{\Psi}_\theta(\varphi_i)$.

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. De-

fine a uniformly consistent estimator of $\varphi_i(\cdot)$:

$$\widehat{\varphi}_i(\cdot) = \frac{t_i(\widehat{\zeta}_{i,\cdot} - \widehat{\theta}(\cdot))}{\widehat{\varrho}} + \frac{\widehat{\varrho}_i}{\widehat{\varrho}} \frac{(1-t_i)(y_{i,\cdot} - \widehat{\zeta}_{i,\cdot})}{1-\widehat{\varrho}_i}.$$

The following corollary shows that conditional on the data, $\frac{1}{\sqrt{n}} \sum_{i=1}^n \omega_i \widehat{\varphi}_i(\cdot)$ weakly converges to the same limiting process $\mathsf{Z}(\cdot)$ 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,* $n^{-1/2} \sum_{i=1}^n \omega_i \widehat{\varphi}_i(\cdot) \rightsquigarrow \mathsf{Z}(\cdot)$ *that is the Gaussian process defined in Theorem 5.1 with probability approaching one.*

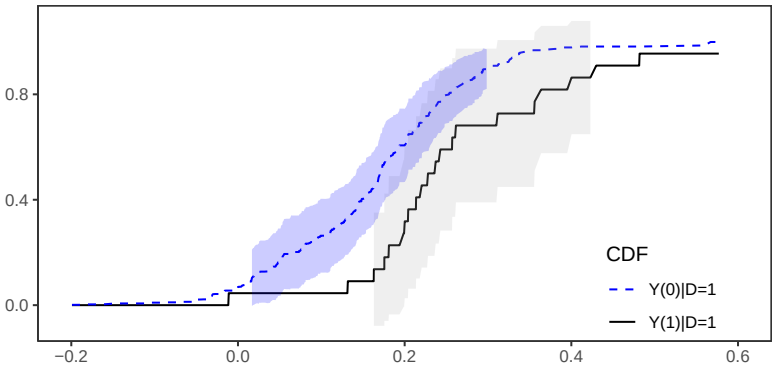
To showcase the uniform inference procedure, we use the data of Acemoglu et al. (2016) to check the (first-order) stochastic dominance (SD) of $\theta_1(\cdot)$ over $\theta_0(\cdot)$, where $\theta_1(\cdot)$ and $\theta_0(\cdot)$ respectively denote the cumulative distribution functions (CDFs) of potential stock returns of firms connected to Geithner if they were connected and not connected with him. By definition of SD, the null hypothesis is $\theta_1(\tau) \leq \theta_0(\tau)$ for all $\tau \in \mathcal{Y}$. An intuitive test statistic is $\sqrt{n} \sup_{\tau\in\mathcal{Y}}(\theta_1(\tau) - \theta_0(\tau))$. The null hypothesis is rejected if the test statistic is greater than a certain critical value. Given the asymptotic expansions of the corresponding estimators $\widehat{\theta}_1(\cdot)$ and $\widehat{\theta}_0(\cdot)$, the critical value can be obtained by simulating the supremum of the approximation process. 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 4. 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 5.81 for a confidence level of 0.95 obtained through simulation. Thus, SD of $\theta_1(\cdot)$ over $\theta_0(\cdot)$ 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 connected firms, which is a stronger conclusion than that based simply on the mean in Section 3.

Figure 4: Estimated CDFs of Potential Outcomes for the Treated



Notes: This figure plots the estimated cumulative distribution functions of potential returns for connected firms. Shaded regions are 95% confidence bands.

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 proxies 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., 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.

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.

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

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.

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

Busso, M., DiNardo, J., and McCrary, J. (2014), "New Evidence on the Finite Sample Properties of Propensity Score Matching and Reweighting Estimators," *The Review of Economics and Statistics*, 96, 885–897.

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.

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2009), "Dealing with Limited Overlap in Estimation of Average Treatment Effects," *Biometrika*, 96, 187–199.

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

Dobriban, E., and Owen, A. B. (2019), "Deterministic Parallel Analysis: An Improved Method for Selecting Factors and Principal Components," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81, 163–183.

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.

Farrell, M. H., Liang, T., and Misra, S. (2021), "Deep Neural Networks for Estimation and Inference," *Econometrica*, 89, 181–213.

Feng, Y. (2024), "Optimal Estimation of Large-Dimensional Nonlinear Factor Models," *arXiv preprint arXiv:2311.07243*.

Ferraty, F., and Vieu, P. (2006), *Nonparametric Functional Data Analysis: Theory and Practice*, New York: Springer Science & Business Media.

Gao, C., Lu, Y., and Zhou, H. H. (2015), "Rate-Optimal Graphon Estimation," *Annals of Statistics*, 43, 2624–2652.

Härdle, W. K., Müller, M., Sperlich, S., and Werwatz, A. (2004), *Nonparametric and Semiparametric Models*, Berlin: Springer.

Holmström, B., and Tirole, J. (1993), "Market Liquidity and Performance Monitoring," *Journal of Political economy*, 101, 678–709.

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

Johnstone, I. M. (2001), "On the Distribution of the Largest Eigenvalue in Principal Components Analysis," *The Annals of statistics*, 29, 295–327.

Ma, X., and Wang, J. (2020), "Robust Inference Using Inverse Probability Weighting," *Journal of the American Statistical Association*, 115, 1851–1860.

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*.

Pan, Y., Wang, T. Y., and Weisbach, M. S. (2015), "Learning about CEO ability and stock return volatility," *The Review of Financial Studies*, 28, 1623–1666.

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.

Singh, R. (2020), "Kernel Methods for Unobserved Confounding: Negative Controls, Proxies, and Instruments," *arXiv preprint arXiv:2012.10315*.

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.