Econometric Theory, **39**, 2023, 357–388.
doi:10.1017/S0266466621000554

NONPARAMETRIC ESTIMATION OF GENERALIZED TRANSFORMATION MODELS WITH FIXED EFFECTS

SONGNIAN CHEN
Hong Kong University of Science and Technology

XUN LU
The Chinese University of Hong Kong

XI WANG
Jinan University

This paper considers a generalized panel data transformation model with fixed effects where the structural function is assumed to be additive. In our model, no parametric assumptions are imposed on the transformation function, the structural function, or the distribution of the idiosyncratic error term. The model is widely applicable and includes many popular panel data models as special cases. We propose a kernel-based nonparametric estimator for the structural function. The estimator has a closed-form solution and is easy to implement. We study the asymptotic properties of our estimator and show that it is asymptotically normally distributed. The Monte Carlo simulations demonstrate that our new estimator performs well in finite samples.

1. INTRODUCTION

The transformation model has received considerable attention since the pioneering work of Box and Cox (1964). It includes many popular models as special cases, including the accelerated failure time model, the Weibull hazard model, the proportional hazard model, and the mixed proportional hazard model, among others. The transformation model has been widely applied in many fields of economics, such as labor economics, insurance, and industrial organization (see, e.g., Ham and Rea Jr., 1987; Lancaster, 1997; Bhattacharjee et al., 2007). The panel data transformation model is particularly appealing, as it allows fixed effects

We would like to thank the Editor, Peter C.B. Phillips, a Co-Editor, and two anonymous referees for their many constructive comments. We also thank Songxi Chen, Yingyao Hu, Whitney Newey, and the conference and seminar participants at the CEMP Econometric Workshop, 2019 Guangzhou Econometrics Workshop, and Peking University for their helpful comments. Lu acknowledges support from the Hong Kong Research Grants Council (RGC) under grant number 16504819, National Natural Science Foundation of China (NSFC) under grant number 72133002, and the Chinese University of Hong Kong for the start-up fund. Wang gratefully acknowledges the support by the National Natural Science Foundation of China Grant through NSFC7190312. Address correspondence to Songnian Chen, Department of Economics, Hong Kong University of Science and Technology, Kowloon, Hong Kong; e-mail: snchen@ust.hk.

to account for unobservable individual heterogeneity. For example, Topel and Ward (1992) and Pitt (2007) use the panel data transformation models to study the issues of unemployment spells and insurance claim durations, respectively. However, most of the existing panel data transformation models with fixed effects need to parametrically specify the transformation function, the structural function, or the distribution of the idiosyncratic error term. In practice, those parametric specifications can be misspecified. It is well known that estimation and inference based on misspecified models can be seriously misleading. To this end, this paper considers a generalized panel data transformation model with fixed effects where no parametric assumptions are imposed on the transformation function, the structural function, or the distribution of the idiosyncratic error term.

Specifically, we consider the following model:

$$\Lambda(Y_{it}) = \sum_{k=1}^{K} m_k(X_{k,it}) + \alpha_i + \varepsilon_{it}, i = 1, \ldots, n \text{ and } t = 1, \ldots, T, \tag{1.1}$$

where Y_{it} is the observed dependent variable, $(X_{1,it}, X_{2,it}, \ldots, X_{K,it})'$ is a $K \times 1$ vector of observed covariates, α_i is the individual fixed effect, ε_{it} is the idiosyncratic error term, and $\Lambda(\cdot), m_k(\cdot), k = 1, \ldots, K$, are all general unknown functions: $\mathbb{R} \to \mathbb{R}$. Here, $\Lambda(\cdot)$ is a strictly increasing transformation function and $(m_1(\cdot), \ldots, m_K(\cdot))$ are the K structural functions.[1] In applications, the K structural functions $(m_1(\cdot), \ldots, m_K(\cdot))$ are often the parameters of interest. In addition, their derivatives, $(m_1'(\cdot), \ldots, m_K'(\cdot))$, which measure the marginal effects, may be of interest. For example, $m_k'(X_{k,it})$ can be interpreted as the marginal effect of $X_{k,it}$ on $\Lambda(Y_{it})$, where $\Lambda(\cdot)$ is the cumulative (integrated) hazard function in survival models. In addition, based on the estimator of $m_k'(\cdot), k = 1, \ldots, K$, we can obtain the relative marginal effects as

$$\frac{\partial Y_{it}}{\partial X_{k,it}} \Big/ \frac{\partial Y_{it}}{\partial X_{k',it}} = \frac{m_k'\left(X_{k,it}\right)}{m_{k'}'\left(X_{k',it}\right)} \text{ for any } k \text{ and } k'.$$

The main goal of this paper is to estimate $(m_1(\cdot), \ldots, m_K(\cdot))$ and their derivatives $(m_1'(\cdot), \ldots, m_K'(\cdot))$.

Here, we impose an additive assumption on the structural function to partially address the issue of "curse of dimensionality." In the extension section, we also briefly discuss a more general model where we do not impose the additive structure, i.e.,

$$\Lambda(Y_{it}) = m(X_{1,it}, X_{2,it}, \ldots, X_{K,it}) + \alpha_i + \varepsilon_{it}, \; i = 1, \ldots, n \text{ and } t = 1, \ldots, T, \tag{1.2}$$

with an unknown general function $m(\cdot) : \mathbb{R}^K \to \mathbb{R}$.

We consider a short panel where T is fixed, and allow the fixed-effects α_i to be correlated with the covariates arbitrarily. Here, the main challenge is to deal with

[1] For simplicity, we refer to $m_k(\cdot), k = 1, \ldots, K$, as structural functions. Strictly speaking, they are only components of the structural function, as the relationship between $X_{k,it}$ and Y_{it} also depends on the transformation function.

https://doi.org/10.1017/S0266466621000554 Published online by Cambridge University Press

the fixed-effects α_i. Note that our model can be written as

$$Y_{it} = \Lambda^{-1}\left(\sum_{k=1}^{K} m_k(X_{k,it}) + \alpha_i + \varepsilon_{it}\right),$$

where $\Lambda^{-1}(\cdot)$ is the inverse function of $\Lambda(\cdot)$. We cannot use the simple first difference or within transformation to remove α_i due to the nonlinear function $\Lambda^{-1}(\cdot)$. Our model is different from the following model:

$$Y_{it} = \Lambda^{-1}\left(\sum_{k=1}^{K} m_k(X_{k,it})\right) + \alpha_i + \varepsilon_{it},$$

where α_i can be easily removed by the first difference. We impose the key assumption that $\{X_{1,it}, X_{2,it}, \ldots, X_{K,it}\}_{t=1,\ldots,T}$ and $\{\varepsilon_{it}\}_{t=1,\ldots,T}$ are independent over the cross-sectional dimension i, and $\{\varepsilon_{it}\}_{t=1,\ldots,T}$ are exchangeable. Let $\mathbf{1}\{\cdot\}$ be the indicator function. The main innovation of our method is that we consider the binary variable $\mathbf{1}\{Y_{it} > Y_{i,t-1}\}$ that compares Y_{it} and $Y_{i,t-1}$ instead of considering Y_{it} and $Y_{i,t-1}$ themselves. We show that the conditional expectation of $\mathbf{1}\{Y_{it} > Y_{i,t-1}\}$ does not depend on the fixed-effects α_i, and can be approximated by a linear function locally in a neighborhood. Based on the local linear approximation, we propose a kernel-based estimator. Our estimator is easy to implement, as it is based on a closed-form solution and no optimization is required. We show that our estimator is asymptotically normally distributed.

To the best of our knowledge, our paper is the first attempt to address nonparametric generalized panel data transformation models with fixed effects. Our paper is related to two strands of the panel data literature: the nonparametric models and the transformation models. We provide a brief review of each.[2] There is a relatively large literature on the nonparametric estimation of panel data models without the transformation function (see, e.g., Chapter 19 in Li and Racine, 2007 for a review). Henderson, Carroll, and Li (2008) and Su and Lu (2013) propose nonparametric estimators for the model

$$Y_{it} = m(X_{1,it}, X_{2,it}, \ldots, X_{K,it}) + \alpha_i + \varepsilon_{it}, \tag{1.3}$$

where $m(\cdot)$ is a general function: $\mathbb{R}^K \to \mathbb{R}$. Li and Stengos (1996), Baltagi and Li (2002), You, Zhou, and Zhou (2011), Ai, You, and Zhou (2014), and Su and Zhang (2016) study the partially linear model

$$Y_{it} = \beta' Z_{it} + m(X_{1,it}, X_{2,it}, \ldots, X_{K,it}) + \alpha_i + \varepsilon_{it}, \tag{1.4}$$

where Z_{it} is a vector of covariates that enters the linear component of the model, β is a vector of the corresponding coefficients, and $m(\cdot)$ is a general function: $\mathbb{R}^K \to \mathbb{R}$. Mammen, Støve, and Tjøstheim (2009) consider the nonparametric additive

[2]To save space, we do not review the two large bodies of literature: linear panel data models and transformation models with cross-sectional data, which are less relevant to our model.

https://doi.org/10.1017/S0266466621000554 Published online by Cambridge University Press

model

$$Y_{it} = \sum_{k=1}^{K} m_k(X_{k,it}) + \alpha_i + \varepsilon_{it}. \tag{1.5}$$

For models (1.3), (1.4), and (1.5), the fixed-effects α_i can be easily differenced out. Evdokimov (2011) considers a model

$$Y_{it} = m(X_{1,it}, X_{2,it}, \ldots, X_{K,it}, \alpha_i) + \varepsilon_{it},$$

where $m(\cdot)$ is a general function: $\mathbb{R}^{K+1} \to \mathbb{R}$ and the fixed-effects α_i enter the structural function nonseparably. Evdokimov (2011) proposes an estimator of the structural function $m(\cdot)$ based on the method of conditional deconvolution. Hoderlein and White (2012) and Chernozhukov et al. (2013) consider a more general nonseparable model

$$Y_{it} = m(X_{1,it}, X_{2,it}, \ldots, X_{K,it}, \alpha_i, \varepsilon_{it}),$$

where $m(\cdot)$ is a general function: $\mathbb{R}^{K+2} \to \mathbb{R}$. They estimate certain average effects, such as average treatment effects and local average responses. Here, we impose the assumption on the general structural function $m(\cdot)$ as in model (1.1), and estimate the components of the structural functions $(m_1(\cdot), \ldots, m_K(\cdot))$.

Another strand of literature related to our paper is panel data transformation models. So far, the existing panel data transformation models in the literature are usually parametric where the structural function is assumed to take a single index form, that is,

$$\Lambda(Y_{it}) = \mathbf{X}_{it}'\beta + \alpha_i + \varepsilon_{it}, \tag{1.6}$$

where $\mathbf{X}_{it} = (X_{1,it}, X_{2,it}, \ldots, X_{K,it})'$ and $\beta = (\beta_1, \ldots, \beta_K)$ is the corresponding slope coefficients. Horowitz and Lee (2004) and Lee (2008) propose partial and weighted likelihood estimators assuming that the distribution of the error term is parametrically specified. Abrevaya (1999) proposes a leapfrog estimator for β based on the ranking principle. Chen (2010) imposes a strict stationary assumption that marginal distributions of ε_{it} are identical over time and proposes a $\sqrt{n}$-consistent estimator for β. Botosaru and Muris (2018) consider an estimator via binarization. Wang and Chen (2020) relax the stationarity assumption and extend the method to nonstationary panel data. Abrevaya (2000) and Chen and Wang (2018) consider a more general model which allows α_i to enter the structural function nonseparably

$$Y_{it} = m\left(\mathbf{X}_{it}'\beta, \alpha_i, \varepsilon_{it}\right),$$

for a general function $m(\cdot) : \mathbb{R}^3 \to \mathbb{R}$. Although the linear form $\mathbf{X}_{it}'\beta$ can simplify the estimation procedure, it can be restrictive and misspecified in practice. For example, it is well acknowledged that the single index form $\mathbf{X}_{it}'\beta$ precludes the ratio of marginal effects of different covariates (i.e., relative effects) to vary. That

https://doi.org/10.1017/S0266466621000554 Published online by Cambridge University Press

is, the ratio of marginal effects of $X_{k,it}$ and $X_{k',it}$ on Y_{it} is

$$\frac{dY_{it}}{dX_{k,it}} \Big/ \frac{dY_{it}}{dX_{k',it}} = \frac{\beta_k}{\beta_{k'}},$$

which does not depend on the level of $X_{it,k}$ and $X_{it,k'}$. To overcome the limitations of the existing methods, our paper relaxes the linear index assumption in the transformation models. Note that even in model (1.1) with an additive structure, the marginal effects or the ratio can vary with the level of covariates. In addition, many existing estimators are solutions to nonconvex optimization problems; therefore, the computational burden can be heavy. In contrast, our estimator is based on a closed-form solution and is computationally easy.

The rest of this paper is organized as follows. Section 2 proposes a nonparametric estimator for the structural function. In Section 3, we study the large sample properties of the proposed estimator. In Section 4, we conduct Monte Carlo simulations to illustrate the finite-sample performance of our estimator. Section 5 discusses two extensions of the model. First, we extend our method to a more general model where the structural model is not necessarily additive. Second, we discuss how to apply our method to cross-sectional data. Section 6 concludes our paper.

2. MODELS AND ESTIMATORS

In this section, we propose our estimators for the structural function $m_k(\cdot)$ and its derivative $m_k'(\cdot)$ for model (1.1). Throughout the paper, random variables are denoted using uppercase letters (e.g., $X_{k,it}$) and their realizations using lowercase (e.g., x_k). Calligraphic fonts are used to denote the supports of the corresponding variables (e.g., $\mathcal{X}_k$).

We first consider the simple case where $T = 2$. We assume that all covariates are continuous variables. Accordingly, model (1.1) can be written as

$$\begin{aligned}\Lambda(Y_{i1}) &= m(\mathbf{X}_{i1}) + \alpha_i + \varepsilon_{i1} = \sum_{k=1}^{K} m_k(X_{k,i1}) + \alpha_i + \varepsilon_{i1}, \text{ and} \\ \Lambda(Y_{i2}) &= m(\mathbf{X}_{i2}) + \alpha_i + \varepsilon_{i2} = \sum_{k=1}^{K} m_k(X_{k,i2}) + \alpha_i + \varepsilon_{i2}, \end{aligned} \tag{2.1}$$

where $\mathbf{X}_{it} = (X_{1,it}, X_{2,it}, \ldots, X_{K,it})'$ and $m(\mathbf{X}_{it}) = \sum_{k=1}^{K} m_k(X_{k,it})$ is assumed to be additive. We assume that each structural function $m_k(\cdot), k = 1, \ldots, K$, is unknown and continuously differentiable. The fixed-effects α_i are time-invariant and capture individuals' unobserved heterogeneity. Without loss of generality, we focus on the first structural function $m_1(\cdot)$. For ease of notation, we write $\mathbf{X}_{it} = (X_{1,it}, \tilde{X}_{it}')'$ and $m(\mathbf{X}_{it}) = m_1(X_{1,it}) + \tilde{m}(\tilde{X}_{it})$, where $\tilde{X}_{it} = (X_{2,it}, \ldots, X_{K,it})'$ and $\tilde{m}(\tilde{X}_{it}) = \sum_{k=2}^{K} m_k(X_{k,it})$.

https://doi.org/10.1017/S0266466621000554 Published online by Cambridge University Press

We first discuss the main idea of estimation and provide some heuristic arguments. To facilitate our discussion, we introduce several key assumptions. More regularity conditions will be provided in Section 3.

Assumption 1. $\{\mathbf{X}_{i1}, \mathbf{X}_{i2}, Y_{i1}, Y_{i2}, i = 1, 2, \ldots, n\}$ is a random sample generated from (2.1) with the support of $\mathcal{X} \times \mathcal{X} \times \mathcal{Y} \times \mathcal{Y}$, where $\mathcal{X} \times \mathcal{X}$ is compact. The support of $(\mathbf{X}_{i2} - \mathbf{X}_{i1})$ is compact, which includes 0 as an interior point.

Assumption 2. (i) $(\varepsilon_{i1}, \varepsilon_{i2})$ is independent of $(\mathbf{X}_{i1}, \mathbf{X}_{i2})$; (ii) ε_{i1} and ε_{i2} are exchangeable in the sense that $\{\varepsilon_{i1}, \varepsilon_{i2}\}$ and $\{\varepsilon_{i2}, \varepsilon_{i1}\}$ have the same joint distribution, and their density function is positive almost everywhere.

Assumption 3. The transformation function $\Lambda(\cdot)$ is strictly increasing.

Assumption 1 describes the panel data generating process (DGP). Assumption 1 assumes that our data $\{Y_{it}, \mathbf{X}_{it}\}, i = 1, \ldots, n, t = 1$ and 2, are an i.i.d. sample over the cross-sectional dimension i. For convenience, the support of $(\mathbf{X}_{i1}, \mathbf{X}_{i2})$ is assumed to be bounded.[3] We assume that $\mathbf{X}_{i1}$ and $\mathbf{X}_{i2}$ have the same support; therefore; they cannot be trending over time. The support of $(\mathbf{X}_{i2} - \mathbf{X}_{i1})$ is also assumed to be compact with 0 as an interior point to rule out time-invariant regressors. Otherwise, the time-invariant regressors would be absorbed into the fixed-effects α_i. Assumption 2 is a key assumption for our estimation procedure. Assumption 2(i) requires that the error term $(\varepsilon_{i1}, \varepsilon_{i2})$ be independent of the covariates $(\mathbf{X}_{i1}, \mathbf{X}_{i2})$. The full independence is often required for nonlinear models. Note that we allow arbitrary correlation between $(\mathbf{X}_{i1}, \mathbf{X}_{i2})$ and fixed-effects α_i. The independence assumption is commonly imposed in short panel data transformation models (see, e.g., Abrevaya, 1999; Chen, 2010).[4] Assumption 2(ii) imposes an exchangeability assumption on $(\varepsilon_{i1}, \varepsilon_{i2})$. This is a reasonable assumption for a short panel. Note that a more general condition, like strict stationarity, might be useful in cases where the sample size T is large or indefinite, and in dynamic settings. Assumption 3 assumes that the unknown transformation function $\Lambda(\cdot)$ is strictly monotone, which is standard in the literature (see, e.g., Cavanagh and Sherman, 1998; Abrevaya, 1999; Chen, 2002).

To estimate transformation models, we often need to impose certain scale and location normalizations. In our model, note that equation (1.1) holds if we replace $\Lambda(y)$, $\sum_{k=1}^{K} m_k(x_k)$, and ε_{it} with $\gamma\left(\Lambda(y) + \delta_0\right)$, $\gamma\left(\sum_{k=1}^{K}\left(m_k(x_k) + \delta_k\right)\right)$, and $\gamma\left(\varepsilon_{it} + \delta_0 - \sum_{k=1}^{K}\delta_k\right)$, respectively, for any constants $\gamma, \delta_0, \delta_1, \ldots,$ and δ_K. Therefore, we need to impose one scale and $K+1$ location normalizations. For convenience, we impose Assumption 4. Define $\Delta_i = \varepsilon_{i1} - \varepsilon_{i2}$, and let $f_\Delta(\cdot)$ be the probability density function (pdf) of Δ_i.

[3] This can be weakened to allow for unbounded support by using a trimming function in our estimators below, as in Härdle and Stoker (1989) and Powell, Stock, and Stoker (1989).

[4] Here, we assume that $\varepsilon_i \perp \mathbf{X}_i$, where $\varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2}), \mathbf{X}_i = (\mathbf{X}_{i1}, \mathbf{X}_{i2})$, and $\perp$ is denoted as independence. Abrevaya (1999) and Chen (2010) impose the conditional independence assumption that $\varepsilon_i \perp \mathbf{X}_i \mid \alpha_i$. Neither $\varepsilon_i \perp \mathbf{X}_i$ nor $\varepsilon_i \perp \mathbf{X}_i \mid \alpha_i$ implies the other. For our model, we find that $\varepsilon_i \perp \mathbf{X}_i$ is more appropriate.

https://doi.org/10.1017/S0266466621000554 Published online by Cambridge University Press

Assumption 4. The scale normalization is that $f_\Delta(0) = 1$. Let $\left(x_{1,0}, x_{2,0}, \ldots, x_{k,0}\right)$ and y_0 be interior points of $\mathcal{X}$ and $\mathcal{Y}$, respectively. The $K+1$ location normalizations are that $\Lambda(y_0) = 0$, and $m_k\left(x_{k,0}\right) = 0$, for $k = 1, \ldots, K$.

We compare the two dependent variables and define $d_i = \mathbf{1}\{Y_{i2} > Y_{i1}\}$, where $\mathbf{1}\{\cdot\}$ is the indicator function. The strict monotonicity of $\Lambda(y)$ implies that

$$\begin{aligned} d_i &= \mathbf{1}\{\Lambda(Y_{i2}) > \Lambda(Y_{i1})\} \\ &= \mathbf{1}\{m(\mathbf{X}_{i2}) + \alpha_i + \varepsilon_{i2} > m(\mathbf{X}_{i1}) + \alpha_i + \varepsilon_{i1}\} \\ &= \mathbf{1}\{\Delta_i < m(\mathbf{X}_{i2}) - m(\mathbf{X}_{i1})\}. \end{aligned}$$

The advantage of considering d_i is that the fixed-effects α_i are removed by using the strict monotonicity property of the transformation function. Our estimator utilizes the information of the symmetric distribution of Δ_i, which is shown in the lemma below.

LEMMA 2.1. *If ε_{i1} and ε_{i2} are exchangeable, then the distribution of Δ_i is symmetric around zero, and $F_\Delta(0) = 1/2$, where $F_\Delta(\cdot)$ is the cumulative distribution function (cdf) of Δ_i.*

For a fixed evaluation value $x_1 \in \mathcal{X}_1$, let $m_1'(x_1)$ be the first-order derivative of $m_1(x_1)$. Consider a neighborhood where $X_{1,i2} \approx x_1$, $X_{1,i1} \approx x_1$, and $\tilde{X}_{i2} - \tilde{X}_{i1} \approx 0$. Note that $\tilde{X}_{i2} - \tilde{X}_{i1} \approx 0$ suggests that we need to match $\tilde{X}_{i1}$ and $\tilde{X}_{i2}$. A Taylor expansion gives

$$\begin{aligned} &E(d_i|\mathbf{X}_{i1}, \mathbf{X}_{i2}) = \Pr(\Delta_i < m(\mathbf{X}_{i2}) - m(\mathbf{X}_{i1})|\mathbf{X}_{i1}, \mathbf{X}_{i2}) \\ &= F_\Delta(m_1(X_{1,i2}) - m_1(X_{1,i1}) + \tilde{m}(\tilde{X}_{i2}) - \tilde{m}(\tilde{X}_{i1})) \\ &\approx F_\Delta(0) + f_\Delta(0) m_1'(x_1)(X_{1,i2} - X_{1,i1}) + f_\Delta(0) \cdot \left[\tilde{m}(\tilde{X}_{i2}) - \tilde{m}(\tilde{X}_{i1})\right] \\ &\approx F_\Delta(0) + f_\Delta(0) m_1'(x_1)(X_{1,i2} - X_{1,i1}) \\ &= \frac{1}{2} + m_1'(x_1)(X_{1,i2} - X_{1,i1}). \end{aligned} \tag{2.2}$$

The first equality holds due to the independence assumption between $(\varepsilon_{i1}, \varepsilon_{i2})$ and $(\mathbf{X}_{i1}, \mathbf{X}_{i2})$. The second equality is by the definition of F_Δ. The third approximation applies the Taylor expansion to the neighborhood where $X_{1,i2}$ and $X_{1,i1}$ are close to x_1 and $\tilde{X}_{i2} - \tilde{X}_{i1}$ is close to 0. The bias term due to the last term in the third line above can be shown to be asymptotically negligible under our regularity conditions, which yields the fourth approximation. The last equality uses the fact that $F_\Delta(0) = \frac{1}{2}$ and $f_\Delta(0) = 1$. Equation (2.2) serves as the basis for the estimation of $m_1'(x_1)$. So far, we have argued that the generalized panel data transformation model (2.1) implies a locally linear relation,

$$d_i = \frac{1}{2} + m_1'(x_1)(X_{1,i2} - X_{1,i1}) + \eta_i, \tag{2.3}$$

https://doi.org/10.1017/S0266466621000554 Published online by Cambridge University Press

where

$$\eta_i = [d_i - E(d_i|\mathbf{X}_{i1}, \mathbf{X}_{i2})] + \left[E(d_i|\mathbf{X}_{i1}, \mathbf{X}_{i2}) - \frac{1}{2} - m'(x_1)\left(X_{1,i2} - X_{1,i1}\right)\right].$$

Here, η_i signifies the prediction error $d_i - E(d_i|\mathbf{X}_{i1}, \mathbf{X}_{i2})$ and the approximation error. Equation (2.3) suggests that we can run a regression of $(d_i - \frac{1}{2})$ on $(X_{1,i2} - X_{1,i1})$ locally to estimate $m_1'(x_1)$. Specifically, we minimize the following weighted least-squares (LS) criterion function:

$$\min_b \sum_{i=1}^n \left\{\left[\left(d_i - \frac{1}{2}\right) - b \cdot (X_{1,i2} - X_{1,i1})\right]^2 k_1\left(\frac{X_{1,i2} - x_1}{h_1}\right) k_1\left(\frac{X_{1,i1} - x_1}{h_1}\right) \tilde{k}\left(\frac{\tilde{X}_{i2} - \tilde{X}_{i1}}{\tilde{h}}\right)\right\}, \tag{2.4}$$

where $k_1(\cdot)$ and $\tilde{k}(\cdot)$ are the kernel functions. Specifically, $k_1(\cdot)$ is a univariate kernel function, and h_1 is its corresponding bandwidth parameter. With a slight abuse of notation, we let $\tilde{k}(\cdot)$ be the $(K-1) \times 1$ product kernel function with $\tilde{h}$ as its bandwidth parameter. That is, $\tilde{k}\left(\frac{\tilde{X}_{i2} - \tilde{X}_{i1}}{\tilde{h}}\right) = \prod_{\ell=2}^K k_\ell\left(\frac{X_{\ell,i2} - X_{\ell,i1}}{\tilde{h}}\right)$, where $k_\ell(\cdot)$ is a univariate kernel function.[5] The solution to the above minimization problem is the estimator of $m_1'(x_1)$, which is denoted as $\hat{m}_1'(x_1)$:

$$\hat{m}_1'(x_1) = \frac{\sum_{i=1}^n (d_i - \frac{1}{2})(X_{1,i2} - X_{1,i1}) k_1\left(\frac{X_{1,i2} - x_1}{h_1}\right) k_1\left(\frac{X_{1,i1} - x_1}{h_1}\right) \tilde{k}\left(\frac{\tilde{X}_{i2} - \tilde{X}_{i1}}{\tilde{h}}\right)}{\sum_{i=1}^n (X_{1,i2} - X_{1,i1})^2 k_1\left(\frac{X_{1,i2} - x_1}{h_1}\right) k_1\left(\frac{X_{1,i1} - x_1}{h_1}\right) \tilde{k}\left(\frac{\tilde{X}_{i2} - \tilde{X}_{i1}}{\tilde{h}}\right)}. \tag{2.5}$$

To estimate the original structural function $m_1(\cdot)$, we can take an integral of $\hat{m}_1'(\cdot)$. Note that we make the normalization that $m_1(x_{1,0}) = 0$ for a fixed value $x_{1,0}$, as stated in Assumption 4. Subsequently, the estimator of $m_1(x_1)$ is simply

$$\hat{m}_1(x_1) = \int_{x_{1,0}}^{x_1} \hat{m}_1'(\xi_1) d\xi_1. \tag{2.6}$$

We can estimate $m_k(\cdot)$ analogously for $k = 2, \ldots, K$.

Remark 2.1. As for most estimators with nonparametric denominators, trimming is imposed for the estimator to be well-behaved. We omit it for simplicity.

Remark 2.2. Our estimator can be easily extended to the case of $T > 2$. Assume that $(\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{iT})$ are independent of $(\mathbf{X}_{i1}, \mathbf{X}_{i2}, \ldots, \mathbf{X}_{iT})$, and exchangeable. Along the same line as equation (2.5), the nonparametric estimator of

[5] For notation simplicity, we use the same bandwidth $\tilde{h}$ for different covariates. They can be different. In actual applications, we should at least let the bandwidths be proportional to the sample standard deviations of their corresponding covariates.

https://doi.org/10.1017/S0266466621000554 Published online by Cambridge University Press

$m_1'(x_1)$ is

$$\hat{m}_1'(x_1) = \frac{\sum_{s<t}\sum_{i=1}^n (d_{its}-\frac{1}{2})(X_{1,it}-X_{1,is})k_1\left(\frac{X_{1,it}-x_1}{h_1}\right)k_1\left(\frac{X_{1,is}-x_1}{h_1}\right)\tilde{k}\left(\frac{\tilde{X}_{it}-\tilde{X}_{is}}{\tilde{h}}\right)}{\sum_{s<t}\sum_{i=1}^n (X_{1,it}-X_{1,is})^2 k_1\left(\frac{X_{1,it}-x_1}{h_1}\right)k_1\left(\frac{X_{1,is}-x_1}{h_1}\right)\tilde{k}\left(\frac{\tilde{X}_{it}-\tilde{X}_{is}}{\tilde{h}}\right)},$$

for $s,t \in \{1,2,\ldots,T\}$, where $d_{its} = \mathbf{1}\{Y_{it} > Y_{is}\}$. $\hat{m}_1(x_1)$ can be estimated in the same way as in equation (2.6).

Remark 2.3. For ease of notation, equation (2.2) only takes a local linear approximation. We can also approximate $E(d_i|\mathbf{X}_{i1},\mathbf{X}_{i2})$ using a higher-order series (e.g., vth-order power series) locally if $m_1(\cdot)$ is sufficiently smooth. Specifically, in the neighborhood where $X_{1,i2} \approx x_1$, $X_{1,i1} \approx x_1$, and $\tilde{X}_{i2} - \tilde{X}_{i1} \approx 0$,

$$E(d_i|\mathbf{X}_{i1},\mathbf{X}_{i2}) \approx \frac{1}{2} + m_1'(x_1)(X_{1,i2}-X_{1,i1}) + \cdots + \frac{1}{v!}m_1^{(v)}(x_1)(X_{1,i2}-X_{1,i1})^v.$$

Then, $m_1'(x_1)$ can be estimated by the first element of $\hat{b}$, where

$$\hat{b} = \arg\min_b \; Q_n(b),$$

with

$$Q_n(b) = \sum_{i=1}^n \left[d_i - \frac{1}{2} - \sum_{j=1}^v b_j (X_{1,i2}-X_{1,i1})^j \right]^2 k_1\left(\frac{X_{1,i2}-x_1}{h_1}\right)k_1\left(\frac{X_{1,i1}-x_1}{h_1}\right)\tilde{k}\left(\frac{\tilde{X}_{i2}-\tilde{X}_{i1}}{\tilde{h}}\right),$$

and $b = (b_1,\ldots,b_v)$.

Remark 2.4. For efficiency consideration, our estimators use the information of the symmetric distribution of Δ_i and impose $F_\Delta(0) = \frac{1}{2}$ for the intercept. The symmetry is implied by the exchangeability assumption of $(\varepsilon_{i1},\varepsilon_{i2})$. However, this exchangeability assumption can be relaxed. If $(\varepsilon_{i1},\varepsilon_{i2})$ is not exchangeable, the distribution of Δ_i may not be symmetric. Therefore, $F_\Delta(0)$ is not necessarily equal to $\frac{1}{2}$. In this instance, we can replace $\frac{1}{2}$ in (2.4) with an additional parameter to be estimated. In particular, we can allow time-fixed effects in our model. Specifically, for the second period, we have $\Lambda(Y_{i2}) = \sum_{k=1}^K m_k(X_{k,i2}) + \alpha_i + (\lambda + \varepsilon_{i2})$, where λ is the time-fixed effect. In this case, $\Delta_i = \varepsilon_{i1} - (\lambda + \varepsilon_{i2})$ does not have symmetric distribution in general and $F_\Delta(0) \neq \frac{1}{2}$; hence, $F_\Delta(0)$ has to be treated as an unknown parameter.

Remark 2.5. Our estimators in equations (2.5) and (2.6) and the asymptotic theory below do not rely on the additive structure of $\tilde{m}(\cdot)$. Therefore, our estimators

https://doi.org/10.1017/S0266466621000554 Published online by Cambridge University Press

actually apply to a slightly more general model:

$$\Lambda(Y_{it}) = m_1(X_{1,it}) + \tilde{m}(\tilde{X}_{it}) + \alpha_i + \varepsilon_{it},\ i = 1,\ldots,n \text{ and } t = 1,\ldots,T,$$

where $\tilde{m}(\cdot)$ is a general function: $\mathbb{R}^{K-1} \to \mathbb{R}$.

3. LARGE SAMPLE PROPERTIES

In this section, we study the large sample properties of the estimators of $m_1'(x_1)$ and $m_1(x_1)$. We first make some assumptions. Let $\mathcal{W}$ be the support of Δ_i. The pdf of Δ_i at w is denoted as $f_\Delta(w)$, for $w \in \mathcal{W}$. Let $p_0(x)$ be the joint density of $\mathbf{X}_{it}$ at x, where $x = (x_1,\ldots,x_K)$.[6] Let x_{10} and x_{20} be two finite interior points of the support of $X_{1,it}$ such that $x_{10} < x_{20}$.

Assumption 5. (i) $m_1(x_1)$ is three times continuously differentiable, and $\tilde{m}(\tilde{x})$ is continuously differentiable up to sth-order. (ii) $\mathbf{X}_{it}$ has a nondegenerate joint density $p_0(\cdot)$ with support $\mathcal{X}$, which is a nondegenerate subset of $\mathbb{R}^K$. $p_0(x)$ is third-order continuously differentiable with respect to x_1 and sth-order continuously differentiable with respect to $\tilde{x}$, namely, there exists a constant M such that $\left|\frac{\partial^s p_0(x)}{\partial x_2^{s_2}\cdots\partial x_K^{s_K}}\right| \le M$, where $s = s_2 + \cdots + s_K$. (iii) $f_\Delta(w)$ is $(s+2)$ times continuously differentiable. (iv) $Q(x_1) = 2K_2\int p_0^2(x_1,\tilde{x}_1)\,d\tilde{x}_1$ with $K_2 = \int k_1(u)\,u^2 du$ is uniformly bounded away from zero on $[x_{10}, x_{20}]$.

Assumption 6. (i) The kernel function $k_1(u)$ is a symmetric second-order kernel function of bounded variation and bounded support such that

$$\int k_1(u)du = 1,\ \int uk_1(u)du = 0,\ \text{and}\ \int u^2 k_1(u)du \neq 0.$$

(ii) The kernel function $\tilde{k}(u)$ is a product kernel with each component as an sth-order kernel of bounded variation. That is, for each element of $\tilde{k}$, $k_\ell(u)$ satisfies

$$\int k_\ell(u)du = 1,\ \int u^p k_\ell(u)du = 0,\ \text{if } p = 1,\ldots,s-1,\ \text{and}\ \int u^s k_\ell(u)du \neq 0.$$

Let $\delta_{sn} = \left(\frac{\ln n}{nh_1^4\tilde{h}^{K-1}}\right)^{1/2} + h_1^2 + \tilde{h}^s$ and $\delta_{qn} = \left(\frac{\ln n}{nh_1^2\tilde{h}^{K-1}}\right)^{1/2} + h_1^2 + \tilde{h}^s$.

Assumption 7A. The bandwidths $h_1 \to 0$ and $\tilde{h} \to 0$ satisfy

$$nh_1^4\tilde{h}^{K-1} \to \infty,\ \sqrt{nh_1^4\tilde{h}^{K-1}}\,\delta_{sn}\delta_{qn} = o(1),\ \sqrt{nh_1^4\tilde{h}^{K-1}}\,\tilde{h}^s = o(1),$$

and $\sqrt{nh_1^4\tilde{h}^{K-1}}\,h_1^2 \to \lambda_1 < \infty$.

[6] Here, we assume $\mathbf{X}_{i1}$ and $\mathbf{X}_{i2}$ have the same joint distribution for ease of notation. In general, their distributions can be different with densities $p_1(x)$ and $p_2(x)$.

https://doi.org/10.1017/S0266466621000554 Published online by Cambridge University Press

Assumption 7B. The bandwidths $h_1 \to 0$ and $\tilde{h} \to 0$ satisfy

$$nh_1^3\tilde{h}^{K-1} \to \infty,\ \sqrt{nh_1^3\tilde{h}^{K-1}}\delta_{sn}\delta_{qn} = o(1),\ \sqrt{nh_1^3\tilde{h}^{K-1}}\tilde{h}^s = o(1),$$

and $\sqrt{nh_1^3\tilde{h}^{K-1}}h_1^2 \to \lambda_2 < \infty$.

Assumption 5 contains certain smoothness conditions for the structural functions, the joint density of $\mathbf{X}_{it}$, and the density of Δ_i. In particular, in part (ii), we require that the joint density of $\mathbf{X}_{it}$ be nondegenerate, which rules out interaction terms, e.g., $X_{1,it} = W_{1,it} \cdot W_{2,it}, \tilde{X}_{it} = \left(W_{1,it},\ W_{2,it}\right)$ for some base covariates $W_{1,it}$, and $W_{2,it}$. In Assumption 6, the kernel function $k_1(\cdot)$ can be a normal density function, but a higher-order kernel for $\tilde{k}(\cdot)$ is required when K is large in order to remove the influence brought by matching $\tilde{X}_{it}$ on the asymptotic bias. Assumptions 7A and 7B specify the conditions for bandwidths $\left(h_1, \tilde{h}\right)$, which are imposed for the estimators of $m_1'(x_1)$ and $m_1(x_1)$, respectively. They can be satisfied, for example, if $h_1 = c_1 n^{-1/5}$ and $\tilde{h} = \tilde{c}n^{-\frac{1}{2(K-1)+2s}}$ with $s = 2(K-1)$, where c_1 and $\tilde{c}$ are some constants.

The following two theorems establish the large sample properties of the estimators in (2.5) and (2.6), respectively.

Theorem 3.1. *Let Assumptions 1–6 and 7A hold. Then,*

$$\hat{m}_1'(x_1) - m_1'(x_1) = h_1^2 b_1(x_1) + o(h_1^2) + O\left(\tilde{h}^s\right) + O_p\left(\frac{\ln n}{nh_1^4\tilde{h}^{K-1}}\right)^{1/2},$$

uniformly in $x_1 \in [x_{10}, x_{20}]$; in addition, for any interior point $x_1 \in \mathcal{X}_1, \hat{m}_1'(x_1)$ is consistent and asymptotically normal,

$$\sqrt{nh_1^4\tilde{h}^{K-1}}(\hat{m}_1'(x_1) - m_1'(x_1) - h_1^2 b_1(x_1)) \xrightarrow{d} N(0, \sigma_1^2(x_1)), \tag{3.1}$$

where $b_1(x_1)$ and $\sigma_1^2(x_1)$ are defined in the Appendix.

Theorem 3.2. *Let Assumptions 1–6 and 7B hold. Then,*

$$\hat{m}_1(x_1) - m_1(x_1) = h_1^2 b_2(x_1) + o(h_1^2) + O\left(\tilde{h}^s\right) + O_p\left(\frac{\ln n}{nh_1^3\tilde{h}^{K-1}}\right)^{1/2},$$

uniformly in $x_1 \in [x_{10}, x_{20}]$; in addition, for any interior point $x_1 \in \mathcal{X}_1, \hat{m}_1(x_1)$ is consistent and asymptotically normal,

$$\sqrt{nh_1^3\tilde{h}^{K-1}}(\hat{m}_1(x_1) - m_1(x_1) - h_1^2 b_2(x_1)) \xrightarrow{d} N(0, \sigma_2^2(x_1)), \tag{3.2}$$

where $b_2(x_1)$ and $\sigma_2^2(x_1)$ are defined in the Appendix.

https://doi.org/10.1017/S0266466621000554 Published online by Cambridge University Press

The formal proofs are given in the Appendix. Theorems 3.1 and 3.2 establish the consistency and asymptotic normality of our estimators of $m_1'(x_1)$ and $m_1(x_1)$ for any fixed evaluation point x_1. The convergence rates of $\hat{m}_1'(x_1)$ and $\hat{m}_1(x_1)$ depend on $\tilde{h}$ and $K-1$. Therefore, our estimators still suffer from the "curse of dimensionality." The intuition is that to estimate $m_1'(x_1)$ and $m_1(x_1)$, we need to match other covariates $\tilde{X}_{i,1}$ and $\tilde{X}_{i,2}$, and the influence of matching cannot be completely removed. Nevertheless, the additive assumption in (2.1) ameliorates the curse of dimensionality to some degree. As discussed below, if we do not impose the additive structure, the convergence rate of the estimator of the structural function will be slower.

For the purpose of conducting statistical inference, we can make the bias terms negligible asymptotically by undersmoothing, i.e., choosing a small bandwidth h_1 with a suboptimal rate. For the asymptotic variance terms, we could estimate them by replacing the various elements with their sample analogs.

For completeness, one may be interested in estimating $\Lambda(\cdot)$. Consider the case with $T=2$. For a fixed evaluation point y, define $d_{yi2} = \mathbf{1}\{Y_{i2} < y\}$ and $d_{y_0 i1} = \mathbf{1}\{Y_{i1} < y_0\}$. For simplicity, assume that $(\varepsilon_{i1}, \varepsilon_{i2})$ is independent of $(\mathbf{X}_{i1}, \mathbf{X}_{i2}, \alpha_i)$, and let F denote the distribution function of ε_{i1} (and also ε_{i2}). Then, following Chen (2002), we can show that

$$E(d_{yi2} - d_{y_0 i1} | \mathbf{X}_{i1}, \mathbf{X}_{i2}, \alpha_i) = F(\Lambda(y) - m(\mathbf{X}_{i2}) - \alpha_i) - F(-m(\mathbf{X}_{i1}) - \alpha_i) > 0,$$

if and only if

$$\Lambda(y) > m(\mathbf{X}_{i2}) - m(\mathbf{X}_{i1}),$$

where we have imposed the location normalization $\Lambda(y_0) = 0$. Denote $\hat{m}(x) = \sum_{k=1}^{K} \hat{m}_k(x_k)$. Consequently, given $\hat{m}(\mathbf{X}_{i1})$ and $\hat{m}(\mathbf{X}_{i2})$, for $i = 1, 2, \ldots, n$, $\Lambda(y)$ can be estimated by $\hat{\Lambda}(y)$, which solves

$$\max_{b \in \mathcal{L}} \Gamma_n(b; y, \hat{m}),$$

where $\mathcal{L}$ is a compact set, and

$$\Gamma_n(b; y, \hat{m}) = \frac{1}{n} \sum_{i=1}^{n} (d_{yi2} - d_{y_0 i1}) \cdot \mathbf{1}\{\big(b > \hat{m}(\mathbf{X}_{i2}) - \hat{m}(\mathbf{X}_{i1})\big)\}.$$

Due to technical and practical complications associated with the step function in the above objective function, we work with a smoothed version

$$\Gamma_{ns}(b; y, \hat{m}) = \frac{1}{n} \sum_{i=1}^{n} (d_{yi2} - d_{y_0 i1}) K_y \left(\frac{\hat{m}(\mathbf{X}_{i2}) - \hat{m}(\mathbf{X}_{i1}) - b}{h_y} \right),$$

where $K_y(t) = \int_{-\infty}^{t} k_y(u) du$, k_y is a kernel density function on $[-1, 1]$ such that $K_y(t) = 0$ when $t \le -1$ and $K_y(t) = 1$ when $t \ge 1$, and h_y is the corresponding bandwidth parameter. For more details on how to construct a smooth function to replace a step function, see, e.g., Kaplan and Sun (Kaplan and Sun (2017, p. 113).

https://doi.org/10.1017/S0266466621000554 Published online by Cambridge University Press

Here, $m(x)$ can be consistently estimated uniformly with $\hat{m}(x) - m(x) = O_p(\delta_n)$, where $\delta_n = h_1^2 + \tilde{h}^s + \left(nh_1^3\tilde{h}^{K-1}/\ln n\right)^{-1/2}$. The following theorem states that under some regularity conditions, $\hat{\Lambda}(\cdot)$ is uniformly consistent for $\Lambda(\cdot)$. Let y_1 and y_2 be two finite interior points of the support of Y_{it} such that $y_1 < y_2$.

Theorem 3.3. *Let Assumptions 1–6 and 7B hold. In addition, assume that $k_y(\cdot)$ is symmetric with $\int k_y(u)du = 1$ and $\delta_n/h_y = o(1)$, then*

$$\sup_{y\in[y_1,y_2]} \left|\hat{\Lambda}(y) - \Lambda(y)\right| = o_p(1).$$

Note that here we only provide the consistency result. Ascertaining the right rate of convergence and limiting distribution involves tedious and delicate arguments, which are not considered here.

Remark 3.1. Compared with the conventional additive model in cross-sectional data, the convergence rates of our estimators are slower. This is due to the panel structure for which there is no averaging aspect over the time dimension to speed up the rate of convergence. Nevertheless, to the best of our knowledge, there does not exist a consistent estimator for the nonparametric regression model subject to unknown transformation in a panel data setting. We can also consider the average effect, such as $E\left(m_1'(X_{1,it})\right)$, which can be estimated by the sample analog: $\frac{1}{n}\sum_{i=1}^n \hat{m}_1'\left(X_{1,it}\right)$. Following the analysis of $\hat{m}_1(x_1)$, we can also show that it converges to $E\left(m_1'(X_{1,it})\right)$ at the rate of $\sqrt{nh_1^3\tilde{h}^{K-1}}$.

Remark 3.2. As suggested by the Co-Editor, we could consider a multistage estimation procedure and use our estimators as initial estimators, which can potentially achieve a faster convergence rate. Specifically, let $\widehat{\tilde{m}}(\cdot)$ and $\hat{\Lambda}(\cdot)$ be the estimators of $\tilde{m}(\cdot)$ and $\Lambda(\cdot)$, respectively, as described above. Then, using them as initial estimators, we could estimate $m_1(\cdot)$ in the following model:

$$\hat{\Lambda}(Y_{it}) - \widehat{\tilde{m}}(\tilde{X}_{it}) = m_1(X_{1,it}) + \alpha_i + \varepsilon_{it} + \zeta_{it}, \tag{3.3}$$

where $\zeta_{it} = \hat{\Lambda}(Y_{it}) - \Lambda(Y_{it}) - \widehat{\tilde{m}}(\tilde{X}_{it}) + \tilde{m}(\tilde{X}_{it})$ signifies the estimation error. Model (3.3) is a standard nonparametric panel model except that we have to take into account the estimation errors of $\hat{\Lambda}(\cdot)$ and $\widehat{\tilde{m}}(\cdot)$. Model (3.3) can be estimated using existing methods, e.g., the kernel estimator proposed in Su and Lu (2013). Multistage estimators are often used in additive models (see, e.g., Horowitz and Mammen, 2004; Ozabaci, Henderson, and Su, 2014). Although it is practical to implement the estimation procedure as (3.3) for our model, it is challenging to study its theoretical properties. The difficulty is to establish the uniform consistency of $\widehat{\tilde{m}}(\cdot)$ and particularly $\hat{\Lambda}(\cdot)$ for their whole support with sufficiently fast convergence rates. For example, Chen's Chen's (2002) rank estimator of $\Lambda(\cdot)$ is only uniformly consistent over a compact set, and thus we can only estimate $\Lambda(\cdot)$ consistently over a finite interval when the support of Y is the real line. Therefore, we leave a detailed study on this multistage estimator for future research.

https://doi.org/10.1017/S0266466621000554 Published online by Cambridge University Press

4. SIMULATIONS

We conduct Monte Carlo simulations to examine the finite-sample performance of the proposed estimator. To save space, we only report the detailed results for the estimator of $m_1(\cdot)$.[7] We consider six DGPs.

DGP I: $\Lambda(Y_{it}) = X_{1,it}^2 + \alpha_i + \varepsilon_{it}$, where $\varepsilon_{it} \sim U(0,1)$;

DGP II: $\Lambda(Y_{it}) = X_{1,it}^2 + \alpha_i + \varepsilon_{it}$, where $(a\varepsilon_{it} + b) \sim \mathcal{X}^2(2)$ with $a = \frac{1}{2}\left(\frac{9}{8}\right)^3$ and $b = \frac{1}{2}\exp\left(-\frac{1}{2a}\right)$;

DGP III: $\Lambda(Y_{it}) = X_{1,it}^3 + 0.5X_{1,it}^2 + \alpha_i + \varepsilon_{it}$, where $\varepsilon_{it} \sim U(0,1)$;

DGP IV: $\Lambda(Y_{it}) = X_{1,it}^2 + X_{2,it}^2 + \alpha_i + \varepsilon_{it}$, where $\varepsilon_{it} \sim U(0,1)$;

DGP V: $\Lambda(Y_{it}) = X_{1,it}^2 + X_{2,it}^2 + \alpha_i + \varepsilon_{it}$, where $(a\varepsilon_{it} + b) \sim \mathcal{X}^2(2)$ with $a = \frac{1}{2}\left(\frac{9}{8}\right)^3$ and $b = \frac{1}{2}\exp\left(-\frac{1}{2a}\right)$;

DGP VI: $\Lambda(Y_{it}) = X_{1,it}^3 + 0.5X_{1,it}^2 + X_{2,it}^2 + \alpha_i + \varepsilon_{it}$, where $\varepsilon_{it} \sim U(0,1)$.

In all DGPs, we take the Box–Cox transformation of Bickel and Doksum (1981) with $\Lambda(y) = \frac{|y|^\lambda sgn(y) - 1}{\lambda}$ for $\lambda = 0.8$. In DGPs I–III, $X_{1,it}$ is drawn from the uniform distribution $U(-1,1)$. In DGPs IV–VI, both $X_{1,it}$ and $X_{2,it}$ follow $U(-1,1)$, and they are correlated with a correlation coefficient of $\rho \approx 0.2$. $\alpha_i = 0.5(X_{1,i1} + X_{1,i2}) + 0.5\eta_i$, where η_i is an $N(0,1)$ random variable. The error term follows either a uniform distribution or a chi-square distribution of freedom 2. The former is symmetric, but the latter is asymmetric. Note that we have made sure that the normalization condition $f_\Delta(0) = 1$ is satisfied for all DGPs. For DGPs I–III, the number of covariates is 1 ($K = 1$); therefore, there is only one structural function, $m_1(x) = x^2$ or $m_1(x) = x^3 + 0.5x^2$. For DGPs IV–VI, $K = 2$, and there are two structural functions which are additive. Specifically, $m(x) = m_1(x_1) + m_2(x_2)$, where $m_1(x_1) = x_1^2$ and $m_2(x_2) = x_2^2$ for DGPs IV and V, and $m_1(x_1) = x_1^3 + 0.5x_1^2$ and $m_2(x_2) = x_2^2$ for DGP VI. We use the rectangular method for integration. The kernel function $k_1(\cdot)$ is a standard normal density. The baseline bandwidth for h_1 is chosen according to Silverman's rule of thumb: $h_1^* = 1.06 \cdot \hat{\sigma}_1 \cdot n^{-1/5}$, where $\hat{\sigma}_1$ is the sample standard deviation of $X_{1,it}$. To examine the robustness of our estimator to the choice of bandwidth, we consider $h_1 = c_1 \times h_1^*$, where c_1 takes three different values from $\{1, 1.5, 2\}$. With regard to $\tilde{k}(\cdot)$, we use the second-order Gaussian kernel, with the bandwidth as $\tilde{h} = \hat{\sigma}_2 \cdot n^{-1/6}$ for simplicity, where $\hat{\sigma}_2$ is the sample standard deviation of $X_{2,i1} - X_{2,i2}$. We can verify that the choice of $(k_1(\cdot), \tilde{k}(\cdot))$, and $(h_1, \tilde{h})$ satisfies Assumptions 6, 7A and 7B.[8] The number of replications for the simulations is 1,000. We consider two sample sizes: $n = 500$ and 2,000.

Tables 1–6 report bias (Bias), standard deviation (SD), and root-mean-square error (RMSE) of $\hat{m}_1(x_1)$ for DGPs I–VI, respectively. We choose eight evaluation

[7] The results for the estimator of its derivative $m_1'(\cdot)$ are available upon request.

[8] We also consider a higher-order kernel for $\tilde{k}(\cdot)$ in our simulations. We find that the higher-order kernels can result in greater variance and be inferior to lower-order kernels. This phenomenon is well documented in the nonparametric literature (see, e.g., Lewbel, 2000).

https://doi.org/10.1017/S0266466621000554 Published online by Cambridge University Press

TABLE 1. Estimation results for DGP I

x_1		−0.8	−0.6	−0.4	−0.2	0.2	0.4	0.6	0.8
$m_1(x_1)$		0.64	0.36	0.16	0.04	0.04	0.16	0.36	0.64
	c_1				$n = 500$				
Bias		−0.0876	−0.0206	0.0106	0.0138	0.0155	0.0148	−0.0154	−0.0834
SD	1	0.1166	0.0963	0.0796	0.0554	0.0544	0.0783	0.0962	0.1188
RMSE		0.1458	0.0984	0.0802	0.0571	0.0566	0.0797	0.0973	0.1451
Bias		−0.1736	−0.0694	−0.0119	0.0068	0.0084	−0.0096	−0.0667	−0.1710
SD	1.5	0.0651	0.0507	0.0393	0.0259	0.0258	0.0389	0.0503	0.0649
RMSE		0.1854	0.0860	0.0411	0.0268	0.0271	0.0401	0.0836	0.1829
Bias		−0.2605	−0.1251	−0.0416	−0.0030	−0.0019	−0.0400	−0.1234	−0.2588
SD	2	0.0436	0.0332	0.0245	0.0154	0.0155	0.0247	0.0333	0.0434
RMSE		0.2641	0.1294	0.0482	0.0157	0.0156	0.0470	0.1278	0.2624
Bias		−0.1461	−0.0595	−0.0132	0.0039	0.0061	−0.0096	−0.0548	−0.1425
SD	CV	0.1067	0.0801	0.0604	0.0372	0.0359	0.0591	0.0808	0.1085
RMSE		0.1809	0.0997	0.0618	0.0374	0.0364	0.0598	0.0976	0.1791
c_1^{cv}		1.4978	1.4182	1.6142	1.8768	1.8704	1.5772	1.4250	1.5056
					$n = 2,000$				
Bias		−0.0466	0.0012	0.0182	0.0145	0.0186	0.0235	0.0053	−0.0428
SD	1	0.0869	0.0741	0.0633	0.0448	0.0466	0.0640	0.0774	0.0918
RMSE		0.0986	0.0741	0.0658	0.0471	0.0502	0.0682	0.0776	0.1012
Bias		−0.1101	−0.0320	0.0055	0.0117	0.0140	0.0087	−0.0291	−0.1077
SD	1.5	0.0463	0.0375	0.0309	0.0211	0.0221	0.0318	0.0393	0.0492
RMSE		0.1194	0.0493	0.0313	0.0241	0.0262	0.0329	0.0489	0.1184
Bias		−0.1750	−0.0700	−0.0121	0.0068	0.0081	−0.0103	−0.0682	−0.1736
SD	2	0.0302	0.0237	0.0186	0.0124	0.0127	0.0190	0.0245	0.0318
RMSE		0.1776	0.0739	0.0222	0.0142	0.0150	0.0216	0.0725	0.1765
Bias		−0.0940	−0.0298	0.0009	0.0087	0.0114	0.0045	−0.0272	−0.0913
SD	CV	0.0768	0.0595	0.0459	0.0288	0.0312	0.0467	0.0618	0.0797
RMSE		0.1213	0.0665	0.0459	0.0301	0.0332	0.0469	0.0675	0.1212
c_1^{cv}		1.4424	1.4558	1.7490	2.0408	2.0038	1.7394	1.4796	1.4340

Notes: c_1 refers to the constant for the bandwidth choice. $c_1^c v$ refers to the average value of the constant chosen by our CV method (equation (4.1)).

points of x_1: $(-0.8, -0.6, -0.4, -0.2, 0.2, 0.4, 0.6, 0.8)$. We summarize the main findings. First, in general, Bias, SD, and RMSE all decrease when the sample size increases from 500 to 2,000. For example, consider DGP I and the evaluation point $x_1 = -0.8$ in Table 1. The RMSE reduces from (0.1458, 0.1854, 0.2641) for $c_1 = (1, 1.5, 2)$ to (0.0986, 0.1194, 0.1776), respectively, when n increases from

https://doi.org/10.1017/S0266466621000554 Published online by Cambridge University Press

TABLE 2. Estimation results for DGP II

x_1		−0.8	−0.6	−0.4	−0.2	0.2	0.4	0.6	0.8
$m_1(x_1)$		0.64	0.36	0.16	0.04	0.04	0.16	0.36	0.64
	c_1				$n = 500$				
Bias		−0.1494	−0.0560	−0.0047	0.0103	0.0094	−0.0034	−0.0550	−0.1514
SD	1	0.1197	0.0969	0.0793	0.0555	0.0559	0.0804	0.0987	0.1233
RMSE		0.1914	0.1119	0.0794	0.0564	0.0566	0.0805	0.1129	0.1952
Bias		−0.2326	−0.1048	−0.0286	0.0024	0.0019	−0.0292	−0.1063	−0.2363
SD	1.5	0.0658	0.0497	0.0380	0.0251	0.0254	0.0386	0.0504	0.0673
RMSE		0.2417	0.1159	0.0475	0.0252	0.0255	0.0484	0.1176	0.2457
Bias		−0.3106	−0.1552	−0.0560	−0.0070	−0.0075	−0.0571	−0.1573	−0.3141
SD	2	0.0450	0.0335	0.0244	0.0153	0.0155	0.0249	0.0342	0.0462
RMSE		0.3138	0.1588	0.0611	0.0168	0.0172	0.0623	0.1610	0.3175
Bias		−0.2098	−0.0954	−0.0284	0.0008	−0.0005	−0.0284	−0.0957	−0.2125
SD	CV	0.1078	0.0805	0.0602	0.0381	0.0382	0.0622	0.0828	0.1115
RMSE		0.2359	0.1248	0.0666	0.0381	0.0382	0.0683	0.1265	0.2400
c_1^{cv}		1.6012	1.4946	1.6460	1.9044	1.8806	1.6030	1.5052	1.6210
					$n = 2{,}000$				
Bias		−0.1055	−0.0301	0.0083	0.0146	0.0128	0.0079	−0.0296	−0.1052
SD	1	0.0892	0.0765	0.0627	0.0446	0.0447	0.0635	0.0761	0.0892
RMSE		0.1381	0.0822	0.0633	0.0469	0.0464	0.0640	0.0817	0.1379
Bias		−0.1747	−0.0691	−0.0098	0.0090	0.0080	−0.0103	−0.0695	−0.1754
SD	1.5	0.0488	0.0389	0.0310	0.0211	0.0213	0.0313	0.0391	0.0491
RMSE		0.1814	0.0793	0.0325	0.0229	0.0227	0.0330	0.0797	0.1821
Bias		−0.2373	−0.1072	−0.0294	0.0023	0.0019	−0.0296	−0.1075	−0.2378
SD	2	0.0331	0.0252	0.0192	0.0125	0.0125	0.0192	0.0253	0.0332
RMSE		0.2396	0.1101	0.0351	0.0127	0.0127	0.0353	0.1104	0.2401
Bias		−0.1566	−0.0641	−0.0120	0.0067	0.0054	−0.0120	−0.0633	−0.1561
SD	CV	0.0775	0.0610	0.0458	0.0292	0.0289	0.0466	0.0615	0.0785
RMSE		0.1748	0.0885	0.0473	0.0299	0.0294	0.0481	0.0882	0.1747
c_1^{cv}		1.4914	1.4950	1.7184	1.9618	1.9690	1.7056	1.4974	1.4988

Notes: c_1 refers to the constant for the bandwidth choice. $c_1^c v$ refers to the average value of the constant chosen by our CV method (equation (4.1)).

500 to 2,000. Second, although the performance of our estimators depends on the choice of bandwidth (i.e., c_1), it is not too sensitive. As suggested by the theory, when the bandwidth increases, the variance tends to decrease, and the bias tends to increase. This shows the trade-off between the bias and the variance. In general, Silverman's rule of thumb works reasonably well. Third, as expected, we usually

https://doi.org/10.1017/S0266466621000554 Published online by Cambridge University Press

TABLE 3. Estimation results for DGP III

x_1		−0.8	−0.6	−0.4	−0.2	0.2	0.4	0.6	0.8
$m_1(x_1)$		−0.192	−0.036	0.016	0.012	0.028	0.144	0.396	0.832
	c_1				$n = 500$				
Bias		−0.0547	−0.0742	−0.0512	−0.0244	0.0384	0.0534	0.0082	−0.1502
SD	1	0.1215	0.0989	0.0799	0.0551	0.0542	0.0777	0.0932	0.1089
RMSE		0.1332	0.1236	0.0949	0.0603	0.0664	0.0942	0.0935	0.1855
Bias		−0.0508	−0.1036	−0.0901	−0.0524	0.0544	0.0542	−0.0311	−0.2526
SD	1.5	0.0739	0.0564	0.0419	0.0264	0.0255	0.0382	0.0480	0.0580
RMSE		0.0896	0.1180	0.0994	0.0587	0.0600	0.0664	0.0572	0.2592
Bias		−0.0480	−0.1251	−0.1205	−0.0754	0.0642	0.0487	−0.0679	−0.3341
SD	2	0.0532	0.0390	0.0271	0.0160	0.0151	0.0236	0.0311	0.0387
RMSE		0.0716	0.1311	0.1235	0.0771	0.0660	0.0542	0.0747	0.3364
Bias		−0.0577	−0.0999	−0.0860	−0.0478	0.0506	0.0529	−0.0047	−0.1739
SD	CV	0.1010	0.0776	0.0647	0.0464	0.0372	0.0563	0.0773	0.1002
RMSE		0.1163	0.1265	0.1076	0.0666	0.0628	0.0772	0.0774	0.2007
c_1^{cv}		1.7034	1.9940	2.1970	1.7696	2.2012	1.5706	1.1922	1.1392
					$n = 2,000$				
Bias		−0.0559	−0.0562	−0.0331	−0.0137	0.0315	0.0499	0.0250	−0.0940
SD	1	0.0918	0.0757	0.0626	0.0442	0.0470	0.0648	0.0768	0.0875
RMSE		0.1075	0.0943	0.0708	0.0463	0.0565	0.0817	0.0807	0.1284
Bias		−0.0539	−0.0826	−0.0618	−0.0322	0.0433	0.0539	−0.0041	−0.1831
SD	1.5	0.0521	0.0402	0.0313	0.0209	0.0223	0.0320	0.0383	0.0449
RMSE		0.0749	0.0918	0.0693	0.0384	0.0487	0.0627	0.0385	0.1885
Bias		−0.0511	−0.1045	−0.0914	−0.0534	0.0549	0.0542	−0.0327	−0.2565
SD	2	0.0366	0.0270	0.0197	0.0124	0.0126	0.0188	0.0234	0.0282
RMSE		0.0628	0.1080	0.0935	0.0548	0.0563	0.0573	0.0402	0.2580
Bias		−0.0621	−0.0801	−0.0633	−0.0336	0.0441	0.0530	0.0158	−0.1107
SD	CV	0.0770	0.0581	0.0496	0.0360	0.0334	0.0464	0.0618	0.0782
RMSE		0.0989	0.0989	0.0804	0.0493	0.0553	0.0704	0.0638	0.1355
c_1^{cv}		1.4826	1.9178	2.1358	1.7880	2.1952	1.7030	1.2134	1.0812

Notes: c_1 refers to the constant for the bandwidth choice. $c_1^c v$ refers to the average value of the constant chosen by our CV method (equation (4.1)).

observe a relatively large bias when the evaluation point is close to the boundary (−1 or 1), e.g., $x_1 = -0.8$ or $x_1 = 0.8$. Figure 1 plots the true functions of $m_1(\cdot)$ for the six DGPs and their estimates averaging more than 1,000 replications when $n = 2,000$ and $c_1 = 1$. In general, the estimates and true parameters match well except those close to the boundary.

https://doi.org/10.1017/S0266466621000554 Published online by Cambridge University Press

TABLE 4. Estimation results for DGP IV

x_1		−0.8	−0.6	−0.4	−0.2	0.2	0.4	0.6	0.8
$m_1(x_1)$		0.64	0.36	0.16	0.04	0.04	0.16	0.36	0.64
	c_1				$n = 500$				
Bias		−0.1440	−0.0590	−0.0128	0.0027	0.0111	−0.0040	−0.0485	−0.1310
SD	1	0.1769	0.1454	0.1189	0.0812	0.0822	0.1214	0.1464	0.1745
RMSE		0.2280	0.1568	0.1195	0.0812	0.0829	0.1214	0.1542	0.2181
Bias		−0.2107	−0.0941	−0.0260	0.0015	0.0039	−0.0233	−0.0911	−0.2068
SD	1.5	0.0994	0.0764	0.0581	0.0374	0.0385	0.0600	0.0770	0.0964
RMSE		0.2330	0.1212	0.0637	0.0374	0.0387	0.0643	0.1192	0.2281
Bias		−0.2864	−0.1418	−0.0503	−0.0058	−0.0050	−0.0492	−0.1405	−0.2850
SD	2	0.0686	0.0518	0.0375	0.0231	0.0234	0.0378	0.0512	0.0661
RMSE		0.2945	0.1509	0.0628	0.0238	0.0239	0.0621	0.1495	0.2925
Bias		−0.1942	−0.0910	−0.0300	−0.0025	0.0040	−0.0232	−0.0828	−0.1826
SD		0.1510	0.1180	0.0898	0.0571	0.0575	0.0900	0.1167	0.1492
RMSE	CV	0.2459	0.1489	0.0946	0.0571	0.0576	0.0928	0.1430	0.2358
c_1^{cv}		1.7248	1.6638	1.8098	2.0154	1.9966	1.8068	1.6412	1.7128
$\tilde{c}^{cv}$		1.6904	1.7166	1.8356	1.9478	1.8490	1.7686	1.6952	1.6912
					$n = 2,000$				
Bias		−0.0918	−0.0291	0.0032	0.0108	0.0095	0.0037	−0.0253	−0.0888
SD	1	0.1511	0.1267	0.1036	0.0735	0.0724	0.1036	0.1265	0.1500
RMSE		0.1767	0.1299	0.1036	0.0742	0.0730	0.1036	0.1290	0.1742
Bias		−0.1446	−0.0549	−0.0067	0.0080	0.0076	−0.0063	−0.0533	−0.1433
SD	1.5	0.0814	0.0649	0.0516	0.0349	0.0349	0.0510	0.0636	0.0800
RMSE		0.1659	0.0850	0.0520	0.0358	0.0357	0.0514	0.0829	0.1641
Bias		−0.2018	−0.0876	−0.0217	0.0038	0.0038	−0.0210	−0.0864	−0.2008
SD	2	0.0531	0.0410	0.0314	0.0205	0.0203	0.0308	0.0402	0.0527
RMSE		0.2087	0.0967	0.0381	0.0208	0.0206	0.0373	0.0953	0.2076
Bias		−0.1381	−0.0568	−0.0102	0.0072	0.0035	−0.0133	−0.0571	−0.1388
SD		0.1196	0.0939	0.0724	0.0493	0.0486	0.0730	0.0951	0.1212
RMSE	CV	0.1826	0.1097	0.0731	0.0498	0.0487	0.0742	0.1109	0.1842
c_1^{cv}		1.7152	1.7444	1.9596	2.1146	2.1090	1.9224	1.7252	1.7318
$\tilde{c}^{cv}$		1.6978	1.6850	1.6958	1.8674	1.8184	1.6604	1.5762	1.7022

Notes: c refers to the constant for the bandwidth choice. $c_1^c v$ and $\tilde{c}^c v$ refers to the average value of the constant chosen by our CV method (equation (4.1)).

https://doi.org/10.1017/S0266466621000554 Published online by Cambridge University Press

TABLE 5. Estimation results for DGP V

x_1		−0.8	−0.6	−0.4	−0.2	0.2	0.4	0.6	0.8
$m_1(x_1)$		0.64	0.36	0.16	0.04	0.04	0.16	0.36	0.64
	c_1				$n = 500$				
Bias		−0.1958	−0.0877	−0.0244	0.0008	0.0036	−0.0247	−0.0909	−0.2025
SD	1	0.1783	0.1445	0.1169	0.0821	0.0846	0.1239	0.1517	0.1855
RMSE		0.2647	0.1689	0.1194	0.0821	0.0846	0.1263	0.1768	0.2746
Bias		−0.2641	−0.1267	−0.0419	−0.0032	−0.0020	−0.0419	−0.1286	−0.2682
SD	1.5	0.1032	0.0780	0.0588	0.0385	0.0395	0.0615	0.0813	0.1064
RMSE		0.2835	0.1488	0.0722	0.0386	0.0395	0.0744	0.1521	0.2885
Bias		−0.3330	−0.1704	−0.0646	−0.0101	−0.0099	−0.0648	−0.1716	−0.3357
SD	2	0.0727	0.0541	0.0390	0.0241	0.0244	0.0396	0.0552	0.0742
RMSE		0.3409	0.1788	0.0754	0.0262	0.0263	0.0759	0.1803	0.3438
Bias		−0.2476	−0.1213	−0.0422	−0.0044	−0.0035	−0.0435	−0.1243	−0.2534
SD		0.1490	0.1145	0.0865	0.0564	0.0593	0.0915	0.1181	0.1535
RMSE	CV	0.2890	0.1668	0.0962	0.0566	0.0594	0.1013	0.1714	0.2962
c_1^{cv}		1.8192	1.7342	1.8386	2.0320	2.0262	1.8658	1.7530	1.8316
$\tilde{c}^{cv}$		1.7698	1.8326	1.8216	1.9392	1.9074	1.8248	1.8176	1.7712
					$n = 2,000$				
Bias		−0.1558	−0.0673	−0.0147	0.0070	0.0018	−0.0215	−0.0758	−0.1713
SD	1	0.1468	0.1263	0.1065	0.0756	0.0774	0.1037	0.1226	0.1449
RMSE		0.2140	0.1431	0.1074	0.0759	0.0774	0.1058	0.1441	0.2244
Bias		−0.2087	−0.0939	−0.0255	0.0030	0.0011	−0.0282	−0.0980	−0.2160
SD	1.5	0.0823	0.0665	0.0538	0.0366	0.0360	0.0517	0.0641	0.0802
RMSE		0.2244	0.1150	0.0595	0.0367	0.0360	0.0588	0.1171	0.2304
Bias		−0.2610	−0.1240	−0.0397	−0.0015	−0.0024	−0.0410	−0.1264	−0.2653
SD	2	0.0563	0.0429	0.0325	0.0211	0.0208	0.0316	0.0416	0.0546
RMSE		0.2670	0.1312	0.0513	0.0211	0.0209	0.0517	0.1331	0.2709
Bias		−0.2030	−0.0958	−0.0292	0.0017	−0.0017	−0.0349	−0.1038	−0.2181
SD		0.1159	0.0940	0.0758	0.0519	0.0522	0.0717	0.0898	0.1128
RMSE	CV	0.2337	0.1342	0.0812	0.0519	0.0522	0.0797	0.1372	0.2455
c_1^{cv}		1.7968	1.7950	1.9948	2.1012	2.1120	1.9848	1.8176	1.8380
$\tilde{c}^{cv}$		1.7366	1.7036	1.7400	1.8100	1.8204	1.7378	1.7698	1.8006

Notes: c refers to the constant for the bandwidth choice. c_1^cv and $\tilde{c}^cv$ refers to the average value of the constant chosen by our CV method (equation (4.1)).

https://doi.org/10.1017/S0266466621000554 Published online by Cambridge University Press

TABLE 6. Estimation results for DGP V

x_1		−0.8	−0.6	−0.4	−0.2	0.2	0.4	0.6	0.8
$m_1(x_1)$		−0.192	−0.036	0.016	0.012	0.028	0.144	0.396	0.832
	c_1				$n = 500$				
Bias		−0.0211	−0.0609	−0.0501	−0.0280	0.0329	0.0331	−0.0274	−0.1995
SD	1	0.1859	0.1486	0.1206	0.0818	0.0821	0.1207	0.1451	0.1684
RMSE		0.1870	0.1605	0.1305	0.0864	0.0884	0.1251	0.1475	0.2610
Bias		−0.0185	−0.0854	−0.0813	−0.0485	0.0466	0.0373	−0.0580	−0.2894
SD	1.5	0.1129	0.0836	0.0613	0.0383	0.0383	0.0591	0.0751	0.0907
RMSE		0.1144	0.1195	0.1019	0.0618	0.0603	0.0699	0.0948	0.3032
Bias		−0.0185	−0.1059	−0.1087	−0.0689	0.0566	0.0349	−0.0886	−0.3621
SD	2	0.0822	0.0589	0.0402	0.0235	0.0226	0.0363	0.0486	0.0609
RMSE		0.0842	0.1211	0.1159	0.0728	0.0610	0.0503	0.1010	0.3672
Bias		−0.0238	−0.0792	−0.0750	−0.0456	0.0431	0.0344	−0.0405	−0.2282
SD		0.1452	0.1099	0.0889	0.0607	0.0571	0.0870	0.1146	0.1453
RMSE	CV	0.1471	0.1354	0.1163	0.0759	0.0715	0.0935	0.1215	0.2705
c_1^{cv}		1.9334	2.1472	2.1688	1.9584	2.1804	1.8294	1.3944	1.2994
$\tilde{c}^{cv}$		1.7514	1.7952	1.9758	2.1128	1.8926	1.7630	1.6802	1.7254
					$n = 2,000$				
Bias		−0.0242	−0.0448	−0.0296	−0.0125	0.0223	0.0295	−0.0097	−0.1449
SD	1	0.1586	0.1331	0.1076	0.0744	0.0724	0.1030	0.1241	0.1425
RMSE		0.1604	0.1404	0.1115	0.0755	0.0757	0.1071	0.1244	0.2032
Bias		−0.0285	−0.0700	−0.0561	−0.0295	0.0356	0.0374	−0.0303	−0.2200
SD	1.5	0.0915	0.0716	0.0553	0.0360	0.0350	0.0510	0.0620	0.0738
RMSE		0.0958	0.1001	0.0788	0.0465	0.0499	0.0633	0.0690	0.2320
Bias		−0.0282	−0.0909	−0.0837	−0.0493	0.0482	0.0411	−0.0526	−0.2841
SD	2	0.0634	0.0473	0.0345	0.0214	0.0204	0.0308	0.0389	0.0476
RMSE		0.0693	0.1024	0.0905	0.0537	0.0524	0.0514	0.0654	0.2881
Bias		−0.0227	−0.0617	−0.0525	−0.0286	0.0328	0.0317	−0.0236	−0.1754
SD		0.1203	0.0961	0.0781	0.0551	0.0513	0.0733	0.0951	0.1197
RMSE	CV	0.1224	0.1142	0.0941	0.0620	0.0609	0.0798	0.0979	0.2123
c_1^{cv}		1.8620	2.0892	2.1496	1.9848	2.2004	1.9282	1.4378	1.2528
$\tilde{c}^{cv}$		1.6284	1.7028	1.9088	2.0178	1.8612	1.6730	1.6198	1.7442

Notes: c refers to the constant for the bandwidth choice. $c_1^c v$ and $\tilde{c}^c v$ refers to the average value of the constant chosen by our CV method (equation (4.1)).

https://doi.org/10.1017/S0266466621000554 Published online by Cambridge University Press

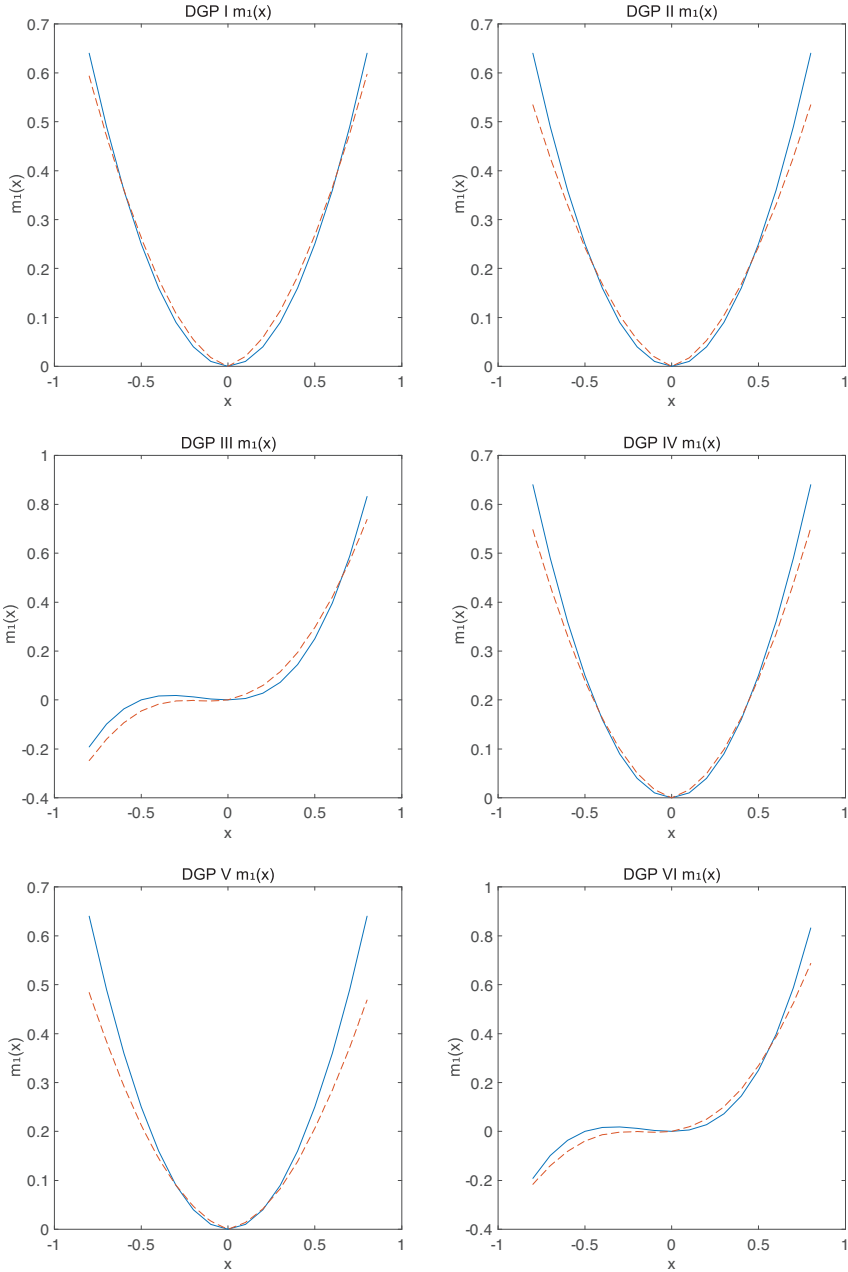

FIGURE 1. $m_1(\cdot)$ and the average of $\hat{m}_1(\cdot)$ for $n = 2,000$ and $c_1 = 1$.

https://doi.org/10.1017/S0266466621000554 Published online by Cambridge University Press

We also propose a data-driven method to choose the bandwidth. Specifically, we set $h_1 = c_1 h_1^0$ and $\tilde{h} = \tilde{c}\tilde{h}^0$, where $(h_1^0, \tilde{h}^0) = (1.06\hat{\sigma}_1 n^{-1/5}, n^{-1/6}\hat{\sigma}_2)$ for $c_1 \in \{0.6, \ldots, 2.6\}$ and $\tilde{c} \in \{0.6, \ldots, 2.6\}$. For each evaluation point x_1, we choose the constant $(c_1, \tilde{c})$ by minimizing the following leave-one-out cross-validation (CV) function

$$CV(c_1, \tilde{c}) = \sum_{i=1}^{n} \left(d_i - \frac{1}{2} - \hat{m}'_{1,-i}(x_1)(X_{1,i2} - X_{1,i1}) \right)^2 k_1\left(\frac{X_{1,i2} - x_1}{h_1^0}\right) k_1\left(\frac{X_{1,i1} - x_1}{h_1^0}\right) \tilde{k}\left(\frac{\tilde{X}_{i2} - \tilde{X}_{i1}}{\tilde{h}^0}\right), \tag{4.1}$$

where $\hat{m}'_{1,-i}(x_1)$ signifies the leave-one-out estimator using the bandwidth $(h_1, \tilde{h})$ with the ith observation deleted. Note that the above CV function uses $(h_1^0, \tilde{h}^0)$ as preliminary bandwidths.[9] The selected values of $(c_1, \tilde{c})$ are denoted as $(c_1^{cv}, \tilde{c}^{cv})$. At the bottom of Tables 1–6, we report the estimation results and $(c_1^{cv}, \tilde{c}^{cv})$ when the CV method is implemented. In general, our CV method performs well. In most DGPs, the bandwidth selected by the CV method tends to be relatively large, which can lead to a relatively large bias.

5. EXTENSIONS

In this section, we briefly discuss two possible extensions. First, we allow the structural function to be nonadditive. Second, we consider how to estimate the structural function when we only have cross-sectional data.

5.1. Models with a Nonadditive Structure

The generalized panel data transformation model with a nonadditive structural function is specified as

$$\Lambda(Y_{it}) = m(\mathbf{X}_{it}) + \alpha_i + \varepsilon_{it}, \tag{5.1}$$

where $m(\cdot)$ is a general function: $\mathbb{R}^K \to \mathbb{R}$. The nonadditive structure allows for arbitrary interactions among observed covariates. For simplicity, we take $K = 2$ and $T = 2$. Therefore, the model is

$$\Lambda(Y_{it}) = m(X_{1,it}, X_{2,it}) + \alpha_i + \varepsilon_{it},$$

for $t = 1, 2$.

[9] The CV function is quite different from the standard CV function for nonparametric kernel regressions. The reason is that our approximation equation (2.3) only holds in a particular neighborhood where $X_{1,i2} \approx x_1, X_{1,i1} \approx x_1$, and $\tilde{X}_{i2} - \tilde{X}_{i1} \approx 0$. To control for that, we use kernel functions with preliminary bandwidths in our CV function. Although our CV function appears intuitive, it is difficult to justify theoretically.

https://doi.org/10.1017/S0266466621000554 Published online by Cambridge University Press

We consider a neighborhood where $(X_{1,i2}, X_{2,i2}) \approx (x_1, x_2)$ and $(X_{1,i1}, X_{2,i1}) \approx (x_1, x_2)$, where (x_1, x_2) is an evaluation point. A Taylor expansion yields

$$m(X_{1,i2}, X_{2,i2}) - m(X_{1,i1}, X_{2,i1}) \approx \mathfrak{m}_1'(x_1, x_2)(X_{1,i2} - X_{1,i1}) + \mathfrak{m}_2'(x_1, x_2)(X_{2,i2} - X_{2,i1}),$$

where $\mathfrak{m}_1'(x_1, x_2) = \frac{\partial m(x_1, x_2)}{\partial x_1}$ and $\mathfrak{m}_2'(x_1, x_2) = \frac{\partial m(x_1, x_2)}{\partial x_2}$. It follows that

$$\begin{aligned} &E(d_i | \mathbf{X}_{i1}, \mathbf{X}_{i2}) \\ &= F_\Delta(m(X_{1,i2}, X_{2,i2}) - m(X_{1,i1}, X_{2,i1})) \\ &\approx F_\Delta(0) + f_\Delta(0)[\mathfrak{m}_1'(x_1, x_2)(X_{1,i2} - X_{1,i1}) + \mathfrak{m}_2'(x_1, x_2)(X_{2,i2} - X_{2,i1})] \\ &= \frac{1}{2} + \mathfrak{m}_1'(x_1, x_2)(X_{1,i2} - X_{1,i1}) + \mathfrak{m}_2'(x_1, x_2)(X_{2,i2} - X_{2,i1}). \end{aligned} \tag{5.2}$$

Let $w_i(x_1, x_2) = k_1\left(\frac{X_{1,i2} - x_1}{h_1}\right) k_1\left(\frac{X_{1,i1} - x_1}{h_1}\right) k_2\left(\frac{X_{2,i2} - x_2}{h_2}\right) k_2\left(\frac{X_{2,i1} - x_2}{h_2}\right)$, where $k_1(\cdot)$ and $k_2(\cdot)$ are kernel functions with respect to the first and second covariates, respectively. h_1 and h_2 are the corresponding bandwidth parameters. With $w_i(x_1, x_2)$ as a weight, equation (5.2) suggests that $\mathfrak{m}_1'(x_1, x_2)$ and $\mathfrak{m}_2'(x_1, x_2)$ can be estimated by a local linear regression of $\left(d_i - \frac{1}{2}\right)$ on $(X_{1,i2} - X_{1,i1})$ and $(X_{2,i2} - X_{2,i1})$. Therefore, we minimize the following LS criterion function:

$$\min_{(b_1, b_2)} \sum_{i=1}^{n} \left\{ \left[\left(d_i - \frac{1}{2} \right) - \begin{pmatrix} \Delta X_{1,i} \\ \Delta X_{2,i} \end{pmatrix}' \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \right]^2 w_i(x_1, x_2) \right\},$$

where $\Delta X_{1,i} = (X_{1,i2} - X_{1,i1})$ and $\Delta X_{2,i} = (X_{2,i2} - X_{2,i1})$. Define $\Delta \mathbf{X}_i = \left(\Delta X_{1,i}, \Delta X_{2,i}\right)'$. The solution to the above minimization problem is the estimator of $\left(\mathfrak{m}_1'(x_1, x_2),\ \mathfrak{m}_2'(x_1, x_2)\right)$:

$$\begin{pmatrix} \hat{\mathfrak{m}}_1'(x_1, x_2) \\ \hat{\mathfrak{m}}_2'(x_1, x_2) \end{pmatrix} = \left(\sum_{i=1}^{n} \Delta \mathbf{X}_i \cdot \Delta \mathbf{X}_i' \cdot w_i \right)^{-1} \left(\sum_{i=1}^{n} \Delta \mathbf{X}_i \cdot \left(d_i - \frac{1}{2} \right) \cdot w_i \right).$$

We impose the location normalization that $m\left(x_{1,0}, x_{2,0}\right) = 0$ for an interior point $\left(x_{1,0}, x_{2,0}\right)$. For a fixed interior evaluation point (x_1, x_2), note that

$$\int_{x_{1,0}}^{x_1} \mathfrak{m}_1'(\xi_1, x_{2,0}) d\xi_1 = m(x_1, x_{2,0}) - m(x_{1,0}, x_{2,0}),$$

and

$$\int_{x_{2,0}}^{x_2} \mathfrak{m}_2'(x_1, \xi_2) d\xi_2 = m(x_1, x_2) - m(x_1, x_{2,0}).$$

Subsequently,

$$m(x_1, x_2) = \int_{x_{2,0}}^{x_2} \mathfrak{m}_2'(x_1, \xi_2) d\xi_2 + \int_{x_{1,0}}^{x_1} \mathfrak{m}_1'(\xi_1, x_{2,0}) d\xi_1. \tag{5.3}$$

https://doi.org/10.1017/S0266466621000554 Published online by Cambridge University Press

By replacing the integrands with $\left(\hat{\mathfrak{m}}_1'(x_1,x_2),\hat{\mathfrak{m}}_2'(x_1,x_2)\right)$, equation (5.3) suggests the nonparametric estimator of $m(x_1,x_2)$ as

$$\hat{m}(x_1,x_2)=\int_{x_{2,0}}^{x_2}\hat{\mathfrak{m}}_2'(x_1,\xi_2)d\xi_2+\int_{x_{1,0}}^{x_1}\hat{\mathfrak{m}}_1'(\xi_1,x_{2,0})d\xi_1. \tag{5.4}$$

To save space, we omit details about the asymptotic properties of $\hat{m}(x_1,x_2)$. Although the nonadditive model (5.1) is more general, it has several disadvantages. First, the estimator of the nonadditive model has a slower convergence rate than that of the additive model. For example, we can show that $\hat{m}(x_1,x_2)$ in (5.4) is asymptotically normally distributed with a convergence rate of $\sqrt{nh_1^3h_2^2+nh_1^2h_2^3}$, which is slower than that of $\hat{m}(x_1)$ in (2.6) $\left(\sqrt{nh_1^3h_2}\text{ for }K=2\right)$. Second, here, we are only considering the simple case where $K=2$; therefore, we only need to calculate two integrals in (5.4). When $K>2$, recovering the original structural function from its derivative estimators is complicated and requires multiple integrations, which can be tedious. Third, the univariate function in the additive model is usually easier to interpret than the general function $m(\cdot):\mathbb{R}^K\to\mathbb{R}$.

5.2. Models with Cross-Sectional Data

Our approach can also be applied to cross-sectional data. With cross-sectional data, we cannot allow fixed effects to account for unobservable individual heterogeneity. Specifically, consider the cross-sectional transformation model with an additive structure,

$$\Lambda(Y_i)=\sum_{k=1}^{K}m_k(X_{k,i})+\varepsilon_i,\ i=1,\ldots,n. \tag{5.5}$$

This model has been studied in several papers, including Horowitz (2001), Horowitz and Mammen Horowitz and Mammen (2004, 2007, 2011), and Jacho-Chávez, Lewbel and Linton (2010). Our approach has two main advantages. First, unlike the existing methods which require both monotonicity and smoothness of the transformation function, our paper relaxes the smoothness condition. Second, our estimator is computationally easy.

Throughout this subsection, we assume an i.i.d. sample of $(\mathbf{X}_i,Y_i)$, and that ε_i is independent of $\mathbf{X}_i$, where $\mathbf{X}_i=(X_{1,i},\ldots,X_{K,i})$. Define $\Delta_{ij}=\left(\varepsilon_j-\varepsilon_i\right)$. Let $F_\Delta(\cdot)$ and $f_\Delta(\cdot)$ be the cdf and pdf of Δ_{ij}, respectively. As in the panel data case, we normalize $f_\Delta(0)=1$ and $m_k(x_{k,0})=0$, for $k=1,\ldots,K$, for an interior point $(x_{1,0},x_{2,0},\ldots,x_{K,0})$ in $\mathcal{X}$. We also focus on the first structural function $m_1(\cdot)$. Write $\mathbf{X}_i=\left(X_{1,i},\tilde{X}_i'\right)'$ and $\tilde{m}\left(\tilde{X}_i\right)=\sum_{k=2}^{K}m_k(X_{k,i})$, where $\tilde{X}_i=(X_{2,i},\ldots,X_{K,i})'$. Define $d_{ij}=\mathbf{1}\left\{Y_i>Y_j\right\}$. Considering a neighborhood where $X_{1,i}\approx x_1,X_{1,j}\approx x_1$, and

https://doi.org/10.1017/S0266466621000554 Published online by Cambridge University Press

$\tilde{X}_i - \tilde{X}_j \approx 0$, we have

$$\begin{aligned} E(d_{ij}|\mathbf{X}_i, \mathbf{X}_j) &= F_\Delta[m_1(X_{1,i}) - m_1(X_{1,j}) + \tilde{m}(\tilde{X}_i) - \tilde{m}(\tilde{X}_j)] \\ &\approx F_\Delta(0) + f_\Delta(0) m_1'(x_1)\left(X_{1,i} - X_{1,j}\right) \\ &= \frac{1}{2} + m_1'(x_1)\left(X_{1,i} - X_{1,j}\right). \end{aligned}$$

The first equality holds because of the independence between Δ_{ij} and $\left(\mathbf{X}_i, \mathbf{X}_j\right)$. In the approximation formula, we remove $\tilde{m}(\cdot)$ by matching $\tilde{X}_i$ with $\tilde{X}_j$ and apply a local linear approximation to the first structural function. The i.i.d. assumption implies that Δ_{ij} is symmetric around zero, and thus $F_\Delta(0) = \frac{1}{2}$. The above local approximation suggests that we can obtain the estimator of $m_1'(x_1)$ by regressing $(d_{ij} - \frac{1}{2})$ on $(X_{1,i} - X_{1,j})$ locally. We minimize the following LS criterion function:

$$\min_b \sum_{i \neq j} \left\{ \left[\left(d_{ij} - \frac{1}{2} \right) - b \cdot (X_{1,i} - X_{1,j}) \right]^2 k_1\left(\frac{X_{1,i} - x_1}{h_1}\right) k_1\left(\frac{X_{1,j} - x_1}{h_1}\right) \tilde{k}\left(\frac{\tilde{X}_i - \tilde{X}_j}{\tilde{h}}\right) \right\},$$

where the weights $k_1\left(\frac{X_{1,i}-x_1}{h_1}\right) k_1\left(\frac{X_{1,j}-x_1}{h_1}\right) \tilde{k}\left(\frac{\tilde{X}_i-\tilde{X}_j}{\tilde{h}}\right)$ are kernel functions defined analogously to the panel case. The solution to the above minimization problem is the estimator of $m_1'(x_1)$:

$$\hat{m}_1'(x_1) = \frac{\sum_{i \neq j} (d_{ij} - \frac{1}{2})(X_{1,i} - X_{1,j}) k_1\left(\frac{X_{1,i}-x_1}{h_1}\right) k_1\left(\frac{X_{1,j}-x_1}{h_1}\right) \tilde{k}\left(\frac{\tilde{X}_i-\tilde{X}_j}{\tilde{h}}\right)}{\sum_{i \neq j} \left(X_{1,i} - X_{1,j}\right)^2 k_1\left(\frac{X_{1,i}-x_1}{h_1}\right) k_1\left(\frac{X_{1,j}-x_1}{h_1}\right) \tilde{k}\left(\frac{\tilde{X}_i-\tilde{X}_j}{\tilde{h}}\right)}.$$

Subsequently, the estimator of $m_1(x_1)$ is

$$\hat{m}_1(x_1) = \int_{x_{1,0}}^{x_1} \hat{m}_1'(\xi) d\xi. \tag{5.6}$$

Analogously, we can estimate $m_k(x_k)$, for $k = 2, \ldots, K$. To save space, we omit the detailed asymptotic analysis.

6. CONCLUSIONS

This paper considers a generalized panel data transformation model with fixed effects where no parametric assumptions are imposed on the transformation function, the structural functions, or the distribution of the error term. Therefore, compared with the existing models which often require certain parametric assumptions, our model is less likely to suffer from model misspecfication. We propose a kernel-based nonparametric estimator for the structural functions. Our proposed estimator has an explicit expression and is easy to implement. These properties are

https://doi.org/10.1017/S0266466621000554 Published online by Cambridge University Press

particularly appealing in practice. We show that our estimator is asymptotically normal.

There are several directions for future research. First, our estimators still suffer from the curse of dimensionality to some degree even if we assume that the structural function is additive. This is due to the fact that we have to match other covariates locally. It will be interesting to develop an estimation procedure to avoid this problem, which may require stronger assumptions. Second, we require full independence between all covariates and all error terms. This type of strict exogeneity assumption rules out dynamic models with lagged dependent variables. Honoré and Kyriazidou (2000) consider the use of the maximum score estimator for a binary choice fixed-effects model with lagged dependent variables. More generally, Honoré and Lewbel (2002) study a binary choice fixed-effects model without the strict exogeneity assumption. It will be useful for our model to allow dynamics in future studies.

REFERENCES

Abrevaya, J. (1999) Leapfrog estimation of a fixed effects models with unknown transformation of the dependent variable. *Journal of Econometrics* 93, 203–228.

Abrevaya, J. (2000) Rank estimation of a generalized fixed-effects regression model. *Journal of Econometrics* 95(1), 1–23.

Ai, C., J. You, & Y. Zhou (2014) Estimation of fixed effects panel data partially linear additive regression models. *Econometrics Journal* 17(1), 83–106.

Baltagi, B.H. & D. Li (2002) Series estimation of partially linear panel data models with fixed effects. *Annals of Economics and Finance* 3(1), 103–116.

Bhattacharjee, S., R.D. Gopal, K. Lertwachara, J.R. Marsden, & R. Telang (2007) The effect of digital sharing technologies on music markets: A survival analysis of albums on ranking charts. *Management Science* 53(9), 1359–1374.

Bickel, P.J. & K.A. Doksum (1981) An analysis of transformations revisited. *Journal of the American Statistical Association*, 76(374), 296–311.

Botosaru, I. & C. Muris (2018) Binarization for Panel Models with Fixed Effects. Cemmap Working paper, CWP31/17, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

Box, G.E. & D.R. Cox (1964) An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)* 26(2), 211–243.

Carroll, R.J., J. Fan, I. Gijbels, & M.P. Wand (1997) Generalized partially linear single-index models. *Journal of the American Statistical Association* 92(438), 477–489.

Cavanagh, C. & R.P. Sherman (1998) Rank estimators for monotonic index models. *Journal of Econometrics* 84, 351–381.

Chen, S. (2002) Rank estimation of transformation models. *Econometrica* 70, 1683–1697.

Chen, S. (2010) Root-*N*-consistent estimation of fixed-effect panel data transformation models with censoring. *Journal of Econometrics* 159, 222–234.

Chen, S. & X. Wang (2018) Semiparametric estimation of panel data models without monotonicity or separability. *Journal of Econometrics* 206, 515–530.

Chernozhukov, V., I. Fernández-Val, J. Hahn, & W. Newey (2013). Average and quantile effects in nonseparable panel models. *Econometrica* 81, 535–580.

Evdokimov, K. (2011) Nonparametric Identification of a Nonlinear Panel Model with Application to Duration Analysis with Multiple Spells. Working paper, Princeton University.

Ham, J.C. & S.A. Rea Jr. (1987) Unemployment insurance and male unemployment duration in Canada. *Journal of Labor Economics* 5, 325–353.

https://doi.org/10.1017/S0266466621000554 Published online by Cambridge University Press

Härdle, W. & T.M. Stoker (1989) Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association* 84(408), 986–995.

Henderson, D.J., R.J. Carroll, & Q. Li (2008) Nonparametric estimation and testing of fixed effects panel data models. *Journal of Econometrics* 144(1), 257–275.

Hoderlein, S. & H. White (2012) Nonparametric identification in nonseparable panel data models with generalized fixed effects. *Journal of Econometrics* 168(2), 300–314.

Honoré, B.E. & E. Kyriazidou (2000) Panel data discrete choice models with lagged dependent variables. *Econometrica* 68, 839–874.

Honoré, B.E. & A. Lewbel (2002) Semiparametric binary choice panel data models without strictly exogeneous regressors. *Econometrica* 70, 2053–2063.

Horowitz, J.L. (2001) Nonparametric estimation of a generalized additive model with an unknown link function. *Econometrica* 69, 499–513.

Horowitz, J.L. & S. Lee (2004) Semiparametric estimation of a panel data proportional hazards model with fixed effects. *Journal of Econometrics* 119(1), 155–198.

Horowitz, J.L. & E. Mammen (2004) Nonparametric estimation of an additive model with a link function. *Annals of Statistics* 32(6), 2412–2443.

Horowitz, J.L. & E. Mammen (2007) Rate-optimal estimation for a general class of nonparametric regression models with unknown link functions. *Annals of Statistics* 35(6), 2589–2619.

Horowitz, J.L. & E. Mammen (2011) Oracle-efficient nonparametric estimation of an additive model with an unknown link function. *Econometric Theory* 27(3), 582–608.

Jacho-Chávez, D., A. Lewbel, & O. Linton (2010) Identification and nonparametric estimation of a transformed additively separable model. *Journal of Econometrics* 156, 392–407.

Kaplan, D.M. & Y. Sun (2017) Smoothed estimating equations for instrumental variables quantile regression. *Econometric Theory* 33(1), 105–157.

Lancaster, T. (1997) Econometric methods for the duration of unemployment. *Econometrica* 47, 939–956.

Lee, S. (2008) Estimating panel data duration models with censored data. *Econometric Theory* 24, 1254–1276.

Lewbel, A. (2000) Semiparametric qualitative response model estimation with instrumental variables and unknown heteroscedasticity. *Journal of Econometrics* 97(1), 145–177.

Li, Q. & J.S. Racine (2007) *Nonparametric Econometrics: Theory and Practice*. Princeton University Press.

Li, Q. & T. Stengos (1996) Semiparametric estimation of partially linear panel data models. *Journal of Econometrics* 71(1–2), 389–397.

Mammen, E., B. Støve, & D. Tjøstheim (2009) Nonparametric additive models for panels of time series. *Econometric Theory* 25(2), 442–481.

Ozabaci, D., D.J. Henderson, & L. Su (2014) Additive nonparametric regression in the presence of endogenous regressors. *Journal of Business & Economic Statistics* 32(4), 555–575.

Pakes, A. & D. Pollard (1989) Simulation and the asymptotics of optimization estimators. *Econometrica* 57, 1027–1057.

Pitt, D.G.W. (2007) Modelling the claim duration of income protection insurance policyholders using parametric mixture models. *Annals of Actuarial Science* 2, 1–24.

Pollard, D. (1995) Uniform ratio limit theorems for empirical processes. *Scandinavian Journal of Statistics* 22, 271–278.

Powell, J.L., J.H. Stock, & T.M. Stoker (1989) Semiparametric estimation of index coefficients. *Econometrica* 57, 1403–1430.

Su, L. & X. Lu (2013) Nonparametric dynamic panel data models: Kernel estimation and specification testing. *Journal of Econometrics* 176(2), 112–133.

Su, L. & Y. Zhang (2016) Semiparametric estimation of partially linear dynamic panel data models with fixed effects. In *Essays in Honor of Aman Ullah* (Gloria Gonzalez-Rivera, R. Carter Hill and Tae-Hwy Lee, editors), pp. 137–204. Emerald Group Publishing Limited.

Topel, R.H. & M.P. Ward (1992) Job mobility and the careers of young men. *Quarterly Journal of Economics* 107, 439–479.
Wang, X. & S. Chen (2020) Semiparametric estimation of generalized transformation panel data models with non-stationary error. *The Econometrics Journal* 23(3), 386–402.
You, J., X. Zhou, & Y. Zhou (2011) Series estimation in partially linear in-slide regression models. *Scandinavian Journal of Statistics* 38(1), 89–107.

APPENDIX

This appendix contains the proofs for the main results.

Proof of Lemma 2.1. For any $w \in \mathcal{W}$, where $\mathcal{W}$ is the support of Δ_i,

$$\begin{aligned} F_\Delta(-w) &= P(\varepsilon_{i1} - \varepsilon_{i2} \le -w) = P(\varepsilon_{i2} - \varepsilon_{i1} \ge w) \\ &= 1 - P(\varepsilon_{i2} - \varepsilon_{i1} < w) = 1 - P(\varepsilon_{i1} - \varepsilon_{i2} < w) \\ &= 1 - F_\Delta(w), \end{aligned}$$

where the fourth equality holds because of the exchangeability assumption. Let $w = 0$, then $F_\Delta(-w) = 1 - F_\Delta(w)$ implies $2F_\Delta(0) = 1$, namely, $F_\Delta(0) = \frac{1}{2}$. □

Proof of Theorem 3.1. Note that

$$\left[\hat{m}_1'(x_1) - m_1'(x_1)\right] = \frac{S_n(x_1) + R_n(x_1)}{Q_n(x_1)},$$

where

$$\begin{aligned} S_n(x_1) = \frac{1}{nh_1}\sum_{i=1}^{n}(P_i - \frac{1}{2} - m_1'(x_1)(X_{1,i2} - X_{1,i1}))\left(\frac{X_{1,i2} - X_{1,i1}}{h_1}\right)\frac{1}{h_1^2}\frac{1}{\tilde{h}^{K-1}} \\ k_1\left(\frac{X_{1,i1} - x_1}{h_1}\right)k_1\left(\frac{X_{1,i2} - x_1}{h_1}\right)\tilde{k}\left(\frac{\tilde{X}_{i2} - \tilde{X}_{i1}}{\tilde{h}}\right), \end{aligned}$$

$$\begin{aligned} R_n(x_1) = \frac{1}{nh_1}\sum_{i=1}^{n}(d_i - P_i)\left(\frac{X_{1,i2} - X_{1,i1}}{h_1}\right)\frac{1}{h_1^2}\frac{1}{\tilde{h}^{K-1}} \\ k_1\left(\frac{X_{1,i1} - x_1}{h_1}\right)k_1\left(\frac{X_{1,i2} - x_1}{h_1}\right)\tilde{k}\left(\frac{\tilde{X}_{i2} - \tilde{X}_{i1}}{\tilde{h}}\right), \end{aligned}$$

and

$$\begin{aligned} Q_n(x_1) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_{1,i2} - X_{1,i1}}{h_1}\right)^2\frac{1}{h_1^2}\frac{1}{\tilde{h}^{K-1}} \\ k_1\left(\frac{X_{1,i1} - x_1}{h_1}\right)k_1\left(\frac{X_{1,i2} - x_1}{h_1}\right)\tilde{k}\left(\frac{\tilde{X}_{i2} - \tilde{X}_{i1}}{\tilde{h}}\right), \end{aligned}$$

https://doi.org/10.1017/S0266466621000554 Published online by Cambridge University Press

with $P_i = E(d_i|\mathbf{X}_{i1},\mathbf{X}_{i2}) = E(d_i|X_{1,i1},X_{1,i2},\tilde{X}_{i1},\tilde{X}_{i2})$. We further decompose the above into

$$\begin{aligned}
&\left[\hat{m}_1'(x_1) - m_1'(x_1)\right] \\
&= \frac{1}{Q(x_1)}E[S_n(x_1)] + \frac{1}{Q(x_1)}R_n(x_1) \\
&\quad + \frac{1}{Q(x_1)}[(S_n(x_1) - E[S_n(x_1)])] + \left(\frac{1}{Q_n(x_1)} - \frac{1}{Q(x_1)}\right)[S_n(x_1) + R_n(x_1)] \\
&\equiv A_1(x_1) + A_2(x_1) + A_3(x_1) + A_4(x_1), \text{ say,}
\end{aligned}$$

where $Q(x_1)$ is the probability limit of $Q_n(x_1)$. It is easy to show that $Q(x_1) = 2K_2\int p_0^2(x_1,\tilde{x}_1)d\tilde{x}_1$ with $K_2 = \int k_1(u)u^2du$, where $p_0(\bar{x}_1,\tilde{x}_1)$ and $p_0(\bar{x}_2,\tilde{x}_2)$ are joint densities of $(X_{1,i1},\tilde{X}_{i1})$ and $(X_{1,i2},\tilde{X}_{i2})$ evaluated at $(\bar{x}_1,\tilde{x}_1)$ and $(\bar{x}_2,\tilde{x}_2)$, respectively. We analyze the four terms separately.

We first analyze $A_1(x_1)$. Define

$$g(\bar{x}_1,\bar{x}_2,\tilde{x}_1,\tilde{x}_2) = F(\bar{x}_1,\bar{x}_2,\tilde{x}_1,\tilde{x}_2) - \frac{1}{2} - m_1'(x_1)(\bar{x}_2 - \bar{x}_1),$$

where $F(\bar{x}_1,\bar{x}_2,\tilde{x}_1,\tilde{x}_2) = E(d_i|X_{1,i1} = \bar{x}_1,\ X_{1,i2} = \bar{x}_2,\ \tilde{X}_{i1} = \tilde{x}_1,\ \tilde{X}_{i2} = \tilde{x}_2)$. Note that $g(\bar{x}_1,\bar{x}_1,\tilde{x}_1,\tilde{x}_1) = 0$. Under Assumptions 5, 6, and 7A, with some tedious algebra, we can show that

$$\begin{aligned}
\frac{1}{Q(x_1)}E(S_n(x_1)) &= \frac{1}{Q(x_1)}\frac{1}{h_1}\int g(\bar{x}_1,\bar{x}_2,\tilde{x}_1,\tilde{x}_2)\left(\frac{\bar{x}_2 - \bar{x}_1}{h_1}\right)\frac{1}{h_1^2}k_1\left(\frac{\bar{x}_1 - x_1}{h_1}\right)k_1\left(\frac{\bar{x}_2 - x_1}{h_1}\right) \\
&\quad \times \frac{1}{\tilde{h}^{K-1}}\tilde{k}\left(\frac{\tilde{x}_2 - \tilde{x}_1}{\tilde{h}}\right)p_0(\bar{x}_1,\tilde{x}_1)p_0(\bar{x}_2,\tilde{x}_2)d\bar{x}_1d\bar{x}_2d\tilde{x}_1d\tilde{x}_2 \\
&= h_1^2 b_1(x_1) + o(h_1^2) + O\left(\tilde{h}^s\right),
\end{aligned}$$

uniformly in $x_1 \in [x_{10},x_{20}]$, where

$$b_1(x_1) = \frac{b_{11}(x_1) + b_{12}(x_1)}{Q(x_1)},$$

with

$$\begin{aligned}
b_{11}(x_1) &= \frac{1}{6}\sum_{\alpha_1+\alpha_2=3}\int \frac{\partial^3}{\partial\bar{x}_1^{\alpha_1}\partial\bar{x}_2^{\alpha_2}}g(x_1,x_1,\tilde{x}_1,\tilde{x}_1)\cdot p_0^2(x_1,\tilde{x}_1)d\tilde{x}_1 \\
&\quad \times \int\int u_1^{\alpha_1}u_2^{\alpha_2}(u_2 - u_1)k_1(u_1)k_1(u_2)du_1du_2,
\end{aligned}$$

and

$$\begin{aligned}
b_{12}(x_1) &= \frac{1}{2}\sum_{\alpha_1+\alpha_2=2,\ \beta_1+\beta_2=1}\int \frac{\partial^2}{\partial\bar{x}_1^{\alpha_1}\partial\bar{x}_2^{\alpha_2}}g(x_1,x_1,\tilde{x}_1,\tilde{x}_1) \\
&\quad \cdot \frac{\partial}{\partial\bar{x}_1^{\beta_1}}p_0(x_1,\tilde{x}_1)\cdot\frac{\partial}{\partial\bar{x}_2^{\beta_2}}p_0(x_1,\tilde{x}_1)d\tilde{x}_1 \\
&\quad \times \int\int u_1^{\alpha_1+\beta_1}u_2^{\alpha_2+\beta_2}(u_2 - u_1)k_1(u_1)k_1(u_2)du_1du_2.
\end{aligned}$$

https://doi.org/10.1017/S0266466621000554 Published online by Cambridge University Press

$b_{11}(x_1)$ and $b_{12}(x_1)$ can be explicitly expressed as

$$b_{11}(x_1) = \frac{1}{3}\left[f''_{\Delta}(0)\cdot\left[m'(x_1)\right]^3\left(K_4+K_2^2\right)+m'''(x_1)K_4\right]\left[\int p_0^2(x_1,\tilde{x}_1)d\tilde{x}_1\right],$$

$$b_{12}(x_1) = m''(x_1)\left(K_4-K_2^2\right)\left(\int \frac{\partial}{\partial \tilde{x}_1}p_0(x_1,\tilde{x}_1)\cdot p_0(x_1,\tilde{x}_1)d\tilde{x}_1\right),$$

where $K_4 = \int k_1(u)u^4du$.

For $A_2(x_1)$ and $A_3(x_1)$, following Pollard (1995), it is straightforward to show that

$$A_2(x_1) = O_p\left(\frac{\ln n}{nh_1^4\tilde{h}^{K-1}}\right)^{1/2},$$

and

$$A_3(x_1) = \frac{1}{Q(x_1)}S_n(x_1) - E\left(S_n(x_1)\right) = O_p\left(\frac{(h_1+\tilde{h})\ln n}{nh_1^4\tilde{h}^{K-1}}\right)^{1/2},$$

uniformly in $x_1 \in [x_{10}, x_{20}]$.

Similarly, we can show that

$$Q_n(x_1) - Q(x_1) = O_p\left(\left(\frac{\ln n}{nh_1^2\tilde{h}^{K-1}}\right)^{1/2} + h_1^2 + \tilde{h}^s\right) \equiv O_p\left(\delta_{qn}\right),$$

and

$$S_n(x_1) + R_n(x_1) = O_p\left(\left(\frac{\ln n}{nh_1^4\tilde{h}^{K-1}}\right)^{1/2} + h_1^2 + \tilde{h}^s\right) \equiv O_p\left(\delta_{sn}\right),$$

uniformly in $x_1 \in [x_{10}, x_{20}]$. Thus

$$A_4(x_1) = O_p\left(\delta_{qn}\right)\cdot O_p\left(\delta_{sn}\right) = O_p\left(\frac{\ln n}{nh_1^3\tilde{h}^{K-1}} + h_1^4 + \tilde{h}^{2s}\right),$$

uniformly in $x_1 \in [x_{10}, x_{20}]$.

By combining the above analysis, we obtain

$$\left[\hat{m}'_1(x_1) - m'_1(x_1)\right] = h_1^2 b_1(x_1) + o(h_1^2) + O\left(\tilde{h}^s\right) + O_p\left(\frac{\ln n}{nh_1^4\tilde{h}^{K-1}}\right)^{1/2},$$

uniformly in $x_1 \in [x_{10}, x_{20}]$.

https://doi.org/10.1017/S0266466621000554 Published online by Cambridge University Press

To show asymptotic normality, note that $E(A_2(x_1)) = 0$ and

$$\begin{aligned}
&\operatorname{var}\left(\sqrt{nh_1^4\tilde{h}^{K-1}}A_2(x_1)\right) = \operatorname{var}\left(\sqrt{nh_1^4\tilde{h}^{K-1}}\frac{R_n(x_1)}{Q(x_1)}\right)\\
&= E\left[\frac{(d_i-P_i)^2}{Q(x_1)^2}\left(\frac{X_{1,i2}-X_{1,i1}}{h_1}\right)^2\frac{1}{h_1^2}\frac{1}{\tilde{h}^{K-1}}\right.\\
&\qquad \left. k_1^2\left(\frac{X_{1,i1}-x_1}{h_1}\right)k_1^2\left(\frac{X_{1,i2}-x_1}{h_1}\right)\tilde{k}^2\left(\frac{\tilde{X}_{i2}-\tilde{X}_{i1}}{\tilde{h}}\right)\right]\\
&= \sigma_1^2(x_1)+o(1),
\end{aligned}$$

where

$$\sigma_1^2(x_1) = \int\frac{\sigma^2(x_1,x_1,\tilde{x}_1,\tilde{x}_1)}{Q(x_1)^2}p_0^2(x_1,\tilde{x}_1)d\tilde{x}_1\int(u_2-u_1)^2k_1^2(u_1)k_1^2(u_2)\tilde{k}^2(\tilde{u})du_1du_2d\tilde{u},$$

with $\sigma^2(x_1,x_2,\tilde{x}_1,\tilde{x}_2) = E\left[(d_i-P_i)^2|X_{1,i1}=x_1,X_{1,i2}=x_2,\tilde{X}_{i1}=\tilde{x}_1,\tilde{X}_{i2}=\tilde{x}_2\right]$. An application of the Lindeberg–Feller Central Limit Theorem yields

$$\sqrt{nh_1^4\tilde{h}^{K-1}}A_2(x_1)\xrightarrow{d}N(0,\sigma_1^2(x_1)).$$

In addition, by the above convergence results for $A_1(x), A_3(x_1)$, and $A_4(x_1)$ and Assumption 7A, we obtain

$$\sqrt{nh_1^4\tilde{h}^{K-1}}(\hat{m}_1'(x_1)-m_1'(x_1)-h_1^2b_1(x_1))\xrightarrow{d}N(0,\sigma_1^2(x_1)).$$ □

Proof of Theorem 3.2. Following the arguments in the proof of Theorem 3.1, we can show that

$$\int_{x_{1,0}}^{x_1}A_1(\xi_1)d\xi_1 = h_1^2b_2(x_1)+o(h_1^2)+O\left(\tilde{h}^s\right),$$

$$\int_{x_{1,0}}^{x_1}A_2(\xi_1)d\xi_1 = O_p\left(\frac{\ln n}{nh_1^3\tilde{h}^{K-1}}\right)^{1/2},$$

$$\int_{x_{1,0}}^{x_1}A_3(\xi_1)d\xi_1 = \frac{1}{Q(x_1)}S_n(x_1)-E(S_n(x_1)) = O_p\left(\frac{(h_1+\tilde{h})\ln n}{nh_1^3\tilde{h}^{K-1}}\right)^{1/2},$$

and

$$\int_{x_{1,0}}^{x_1}A_4(\xi_1)d\xi_1 = o_p\left(\frac{\ln n}{nh_1^3\tilde{h}^{K-1}}\right)^{1/2}+o_p\left(h_1^2+\tilde{h}^s\right),$$

uniformly in $x_1\in[x_{10},x_{20}]$. Consequently, we obtain

$$\int_{x_{1,0}}^{x_1}\hat{m}_1'(\xi_1)d\xi_1-m_1(x_1) = h_1^2b_2(x_1)+o(h_1^2)+O_p\left(\frac{\ln n}{nh_1^3\tilde{h}^{K-1}}\right)^{1/2}+O\left(\tilde{h}^s\right),$$

https://doi.org/10.1017/S0266466621000554 Published online by Cambridge University Press

uniformly in $x_1 \in [x_{10}, x_{20}]$. In addition, we can show that

$$\sqrt{nh_1^3\tilde{h}^{K-1}}\int_{x_{1,0}}^{x_1} A_2(\xi_1)\,d\xi_1 \xrightarrow{d} N(0,\sigma_2^2(x_1)),$$

where

$$\sigma_2^2(x_1) = \int \frac{\mathbf{I}(\xi_1)}{Q^2(\xi_1)}\sigma^2(\xi_1,\xi_1,\tilde{x}_1,\tilde{x}_1)p_0^2(\xi_1,\tilde{x}_1)\,d\xi_1 d\tilde{x}_1$$
$$\int\left[\int k_1(u+u_1)k_1(u_1)du_1\right]^2 u^2du\int\tilde{k}^2(\tilde{u})d\tilde{u},$$

and $\mathbf{I}(\xi_1) = \mathbf{1}\{x_{1,0} < \xi_1 < x_1\}$ when $x_{1,0} < x_1$, and $\mathbf{I}(\xi_1) = \mathbf{1}\{x_1 < \xi_1 < x_{1,0}\}$ when $x_{1,0} > x_1$. Then, by the above rates of convergence results and by Assumption 7B, we obtain

$$\sqrt{nh_1^3\tilde{h}^{K-1}}(\hat{m}_1(x_1) - m_1(x_1) - h_1^2 b_2(x_1)) \xrightarrow{d} N(0,\sigma_2^2(x_1)).$$ □

Proof of Theorem 3.3. A simple Taylor expansion yields

$$\Gamma_{ns}(b;y,\hat{m}) = \Gamma_{ns}(b;y,m) + \tau_{ns},$$

where

$$\tau_{ns} = \frac{1}{n}\sum_{i=1}^{n}(d_{yi2} - d_{y_0i1})\frac{1}{h_y}k_y\left(\frac{\bar{m}(\mathbf{X}_{i2}) - \bar{m}(\mathbf{X}_{i1}) - b}{h_y}\right)$$
$$\left[\left(\hat{m}(\mathbf{X}_{i2}) - \hat{m}(\mathbf{X}_{i1})\right) - (m(\mathbf{X}_{i2}) - m(\mathbf{X}_{i1}))\right],$$

and $\bar{m}(\mathbf{X}_{i2}) - \bar{m}(\mathbf{X}_{i1})$ is between $\hat{m}(\mathbf{X}_{i2}) - \hat{m}(\mathbf{X}_{i1})$ and $m(\mathbf{X}_{i2}) - m(\mathbf{X}_{i1})$. Therefore, we obtain

$$|\tau_{ns}| = O_p\left(h_y^{-1}\delta_n\right) = o_p(1),$$

which, together with an application of the uniform law of large numbers (Pakes and Pollard, 1989), implies that

$$\Gamma_{ns}(b;y,\hat{m}) - \Gamma(b;y,m) = o_p(1),$$

uniformly in $(y,b) \in [y_1,y_2]\times\mathcal{L}$. In addition, it is straightforward to check that $\Gamma(b;y,m)$ is continuous in b and achieves the unique maximum at $\Lambda(y)$. Therefore, $\hat{\Lambda}(y)$ is consistent for $\Lambda(y)$; furthermore, Lemma A.1 in Carroll et al. (1997) implies the uniform consistency. □

https://doi.org/10.1017/S0266466621000554 Published online by Cambridge University Press