Estimating Semiparametric Panel Data Models by Marginal Integration

Junhui Qian and Le Wang

Shanghai Jiao Tong University

10. November 2009

Estimating Semiparametric Panel Data Models by Marginal Integration[1]

Junhui Qian
School of Economics
Shanghai Jiao Tong University
Fua Hua Zhen Road 535
Shanghai, 200052, China
jhqian@sjtu.edu.cn

Le Wang
University of New Hampshire
15 College Road
Durham, NH 03824
Le.Wang@unh.edu

Abstract

We propose a new methodology for estimating semiparametric panel data models, with a primary focus on the nonparametric component. We eliminate individual effects using first differencing transformation and estimate the unknown function by marginal integration. We extend our methodology to treat panel data models with both individual and time effects. And we characterize the asymptotic behavior of our estimators. Monte Carlo simulations show that our estimator behaves well in finite samples in both random effects and fixed effects settings.

This version: November 10, 2009

JEL classification codes: C13; C14; C24

Key words: Semiparametric Panel Data Model, Partially Linear, First Differencing, Marginal Integration

[1] We thank two referees for detailed and constructive comments, which led to a substantial improvement. And we thank Daniel Henderson, Qi Li, Essie Maasoumi, and participants in SETA 2008 for very useful discussions or comments. All remaining errors are ours. This research is partly supported by the Scientific Research Fund of Antai College of Economics & Management, Shanghai Jiao Tong University.

1. Introduction

In many empirical studies involving panel data sets, at least in the initial stage of research, it is useful to consider semiparametric models like the following,

$$Y_{it} = \alpha_i + \beta' Z_{it} + f(X_{it}) + \varepsilon_{it}, \tag{1}$$

where α_i is unobserved individual effect and X_{it} is most likely a low-dimensional covariate vector that relates to Y_{it} via an unknown function f. We call the above model semiparametric since only part of the covariate vector (i.e., Z_{it}) is parameterized; and since the unknown function f is in general nonlinear, the above model is also called a partially linear panel data model. As a middle course between parametric and nonparametric extremes, semiparametric models are very appealing for its flexibility to balance between precision and robustness.

On some occasions, the nonparametric component $f(X_{it})$ is treated as a nuisance term and the functional form f need not be estimated. This may be reasonable if X_{it} only performs a "controlling" role and there is little ambiguity toward the relationship between the variables of interest. However, if X_{it} are indeed among the variables of interest, and when the theoretical predictions regarding the relationship between X_{it} and Y_{it} are ambiguous or controversial, the estimation of f may become the central objective. For example, the classical research on Kuznets curve that investigates the relationship between income and inequality centers on the estimation of a nonlinear (supposedly inverted-U shape) function (See, e.g., Banerjee and Duflo (2003)). This is also true with the recent literature on environmental Kuznets curve which examines the relationship between income and pollution level (See, e.g., Millimet, List, and Stengos (2003)). Other important empirical topics such as the estimation of the Engel curve, production function, earnings-age profile, and so on, also boil down to estimating a possibly nonlinear relationship between two economic variables, controlling for other factors. Of course, the nonparametric estimation may or may not be the end of analysis. But even as an initial analysis before parameterization, a robust and reasonably accurate estimation of the nonlinear component is essential for the success of future modeling and analysis.

In recent years, indeed, a lot of research has been done in estimating semiparametric panel data models, the nonlinear component treated as a term of interest. As in the case of linear panel data models, these efforts can be largely grouped into three categories, RE (Random Effects), FE (Fixed Effects), and FD (First Differencing), depending on how they treat the unobserved effects.

The RE school treats the unobserved effects as exogenous and puts them into the residual. Li and Stengos (1996), following Robinson (1988), develop a root-N consistent IV estimator for the estimation of β, assuming that α_i is uncorrelated with other covariates. Although their focus is on the linear part, the nonlinear part can be easily estimated using a second-stage kernel regression. This simple approach, which we may call the pooled estimation, does not take into account the special covariance structure of the composite error, $\alpha_i + \varepsilon_{it}$. However, Ruckstuhl, Welsh, and Carroll (2000) show that the pooled estimator has better asymptotic properties than the quasi-likelihood estimator which takes into account the covariance structure. Ruckstuhl, Welsh, and Carroll (2000) propose an alternative two-step estimator, which we may call LL-RE (Local Linear Random Effects) estimator, that also accounts for the covariance structure. They show that LL-RE may achieve smaller asymptotic variance than the pooled estimator does, but the bias is in general incomparable. Recently Su and Ullah (2007) generalize the two-step estimator to the multivariate case.

As in linear panel data models, the FE approach treats the unobserved effects as dummy variables. Su and Ullah (2006) propose to estimate the nonlinear component by profile likelihood estimation, which boils down to a locally linear kernel smoothing, controlling for fixed effects by dummies. To use the usual panel data terminology, the estimator by Su and Ullah (2006) may be called LL-LSDV (Local Linear Least Squares with Dummy

Variables) estimator. Recently, Mammen, Støve, and Tjøstheim (2009) develop an iterative procedure based on smooth backfitting algorithm for estimating additive panel data models, treating unobserved effects as dummy variables. Their procedure may be directly applied to semiparametric panel data models with fixed effects.

Finally, the FD approach imposes no assumptions on unobserved effects and eliminates unobserved effects by first differencing. However, this transformation leaves us a structure of the following form, $m(X_{it}, X_{i,t-1}) = f(X_{it}) - f(X_{i,t-1})$, making the recovery of f difficult, even after successful estimation of the linear part. Henderson, Carroll, and Li (2008) solve this problem by using an iterative backfitting procedure based on the first-order condition of a profile likelihood criterion. Alternatively, Lee and Mukherjee (2008) propose to first approximate f using a local Taylor series expansion before taking first differencing (or, alternatively, within transformation). However, the function itself is eliminated from consideration by first differencing, and hence their approach deals only with the first derivative of f.

In this paper we propose a noniterative method that is based on marginal integration. We observe that $m(u, v)$ is an additive function and that marginal integration of an estimate of m recovers f. The technique of marginal integration, under the name of "projection", is introduced by Auestad and Tjøsstheim (1991) in the context of time series regression. A more systematic treatment is given in Tjøsstheim and Auestad (1994). This method is independently invented by Newey (1994) and Linton and Nielsen (1995) in the context of i.i.d. cross-section regression. Linton and Härdle (1996) generalize the method to deal with additive regression with known links. For important developments of this technique, see Masry and Tjøsstheim (1997), Linton (1997), Fan, Härdle, and Mammen (1998), Kim, Linton, and Hengartner (1999), Cai and Masry (2000), and Hengartner and Sperlich (2005).

The estimator we develop is conceptually simple, hence it is straightforward to analyze its statistical properties. Indeed, we derive the asymptotic distribution of our estimator using nothing more than standard arguments in multivariate kernel regression. Furthermore, the computational procedure for our estimator is noniterative, hence it is easy to implement in practice and also fast enough for finite sample investigations using Monte Carlo simulations. The disadvantage of our approach, however, is some efficiency loss during the unconstrained nonparametric estimation of $m(u,v)$. In particular, the information in the antisymmetric structure of $m(u,v)$ is lost. As a preliminary attempt, we propose a sample augmentation technique to make use of the structure. Although simulation results indicate some success for this technique, we are currently unable to validate it theoretically. Hence our asymptotic theory does not rely on sample augmentation.

The rest of the paper is organized as follows. The next section presents the model, describes our methodology, and gives asymptotic properties of our estimators. We first consider panel data models with only individual effects, then we extend our methodology to treat two-way effects models. Section 3 presents some Monte Carlo evidence on how our estimator behaves in the finite sample setting. All mathematical proofs are provided in the appendix.

2. The Model and Estimation

We consider the semiparametric (partially linear) panel data model in (1) which is reproduced here for convenience,

$$Y_{it} = \alpha_i + \beta' Z_{it} + f(X_{it}) + \varepsilon_{it},\ i = 1, \cdots, N,\ t = 0, 1, \cdots, T, \tag{2}$$

where $\beta \in \mathbb{R}^b$, $Z_{it} \in \mathbb{R}^b$, $X_{it} \in \mathbb{R}^d$, and all other variables are scalars. f is an unknown d-dimensional smooth function. Some or all elements in Z_{it} may be correlated with residual ε_{it}. And we allow for arbitrary correlation between the unobservable individual effect α_i and the regressors (X_{it}, Z_{it}). The individual effect may be called fixed effect if it is correlated with regressors or random effect if not. Finally, but importantly, we require $d < 4$ in this paper, considering the curse of dimensionality that the semiparametric model is invented to avoid, and the fact that it is extremely difficult to interpret f if $d \geq 4$.

One extension to the model in (2) is to introduce a time effect into the original model. The extended model, called two-way effects (individual and time effects) model, will be discussed later in the paper.

To estimate the model, we first take the FD (First Differencing) transformation of (2) across time t for each group i ,

$$\Delta Y_{it} = \beta' \Delta Z_{it} + f(X_{it}) - f(X_{i,t-1}) + e_{it}, \ i = 1, \cdots, N, \ t = 1, \cdots, T, \tag{3}$$

where $\Delta Y_{it} = Y_{it} - Y_{i,t-1}$, $\Delta Z_{it} = Z_{it} - Z_{i,t-1}$, and $e_{it} = \varepsilon_{it} - \varepsilon_{i,t-1}$. This first-differencing transformation eliminates the individual effects α_i. Throughout the paper, we assume:

Assumptions A

(1) (X_{it}, Z_{it}, e_{it}) are i.i.d. in i.

(2) For each i, X_{it} is strictly stationary with a well defined density p on a compact support $C \in \mathbb{R}^d$; and the marginal and joint densities of X_{it} are bounded from above and from zero.

(3) e_{it} is independent of (X_{it}), and $\mathbb{E}(e_{it}) = 0$, $\mathbb{E}(e_{it}^2) = \sigma_t^2$, $\mathbb{E}|e_{it}|^{4+\epsilon} < \infty$ for some $\epsilon > 0$.

A(1) is fairly standard for panel data models. The stationarity assumption in A(2) is stronger than necessary and is made for simplifying analysis. We may obtain similar asymptotic results if we assume that $(X_{it}, X_{i,t-1})$ admits a joint density that does not vary over t, which is not sufficient for stationarity. A(3) is fairly weak, allowing for serial correlation in ε_{it}. Indeed, our methodology works best if ε_{it} is a random walk. Finally, note that the covariance matrix of $e_i = (e_{i1}, ..., e_{iT})'$ is non-diagonal in general. Later in this section, we will discuss possibilities of using this fact to improve efficiency. We now turn to the estimation of the model and related issues in implementation.

2.1 The Linear Component

For the linear part parameterized by β, the transformed model in (3) is a special case of the model considered in Li and Stengos (1996). Let $u = (u_1, ..., u_d)'$ and $v = (v_1, ..., v_d)'$, and let $p_2(u, v)$ be the joint density of $(X_{it}, X_{i,t-1})$. The model in (3) implies

$$Y^*_{it} = \beta' Z^*_{it} + u_{it},$$

where $Y^*_{it} = \rho_{it}(\Delta Y_{it} - \mathbb{E}(\Delta Y_{it}|X_{it}, X_{i,t-1}))$, $Z^*_{it} = \rho_{it}(\Delta Z_{it} - \mathbb{E}(\Delta Z_{it}|X_{it}, X_{i,t-1}))$, $u_{it} = \rho_{it} e_{it}$, and $\rho_{it} = p_2(X_{it}, X_{i,t-1})$. Working with this density weighted equation enables us to avoid the random denominator problem typical in nonparametric kernel regression estimation.

Assuming there exists a vector of instrumental variables $W_{it} \in \mathbb{R}^b$, we may construct an IV estimator for β. Let $W^*_{it} = \rho_{it}(W_{it} - \mathbb{E}(W_{it}|X_{it}, X_{i,t-1}))$. And let the capital letters without subscripts denote the matrices of observations of the corresponding variable. More specifically, we denote $Z_i = (Z_{i1}, ..., Z_{iT})'$ and $Z \equiv (Z'_1, ..., Z'_N)'$. Throughout the paper we use the same matrixization and denote $Z = [Z'_{it}]$. Then we have an infeasible IV estimator $\hat{\beta} = (W^{*\prime} Z^*)^{-1} W^{*\prime} Y^*$, assuming that the term in parentheses is invertible.

To make the IV estimator feasible, we estimate ρ_{it} by $\hat{\rho}_{it} = 1/(NT)) \sum_{js} L_g(X_{it} - X_{js}) L_g(X_{i,t-1} - X_{j,s-1})$, where $L_g(u) = g^{-d} \prod_{j=1}^d l(u_j/g)$, l is a univariate kernel and g is the associated bandwidth. The conditional means are estimated by $1/(NT) \sum_{js} \xi_{it} L_g(X_{it} - X_{js}) L_g(X_{i,t-1} - X_{j,s-1})$, where ξ_{it} denotes ΔY_{it}, ΔZ_{it}, and W_{it}. Assuming that $\hat{W}^{*\prime} \hat{Z}^*$ is invertible, we obtain the following feasible IV estimator,

$$\hat{\beta} = (\hat{W}^{*\prime} \hat{Z}^*)^{-1} \hat{W}^{*\prime} \hat{Y}^*. \tag{4}$$

Let the class of kernels $\mathcal{K}_\nu$ and the function class $\mathcal{G}^\alpha_\varsigma$ be defined as Definition 1 and Definition 2 of Robinson (1988), respectively. Kernels in $\mathcal{K}_\nu$ are of order ν, and the functions in $\mathcal{G}^\alpha_\varsigma$ are ς-times partially differentiable with a Lipschitz-continuous remainder. α controls the moment properties of the remainder. In particular, the functions in $\mathcal{G}_\varsigma^\infty$ are bounded. For the root-N consistent estimation of β, we adapt the following assumptions from Li and Stengos (1996),

Assumptions B

(1) $p_2 \in \mathcal{G}_\varsigma^\infty$ for some constant $\varsigma \geq 1$, $f \in \mathcal{G}_\nu^{4+\epsilon}$, $\mathbb{E}(\Delta Z_{it}|X_{it}, X_{i,t-1}) \in \mathcal{G}_\nu^{4+\epsilon}$ for some $\epsilon > 0$ and positive integer ν with $\varsigma < \nu \leq \varsigma + 1$.

(2) There exists an IV vector $W_{it} \in \mathbb{R}^b$ such that W_{it} is i.i.d. in i, $\mathbb{E}(W_{it}^{4+\epsilon}) < \infty$, $\mathbb{E}(W_{it}|X_{it}, X_{i,t-1}) \in \mathcal{G}_\nu^{4+\epsilon}$, and $\mathbb{E}(e_{it}|W_{it}) = 0$ for all t. And $\Xi = \mathbb{E}W_{12}^* Z_{12}^{*}{}'$ is nonsingular.

(3) $l \in \mathcal{K}_\nu$, and as $N \to \infty$, $a \to 0$, and $Na^{4d} \to \infty$, and $Na^{4\nu} \to 0$.

The assumptions in B(2) are fairly standard for an IV vector. For a typical application with $d = 1$, B(3) is satisfied if we let $\nu = 2$ and choose a second-order kernel for l. B(1) is satisfied if p_2 has continuous partial derivatives and f is twice continuously differentiable. If $d > 1$, then we have to use a higher order kernel for l and f needs to have more derivatives, which is one form of the curse of dimensionality.

The assumption on p_2, which states that p_2 is bounded and at least first-order partially differentiable with a Lipschitz-continuous remainder, is stronger than that is made in Li and Stengos (1996), who only require Lipschitz continuity. This stronger assumption is made for estimating the nonlinear part and is not required for the following theorem, which is proved in Li and Stengos (1996).

Theorem 1. Under assumptions A and B, we have

$$\sqrt{N}(\hat{\beta} - \beta) \to_d N(0, \Xi^{-1}\Psi(\Xi^{-1})'), \tag{5}$$

where $\Psi = 1/T^2 \sum_t \sum_s \mathbb{E}\left(e_{1t}e_{1s}W_{1t}^* W_{1s}^{*\prime}\right)$. Furthermore, we can consistently estimate Ψ and Ξ by plugging in the estimates for each term.

2.2 The Nonlinear Component

The major contribution of this paper is in estimating the unobserved function $f(\cdot)$. To simplify the notations, we denote $R_{it} = \Delta Y_{it} - \beta' \Delta Z_{it}$. We rewrite (3) as

$$R_{it} = m(X_{it}, X_{i,t-1}) + e_{it},\ i = 1, \cdots, N,\ t = 1, \cdots, T, \tag{6}$$

where $m : \mathbb{R}^{2d} \to \mathbb{R}$ is an additive function:

$$m(u, v) = f(u) - f(v),\ u, v \in \mathbb{R}^d. \tag{7}$$

Obviously, $m(u, v) = -m(v, u)$. Hence $m(u, v)$ is antisymmetric.

We may easily estimate $m(u, v)$ using multivariate kernel smoothing methods. The popular estimators are Nadaraya-Watson (Nadaraya (1964); Watson (1964)), Gasser-Müller (Gasser and Müller (1984)), and local linear (Stone (1977); Cleveland (1979); Fan (1992); Ruppert and Wand (1994)), among others. In principle, each of these approaches would serve our purpose. We will prove asymptotic properties only for the local linear method, which includes Nadaraya-Watson as a special case.

Let $K(u) = \prod_{i=1}^d k(u_i)$, where k is a univariate second-order symmetric kernel. And denote $K_H(u) = |H|^{-1}K(H^{-1}u)$, where $H = \text{diag}(h_1, ..., h_d)$ is a diagonal bandwidth matrix. The local linear estimator of $m(u, v)$ solves the following problem for α,

$$\min_{\alpha,\gamma_1,\gamma_2} \sum_{i=1}^N \sum_{t=1}^T \left[R_{it} - \alpha - \gamma_1'(X_{it} - u) - \gamma_2'(X_{i,t-1} - v)\right] K_H(X_{it} - u) K_H(X_{i,t-1} - v). \tag{8}$$

It is well known that the problem in (8) is a weighted least square problem. Let $R = [R_{it}]$ and $\Gamma = [1\ (X_{it} - u)'\ (X_{i,t-1} - v)']$ (a $1 + 2d$ column matrix); and $W = \text{diag}[K_H(X_{it} -$

$u)K_H(X_{i,t-1}-v)]$. Note that we suppress the dependence on u and v of Γ and W. Assuming that $\Gamma'W\Gamma$ is invertible, (8) has a solution for $\hat{\alpha}$ (hence $\hat{m}(u,v)$),

$$\hat{m}(u,v) = \hat{\alpha} = \iota'(\Gamma'W\Gamma)^{-1}\Gamma'WR, \tag{9}$$

where $\iota = (1, \overbrace{0, ..., 0}^{2d})'$. Note that $\hat{\gamma}_1$ supplies an estimate of $\frac{\partial f}{\partial u}(u)$. If our primary goal is estimating the partial derivatives, we may stop here. While it is possible to recover f up to an additive constant from partial derivatives by numerical integration, we are obviously not satisfied with this solution; the reason is that although the asymptotic properties of $\hat{\gamma}(u)$ are well known, those of $\int^x \hat{\gamma}(u)du$ are not. And it is conjectured that statistical error may accumulate through numerical integration. We focus on $\hat{m}(u,v)$, which is the stepstone for estimating f.

We proceed to estimate $f(\cdot)$ by marginally integrating $\hat{m}(u,v)$,

$$\hat{f}(u) = \int_C \hat{m}(u,v)q(v)dv, \tag{10}$$

where q is a predetermined density function. For model identification, we follow Hengartner and Sperlich (2005) and assume $\int_C f(u)q(u)du = 0$. Note that this condition reduces to $\mathbb{E}(f(X_{it})) = 0$ as in Linton and Härdle (1996) if we take $q = p$. It is easy to see the rationale behind (10),

$$\int_C m(u,v)q(v)dv = \int_C (f(u) - f(v))q(v)dv = f(u).$$

Note that, in practice, it is not always necessary to impose the identification condition. If the condition does not hold, f can still be identified up to an additive constant.

We may implement the marginal integration in (10) by numerical integration methods such as Simpson's or Trapezoidal rules. Let the number of evaluation points on each "nuisance dimension" (the dimensions in v) be S. Then for the estimation of every point in the dimension of interest, we need compute S^d weighted least squares, each of which requires

$O(N)$ operations. This may become a practical concern and calls for discretion over the balance between numerical accuracy and computational time.

An alternative method of calculating the marginal integration is to generate i.i.d. samples (X_k^*, $k = 1, ..., n$) from the distribution q and to construct $\hat{f}_{mc}(u) = \frac{1}{n}\sum_{k=1}^{n}\hat{m}(u, X_k^*)$. If n is large enough, $\hat{f}_{mc}$ approximates $\hat{f}$ well. We may well choose $q(\cdot)$ to be the density function of $X_{i,t}$. In this case we may use the sample version of (10),

$$\hat{f}_s(u) = \frac{1}{N(T+1)}\sum_{i=1}^{N}\sum_{t=0}^{T}\hat{m}(u, X_{i,t}). \tag{11}$$

Asymptotically, this estimator behaves the same as (10) when q is the density of $X_{i,t}$. We assume:

Assumptions C

(1) q is a bounded density function, defined on the compact support C, twice continuously differentiable, and $\int_C f(u)q(u) = 0$.

(2) k is a second-order kernel that is positive, bounded, symmetric, and defined on the support C.

(3) $H = H_0 N^{-1/(4+d)}$, where H_0 is a diagonal matrix with positive constants on the diagonal.

These assumptions are fairly standard in literature. Let $\hat{f}$ be defined as in (10). And denote $\varphi(k) = \int k(u)^2 du$, $\mu_2(k) = \int u^2 k(u) du$, $\mathcal{D}_f = \frac{\partial f}{\partial u}$, and $\mathcal{H}_f = \frac{\partial^2 f}{\partial u \partial u'}$. The following theorem is the major result of this paper.

Theorem 2. Let u be an interior point of supp(p) and let Assumptions A, B, and C hold. Given a fixed T and as $N \to \infty$, we have

$$N^{2/(4+d)}(\hat{f}(u) - f(u)) \to_d N(B(u), V(u)), \tag{12}$$

where

$$B(u) = \frac{1}{2}\mu_2(k)\left[\operatorname{tr}\left(H_0^2\mathcal{H}_f(u)\right) - \int_C \operatorname{tr}\left(H_0^2\mathcal{H}_f(v)\right) q(v)dv\right], \tag{13}$$

and

$$V(u) = \frac{\varphi^d(k)\bar{\sigma}^2}{T|H_0|}\left(\int_C \frac{q^2(v)}{p_2(u,v)}dv\right), \tag{14}$$

where $\bar{\sigma}^2 = 1/T\sum_t \sigma_t^2$.

The proof is given in the Appendix. Here we make a number of remarks.

Remark 1: If we impose i.i.d. condition on X_{it} across t as well as i, and if we take $q = p$, the asymptotic variance would take an even simpler form:

$$V(u) = \frac{\varphi^d(k)\bar{\sigma}^2}{T|H_0|p(u)}\ .$$

Remark 2: We may consistently estimate $V(u)$ by

$$\hat{V}(u) = N^{-2(2+d)/(4+d)}T^{-2}\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{e}_{it}^2\hat{\theta}_{it}^2, \tag{15}$$

where $\hat{\theta}_{it} = 1/\tilde{n}\sum_{j=1}^{\tilde{n}} w_j(u, X_j^*)$ with

$$w_j(u, X_j^*) = \frac{K_H(u - X_{it})K_H(X_j^* - X_{i,t-1})}{\sum_{i=1}^{N}\sum_{t=1}^{T} K_H(u - X_{it})K_H(X_j^* - X_{i,t-1})},$$

where $(X_j^*,\ j = 1, ..., \tilde{n})$ are drawn from the distribution with density q. If $q = p$, we may simply use (X_{it}) in place of (X_j^*).

Remark 3: We can construct confidence bands for f using the asymptotic result. Denoting the $(1 - \frac{a}{2})$ quantile of the standard normal distribution with $z_{1-\frac{a}{2}}$, we get the $1 - a$

confidence bands,

$$\left[\ \hat{f}(u) - N^{-2/5}\big(B(u) + z_{1-\frac{a}{2}} V^{1/2}(u)\big)\,, \qquad \hat{f}(u) - N^{-2/5}\big(B(u) - z_{1-\frac{a}{2}} V^{1/2}(u)\big)\ \right].$$

With some under-smoothing (i.e., $h = o(N^{-1/5})$), we may ignore the bias term and use only the asymptotic variance to construct confidence bands for $f(u)$. At the cost of computation time, we may also improve the quality of confidence bands by bootstrap. See Härdle and Marron (1991) for more details.

Remark 4: An optimal H_0 may be found by minimizing AMISE (Asymptotic Mean Integrated Squared Error). This is best illustrated by considering the case of $d = 1$, when $H_0 = h_0$ is a positive scalar. We minimize

$$\text{AMISE}(h_0) = \int_C \big(B^2(u) + V(u)\big)\, du.$$

Then we obtain an optimal h_0:

$$h_0 = \left(\frac{\varphi(k)\bar{\sigma}^2}{\mu_2^2(k)T}\frac{\vartheta_2}{\vartheta_1}\right)^{\frac{1}{5}},$$

where

$$\vartheta_1 = \int_C (f''(u) - \int_C f''(v)dv)^2 du, \quad \text{and} \quad \vartheta_2 = \int_C \int_C q^2(v) p_2^{-1}(u,v) du dv.$$

Replacing the unknown quantities ($\bar{\sigma}^2$, ϑ_1, and ϑ_2) with their estimates, we obtain a plug-in bandwidth selector. And the above strategy can be easily extended to multivariate case.

We may also choose bandwidths using delete-one CV (Cross-Validation), generalized CV (Craven and Wabha, 1979), or model selection procedures such as Mallows'(1973) C_p and C_L procedures. See Li (1987) and Andrews (1991) for the asymptotic properties of these selectors.

Remark 5: If the Nadaraya-Watson kernel smoothing is used for estimating $m(u, v)$, the asymptotic variance of $\hat{f}(u)$ would be the same as in (14), but the asymptotic bias takes the following form:

$$\begin{aligned} B(u) &= \mu_2(k)\left[\frac{1}{2}\mathrm{tr}\left(H_0^2\mathcal{H}_f(u)\right) - \frac{1}{2}\int_C \mathrm{tr}\left(H_0^2\mathcal{H}_f(v)\right)q(v)dv \right. \\ &\quad \left. + \mathcal{D}_f'(u)H_0^2\int\frac{\partial\log(p_2(u,v))}{\partial u}q(v)dv - \mathcal{D}_f'(v)H_0^2\int\frac{\partial\log(p_2(u,v))}{\partial v}q(v)dv\right]. \end{aligned}$$

Remark 6: If f is twice partially differentiable, our estimator achieves the best convergence rate possible. However, we require higher-order differentiability of f to estimate the linear component if $d \geq 2$. Recall that in Assumption B, we require $f \in \mathcal{G}_\nu^{4+\epsilon}$ and $\nu > d$. The optimal rate is thus $N^{-\nu/(2\nu+d)}$, higher than our estimator achieves. To achieve the optimal rate, we need higher-order locally polynomial smoothing to reduce bias. We choose not to do so for the attractive properties of local linear estimators (Fan and Gijbels (1992), Ruppert and Wand (1994)). And if there is no linear part, we only need twice differentiability for f, regardless of d. Then the optimal rate of kernel regression estimator is $N^{-2/(4+d)}$, which is achieved by our estimator.

Remark 7: Finally, the form of the asymptotic variance $V(u)$ suggests when our method might fail. That is when X_{it} is accurately predictable by $X_{i,t-1}$. In this case, if we write $p_2(u, v) = p(u|v)p(v)$, $p(u|v)$ would be close to zero except in a small neighborhood of v, hence a large $V(u)$. This happens, for example, if X_{it} is highly persistent in t.

2.3 Efficiency Issues

Now we discuss two issues related with the efficiency of our estimator. The first is concerned with how we may use the covariance matrix of $e_i = (e_{i1}, ..., e_{iT})'$, which is in general not diagonal. The second issue is concerned with how we may use the antisymmetric property of $m(u, v)$.

Covariance Structure

Let Σ denote the covariance matrix of e_i, which is diagonal only when ε_{it} is a random walk. Indeed, if ε_{it} is i.i.d. across t, Σ would be a tridiagonal matrix with 2's on the main diagonal and -1's on the sub-diagonal. Our estimator of $m(u,v)$ does not make use of this structure, and therefore it is possible to be improved.

One way to take advantage of the covariance structure is to estimate $m(u,v)$ by the quasi-likelihood method proposed by Severini and Staniswalis (1994). The quasi-likelihood estimator of $m(u,v)$ is the intercept in the solution of θ in

$$\Gamma' \left(I_N \otimes \Sigma^{-1}\right) \Gamma W (Y - \Gamma\theta) = 0,$$

where I_N is an N-dimensional identity matrix and $\otimes$ denotes Kronecker product.

We may also employ a two-stage procedure similar with Ruckstuhl, Welsh, and Carroll (2000) and Su and Ullah (2007). Let $\eta_{it} = (X_{it}, X_{i,t-1})$ and $\eta_i = (\eta_{i1}, ..., \eta_{iT})'$. The two-stage estimator is based on the following identity

$$\tau\Sigma^{-\frac{1}{2}}R_i - \left(\tau\Sigma^{-\frac{1}{2}} - I_T\right) M(\eta_i) = M(\eta_i) + \tau\Sigma^{-\frac{1}{2}}e_i,$$

where τ is a positive constant, $R_i = (R_{i1}, ..., R_{i,T})'$, and $M(\eta_i) = (m(\eta_{i1}), ..., m(\eta_{iT}))'$. The first step obtains a local linear estimator of m (ignoring the covariance structure) and an estimator of Σ. Then we replace unknown quantities on the left with their estimates and run a second-stage local linear regression.

It is not clear, however, that the above treatments may improve the accuracy of our estimator. In a similar context, Ruckstuhl, Welsh, and Carroll (2000) show that the asymptotic variance of the quasi-likelihood estimator is of higher order than that of the simple "pooled estimator", that is, the estimator ignoring covariance structure. The two-stage estimator may achieve a smaller asymptotic variance with some appropriate choice of τ, but the bias term is complicated and it is difficult to be compared with that of the pooled estimator.

In explaining why the simple pooled estimator performs so well asymptotically, Ruckstuhl, Welsh, and Carroll (2000) point out that the covariance structure is a global property of the residual which may not be important for methods that act locally in the covariate space.

Sample Augmentation

The second issue is concerned with the structure of $m(u, v)$, which is by definition antisymmetric. We do not make use of this information in estimating $m(u, v)$ using unconstrained kernel smoothing methods. Hence our estimator suffers from some efficiency loss. One way to impose the antisymmetric structure on $m(u, v)$ is to generate another copy of data that is antisymmetric to the original data, and to use both the original data and their antisymmetric mirror in kernel smoothing.

To see this, recall that for any triplet $(R_{it}, X_{i,t}, X_{i,t-1})$, we have

$$\mathbb{E}\left(R_{it}|X\right) = f(X_{i,t}) - f(X_{i,t-1}),$$

since ε_{it} is assumed to be independent of X. Then the triplet $(-R_{it}, X_{i,t-1}, X_{i,t})$ must also satisfy the above equation and can be included in the sample. We may call this practice "sample augmentation". However, our development of asymptotic theory does not extend easily to the augmented sample, which obviously contains nonindependent (antisymmetric) observations. Limited simulation results show that sample augmentation may considerably improve our estimator. Undoubtedly it calls for further research into the general theory of imposing prior structures on multivariate nonparametric estimation (not just the antisymmetric structure in our case) by some type of sample augmentation.

2.4 Two-Way Effects Models

For some applications it is desirable to include a time effect in model (2), controlling for unobserved time-varying factors that are common across individuals. Then we have a two-way effects panel data model. To focus on the main idea, we consider the following simplified model,

$$Y_{it} = \alpha_i + \phi_t + f(X_{it}) + \varepsilon_{it},\ i = 0, 1, \cdots, N,\ t = 0, 1, \cdots, T, \tag{16}$$

where (ϕ_t) represent time effects and X_{it} is univariate[2].

The model in (16) contains time effects (ϕ_t) that do not disappear with the first differencing procedure described above. To eliminate ϕ_t, we take another difference across individual i for each time t. Thus we have

$$\Delta^2 Y_{it} = [f(X_{it}) - f(X_{i,t-1}) - f(X_{i-1,t}) + f(X_{i-1,t-1})] + e_{it}, \tag{17}$$

where $i = 1, \cdots, N$, $t = 1, \cdots, T$, and $e_{it} = \varepsilon_{it} - \varepsilon_{i,t-1} - \varepsilon_{i-1,t} + \varepsilon_{i-1,t-1}$. Setting $R_{it} = \Delta^2 Y_{it}$ and $v = (v_1, v_2, v_3)'$, we rewrite (17) as

$$R_{it} = m(X_{it}, X_{i,t-1}, X_{i-1,t}, X_{i-1,t-1}) + e_{it},\ i = 1, \cdots, N,\ t = 1, \cdots, T, \tag{18}$$

where $m : \mathbb{R}^4 \to \mathbb{R}$ is an additive function that satisfies

$$m(u, v) \equiv m(u, v_1, v_2, v_3) = f(u) - f(v_1) - f(v_2) + f(v_3). \tag{19}$$

We estimate m using local linear smoothing. The form of $\hat{m}$ is the same as in (9), but the definition of each term should be modified. Let $\tilde{N} = N$ if N is even, else $\tilde{N} = (N-1)/2$. We set $\iota = (1, 0, 0, 0, 0)'$. Γ is now a 5-column matrix $[1, (X_{2i,t} - u), (X_{2i,t-1} - v_1), (X_{2i-1,t} - v_2), (X_{2i-1,t-1} - v_3)]_{i=1,2,\ldots,\tilde{N}, t=1,2,\ldots,T}$. The diagonal elements of W are now $k_h(X_{2i,t} - u)k_h(X_{2i,t-1} - v_1)k_h(X_{2i-1,t} - v_2)k_h(X_{2i-1,t-1} - v_3)$, where $k_h(u) = k(u/h)/h$ and k is a second-order symmetric kernel. Finally, $R = [R_{2i,t}]$.

Note that in these definitions we only use non-overlapped cross-sections. For example,

[2] For a multivariate X_{it} with dimension d, our methodology would involve nonparametric estimation of a $4d$-dimensional pilot function. The curse of dimensionality would render the two-way effects models with $d \geq 2$ practically irrelevant.

we do not include $[(X_{2i-1,t}-u),(X_{2i-1,t-1}-v_1),(X_{2i-2,t}-v_2),(X_{2i-2,t-1}-v_3)]$ in the definition of Γ. So in a sense we have dropped half of the sample. This is to deal with the technical issue arising from the second differencing transformation across individuals. After this transformation, the observations $U_{it}\equiv(R_{it},X_{it},X_{i,t-1},X_{i-1,t},X_{i-1,t-1})$ are no longer i.i.d. across i, making some of the well known asymptotic results of local linear estimators unapplicable. Undoubtedly this would affect the efficiency of the estimator and should not be rigidly followed in practical applications.

We then estimate $f(u)$ by

$$\hat{f}(u)=\int\hat{m}(u,v)q(v_1)q(v_2)q(v_3)dv_1dv_2dv_3, \tag{20}$$

where $q(v)$ is a predetermined univariate density function. As in the previous section, we may implement (20) by numerical integration or sample integration using actual or simulated data.

Being 4-dimensional, m is endowed with a more complex structure of symmetry or antisymmetry. Specifically, we have

$$\begin{aligned} m(u,v_1,v_2,v_3) &= -m(v_1,u,v_3,v_2),\\ m(u,v_1,v_2,v_3) &= -m(v_2,v_3,u,v_1),\ \text{and}\\ m(u,v_1,v_2,v_3) &= m(v_3,v_2,v_1,u).\end{aligned}$$

We may again use this structural information for sample augmentation to improve efficiency. For the development of asymptotic theory, we assume,

Assumptions D

(1) Both f and q are defined on the compact support C, twice continuously differentiable, and $\int_C f(u)q(u)=0$.

(2) For each i, all joint densities of (X_{it}) are continuously differentiable.

(3) k is a bounded and symmetric second order kernel on C.

(4) $h = h_0 \tilde{N}^{-1/5}$, where h_0 is a positive constant.

Let $p_4(u, v_1, v_2, v_3)$ denote the joint density of $(X_{it}, X_{i,t-1}, X_{i-1,t}, X_{i-1,t-1})$. The following theorem gives the asymptotic properties of the estimator $\hat{f}$ defined in (20).

Theorem 3 Let u be an interior point of supp(p). If Assumptions A and D hold, then

$$\tilde{N}^{2/5}(\hat{f}(u) - f(u)) \to_d N(B(u), V(u)), \tag{21}$$

where

$$B(u) = \frac{1}{2} h_0^2 \mu_2(k) \left(f''(u) - \int_C f''(s) q(s) ds \right), \tag{22}$$

$$V(u) = \frac{\varphi(k)\bar{\sigma}^2}{h_0 T} \left(\int \frac{q^2(v_1) q^2(v_2) q^2(v_3)}{p_4(u, v_1, v_2, v_3)} dv_1 dv_2 dv_3 \right) \tag{23}$$

The proof is a straightforward extension of the proof for Theorem 2 and hence omitted. Note that, unlike the individual effect α_i, the time effect ϕ_t can be consistently estimated, assuming that $\phi_0 = 0$. Let $\Delta Y_{it} = Y_{it} - Y_{i,t-1}$ and $\Delta\phi_t = \phi_t - \phi_{t-1}$, $t \geq 1$. We can consistently estimate $\Delta\phi_t$ by

$$\widehat{\Delta\phi_t} = \frac{1}{N} \sum_{i=1}^{N} (\Delta Y_{it} - (\hat{f}(X_{it}) - \hat{f}(X_{i,t-1})).$$

3. Simulations

In this section we use simulations to answer the following questions: does our estimator perform reasonably well in finite samples under the following settings: (1) when (α_i) are random effects; (2) when (α_i) are fixed effects; (3) and when (X_{it}) is persistent (but still stationary) over time?

3.1 The Setup

We consider the following data generating process (DGP),

$$Y_{it} = \alpha_i + f(X_{it}) + \sigma\varepsilon_{it},\ i = 1, ..., N,\ t = 1, ..., T, \tag{24}$$

where X_{it} is a scalar random variable; ε_{it} is an i.i.d. $N(0, 1)$ random variable; and $f(\cdot)$ is a pre-specified function to be estimated. And we experiment with two specifications of α_i,

(a) Random Effects (RE) : α_i is i.i.d. $N(0, 4)$, independent of (X_{it}), which are i.i.d. uniformly distributed between $[-2, 2]$.

(b) Fixed Effects (FE): α_i is i.i.d. $N(0, 4)$ dependent on (X_{it}); the dependence is imposed by generating X_{it} by $X_{it} = \alpha_i/2+U_{it}$, where U_{it} is i.i.d. uniformly distributed between $[-2, 2]$.

We consider the following functional forms for f:

(1) $f_1(x) = -1/2x^2$,

(2) $f_2(x) = x\cos(\pi x)$,

(3) $f_3(x) = x + 2\exp(-16x^2)$,

(4) $f_4(x) = \sin(2x) + 2\exp(-16x^2)$.

f_1 is of inverted U shape, which is often seen in empirical economics. f_2 is used in Linton and Jacho-Chaávez (2009), and f_3 and f_4 are used in Fan and Gijbels (1992). We use these familiar functional forms to facilitate comparisons in literature. Throughout the simulations, we estimate these function on the support $[-2, 2]$.

3.2 A Graphic Illustration

Figure 1 plots three consecutive FD estimates of each function. The estimated and the true functions are shifted so that they all integrate to zero. We implement the FD estimator using the sample version of marginal integration in (11) with the standard normal kernel. We use the plug-in bandwidth described in Remark 4. And we choose $T = 5$ and $N = 50$, and take α_i to be random effects.

It can be seen that these estimates trace the true functions (bold solid lines) well, even in locations near the boundary. For f_3 and f_4, there is some under-smoothing in the bump area. Recall that the plug-in bandwidth minimizes estimated AMISE, which is a global distance metric. For functions such as f_3 and f_4, it may be better to use bandwidth $h(x)$ that is a function of x and minimizes some local distance metric.

It should be noted that Figure 1 is only of illustrative purpose. We now turn to repeated experiments for a more conclusive view of how our estimator performs in finite samples.

3.3 Comparative Performance

In the repeated experiments, we first compare our estimator with the estimator proposed in Ruckstuhl, Welsh, and Carroll (2000) and Su and Ullah (2007), which works for the random-effects specification; and that proposed in Su and Ullah (2006), which is designed for the fixed-effects specification. As mentioned in the introduction, we may call the former method LL-RE (Local Linear Random Effects) and the latter LL-LSDV (Local Linear LSDV). We call our method FD (First Differencing).

We fix $T = 5$ and examine the finite sample performance of FD, LL-RE, and LL-LSDV when N is 50 and 100. We experiment with both low-noise level ($\sigma = 0.5$) and high-noise level ($\sigma = 1$). We do not impose identification condition in the simulations. Hence f is only identified up to an additive constant. We define the ISE (Integrated Square Error) of an estimate $\hat{f}$ by

$$\mathrm{ISE}_h(\hat{f}) = \int (\hat{f}^*(x) - f^*(x))^2 dx,$$

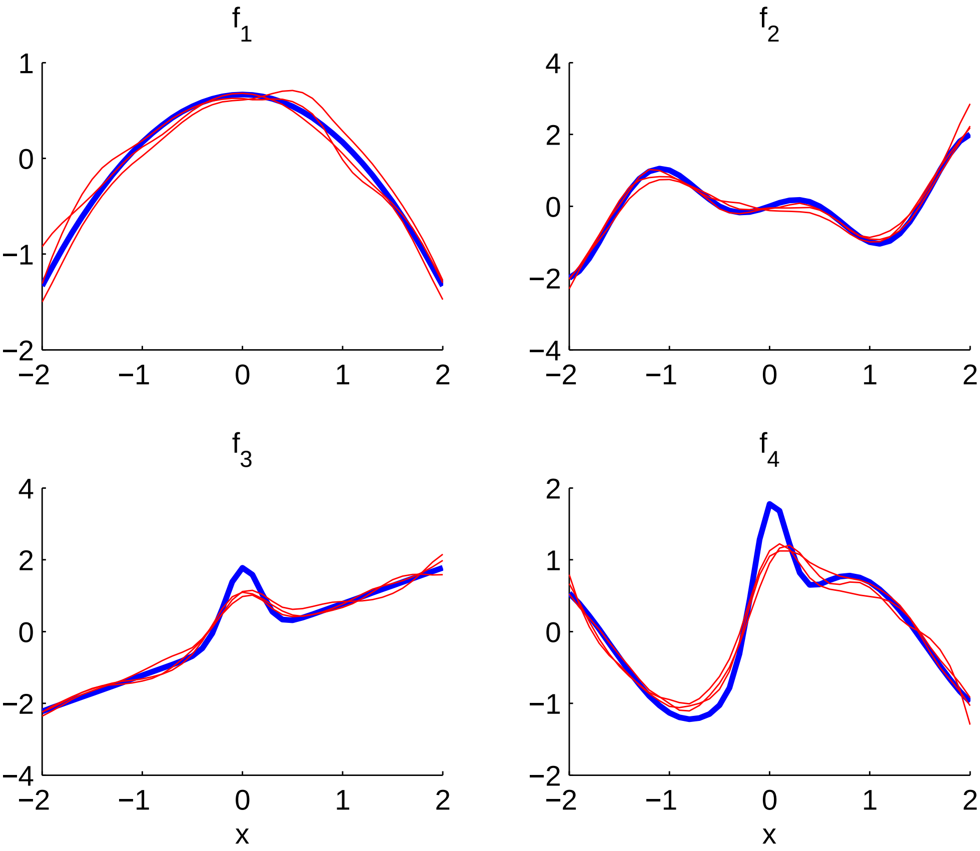

Figure 1: An Illustration of FD Estimates. In each diagram, the bold line is the true function and the three thin lines are FD estimates from three random experiments. f_1, f_2, f_3, and f_4 are defined in the text.

where $f^*(x) = f(x) - \int_a^b f(x)dx/(b-a)$. f^* is the true function shifted by a constant such that f^* integrates to zero. $\hat{f}^*$ is similarly defined. The ISE obviously depends on the choice of bandwidth, hence the subscript. For better comparisons, we use a pre-specified set of bandwidths. More specifically, we vary h from 0.1 to 0.6, with a small step size 0.03 within $[0.1, 0.25]$ and a bigger step size 0.05 in the remaining.

We repeat our experiment 500 times. Taking average of the ISE_h, we obtain an MISE_h for each estimator. Figure 2 compares the logarithm of MISE_h. The first two columns correspond to RE experiments and the third and the fourth columns correspond to FE. The first two rows correspond to f_1, the third and the fourth rows correspond to f_2, and so on. The odd rows correspond to $N = 50$ and the even rows correspond to $N = 100$. The odd columns correspond to $\sigma = 0.5$ and the remaining columns correspond to $\sigma = 1$.

We make the following observations from Figure 2. First, overall, FD works well for both RE and FE specifications. It can be seen that when the underlying DGP is RE, FD is a close competitor of LL-RE. In many cases, FD may even outperform LL-RE. And when the underlying DGP is FE, FD is a close competitor of LL-LSDV. In some cases, FD may also outperform LL-LSDV. Second, FD compares less favorably with its competitors when signal-noise-ratio is low (big σ). This may be explained by the fact that ε_{it} is generated as an i.i.d. $N(0, \sigma^2)$ noise and the first differencing transformation results in a residual term e_{it} with variance $2\sigma^2$. Third, when the bandwidth is very small, FD is unreliable. This indicates that we should worry more about the variance than the bias of our estimator. Finally, we point out the different behavior of MISE_h of FD for RE and FE specifications may be due to the way we generate X_{it}. If the underlying DGP is RE, X_{it} has a compact support $[-2, 2]$. But if the DGP is FE, many observations of X_{it} may fall outside the support, affecting the performance of every estimator.

Table 1 reports the median and the SD (Standard Deviation) of the smallest ISE of each estimator with its most advantageous bandwidth. That is, the median and the SD calculated from $\{\text{ISE}_{j,h_j^*},\ j = 1, ..., 500)\}$ for each each estimator, where the subscript j

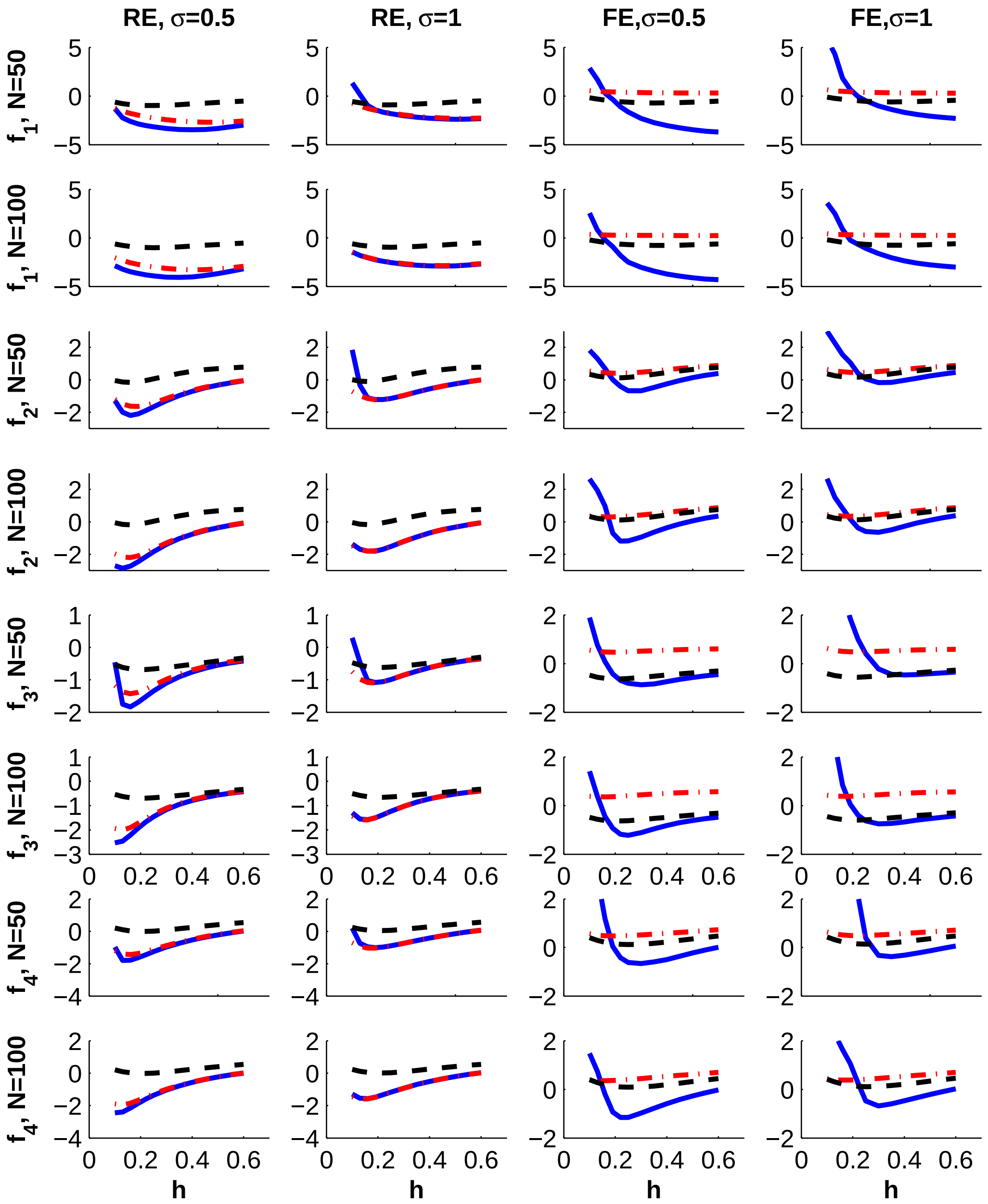

Figure 2: Simulation Results on Comparative Performance. The x-axis is bandwidth h and the y-axis is $\log(\mathrm{MISE}_h)$. In each diagram, the solid line : FD (First Differencing), the dot-dashed line : LL-RE (Local Linear Random Effects), and the dashed line : LL-LSDV (Local Linear LSDV). f_1, f_2, f_3, and f_4 are defined in the text. $T = 5$, and the number of repetition is 500. More details are in the text.

denotes each repetition of simulation and h_j^* is the bandwidth that, among all pre-specified bandwidth, achieves the minimal ISE. Hence Table 1 compares the best performance of each estimator.

It can be seen that the median (best) performance tells a similar story with what is reported in Figure 2, which compares the mean performance across a spectrum of bandwidth. Furthermore, the results on the SD of ISE reassure us that the dispersions of ISE around the mean are reasonable. Hence, with an appropriately chosen bandwidth, our estimator may be safe for practical applications. Finally, we may check that the square root of median ISE decreases at roughly a rate of $N^{-2/5}$, consistent with what is suggested by Theorem 2 for $d = 1$.

3.4 Experiments on Persistence

Now we consider the case when (X_{it}) is persistent. We generate zero-mean AR time series as follows,

$$X_{i,0} \sim N(0,2), \quad \text{and} \quad X_{it} = aX_{i,t-1} + \eta_t,$$

where the $a \in [0,1)$ controls the persistence level and $\eta_t \sim$ i.i.d. $N(0, 2(1-a^2))$. Hence for each i, X_{it} is strictly stationary with marginal distribution $N(0,2)$. α_i is generated as random effect, that is, $\alpha_i \sim$ i.i.d. $N(0,4)$ independent of X. We use the plug-in bandwidth described in Remark 4 following Theorem 2. We vary a from 0 to 0.95. The number of repetitions is set to be 500, and the mean, the median, and the SD (Standard Deviation) of the ISE's of FD estimator are reported in Table 2. We choose to report results from the experiments on f_2. The results from other functional forms are similar.

Table 2 shows, not surprisingly, that our estimator is unstable at high persistence levels. Under mild persistence ($a \leq 0.6$), the mean, the median, and the SD of ISE still remain in a reasonable range. This is not the case when $a \geq 0.8$. Note that the median's are generally lower than the mean, indicating there are instances where our estimator is way off the mark, bringing down the average performance. This set of simulation results, hence, would serve

Table 1: **Monte Carlo Results I: On Comparative Performance**

This table compares the best performance of each estimator. The median and the SD of the smallest ISE of each estimator are reported. $T = 5$, the number of repetition is 500, and more details are in the text.

	RE				FE			
	$\sigma = 0.5$		$\sigma = 1$		$\sigma = 0.5$		$\sigma = 1$	
	Median	SD	Median	SD	Median	SD	Median	SD
f_1, N=50								
FD	0.0232	0.0198	0.0668	0.0673	0.0160	0.0216	0.0656	0.1058
LL-RE	0.0424	0.0541	0.0713	0.0736	1.2478	0.5316	1.2001	0.6261
LL-LSDV	0.3668	0.0809	0.3932	0.1099	0.4641	0.1710	0.4934	0.2141
f_1, N=100								
FD	0.0123	0.0110	0.0364	0.0393	0.0089	0.0116	0.0311	0.0470
LL-RE	0.0262	0.0252	0.0389	0.0405	1.2059	0.3550	1.2287	0.3954
LL-LSDV	0.3618	0.0522	0.3813	0.0784	0.4386	0.1142	0.4478	0.1334
f_2, N=50								
FD	0.0964	0.0451	0.2501	0.1272	0.2566	0.2591	0.5421	0.4094
LL-RE	0.1658	0.0917	0.2602	0.1292	1.3758	0.5950	1.4073	0.6982
LL-LSDV	0.8486	0.1279	0.8781	0.1796	1.1034	0.2168	1.1536	0.2455
f_2, N=100								
FD	0.0515	0.0220	0.1436	0.0652	0.1868	0.1570	0.3506	0.2380
LL-RE	0.0981	0.0450	0.1458	0.0684	1.3030	0.4149	1.3522	0.4595
LL-LSDV	0.8296	0.0840	0.8394	0.1220	1.1009	0.1310	1.1197	0.1585
f_3, N=50								
FD	0.1358	0.0533	0.2991	0.1244	0.2722	0.1264	0.4955	0.1869
LL-RE	0.2126	0.0926	0.2972	0.1293	1.4589	0.5992	1.4709	0.6333
LL-LSDV	0.4988	0.0721	0.5323	0.0913	0.5313	0.0796	0.5630	0.0964
f_3, N=100								
FD	0.0731	0.0275	0.1832	0.0694	0.1974	0.0877	0.3921	0.1329
LL-RE	0.1262	0.0498	0.1875	0.0738	1.3840	0.4042	1.3636	0.4603
LL-LSDV	0.4965	0.0562	0.5134	0.0700	0.5247	0.0523	0.5530	0.0693
f_4, N=50								
FD	0.1371	0.0572	0.3153	0.1293	0.2852	0.1446	0.5131	0.2582
LL-RE	0.2142	0.0959	0.3167	0.1320	1.4372	0.6158	1.4939	0.6304
LL-LSDV	0.9792	0.1285	1.0267	0.1659	1.1119	0.1616	1.1327	0.1933
f_4, N=100								
FD	0.0776	0.0279	0.1879	0.0704	0.2022	0.1146	0.3897	0.1673
LL-RE	0.1288	0.0519	0.1867	0.0720	1.3883	0.4157	1.3787	0.4500
LL-LSDV	0.9833	0.0888	1.0097	0.1161	1.0866	0.1067	1.1072	0.1398

Table 2: **Monte Carlo Results II: On Persistence**

This table reports how persistence in X_{it} may influence the performance of FD. Mean, Median, and SD of ISE are reported. a is the level of persistence. $N = 50$, the underlying function is f_2, the number of repetition is 500, and more details are in the text.

a	0	0.2	0.4	0.6	0.8	0.95
T=5						
Mean	0.2430	0.2900	0.2883	0.6832	1.2472	4.7058
Median	0.2101	0.2133	0.2316	0.3459	0.8525	4.4740
SD	0.0835	0.1503	0.1499	0.9626	2.2988	8.2646
T=10						
Mean	0.1256	0.1337	0.1711	0.5030	1.5994	6.2064
Median	0.1104	0.1184	0.1262	0.1996	0.7148	4.9632
SD	0.0384	0.0479	0.1125	0.8466	3.6882	10.1938
T=20						
Mean	0.0816	0.0857	0.0956	0.2184	1.4612	7.5214
Median	0.0769	0.0776	0.0830	0.1116	0.4940	5.0920
SD	0.0206	0.0257	0.0424	0.3216	3.3534	11.9760

to caution against applying our estimator to panels of highly persistent time series.

To defend our methodology, we point out that the above data generating process is close to the worst scenario of our estimator. When $a = 0.95$, the conditional distribution of X_{it} given $X_{i,t-1} = v$ is Gaussian with mean $0.95v$ and standard deviation 0.44. Let $v = 0$, for example, the conditional density $p(u|v)$ is close to zero at $\{u : |u| > 0.44 \cdot 4 = 1.56\}$. Recall that we estimate functions on $[-2, 2]$ and that $p(u|v)$ is implicitly on the denominator of the asymptotic variance.

4. Conclusions

In this paper we present a new methodology for estimating the nonlinear component of semiparametric panel data models. Technically, we use first differencing transformation to eliminate individual effects and use marginal integration to recover the nonlinear function of interest. We give the asymptotic properties of our estimator. And Monte Carlo simulations show that our estimator performs reasonably well for finite samples.

References

Andrews, D. W. K., 1991, Asymptotic Optimality of Generalized C_L, Cross-Validation, and Generalized Cross-Validation in Regression with Heteroskedastic Errors, Journal of Econometrics, 47, 359-377

Auestad, B., D. Tjøstheim (1991), Functional identification in nonlinear time series. In G.G. Roussas (ed.), Nonparametric Functional Estimation and Related Topics 493-507, Amsterdam: Kluwer Academic.

Banerjee, A.V., and Duflo, E., 2003. Inequality and Growth: What Can the Data Say?, Journal of Economic Growth, Springer, vol. 8(3), pages 267-99

Cleveland, W.S., 1979, Robust locally weighted regression and smoothing scatterplots, Journal of the American Statistical Association, 74, 829-836.

Cai, Z., Masry, E., 2000, Nonparametric estimation in nonlinear ARX time series models: Projection and linear fitting, Econometric Theory, 16, 465-501

Craven P., and G. Wabha, 1979, Smoothing noisy data with spline functions: Estimating the correct degree of smoothing with generalized cross-validation, Numerische Mathematik, 31, 377-403.

Fan, J., 1992, Design-adaptive nonparametric regression, Journal of the American Statistical Association, 87(420), 998-1004.

Fan, J. and Gijbels, I., 1992, Variable Bandwidth and Local Linear Regression Smoothers, The Annals of Statistics, 20, 2008-2036

Fan, J., Hardle, W., Mammen, E., 1998, Direct Estimation of Low-Dimensional Components in Additive Models, The Annals of Statistics, 26, 943-971

Gasser, T., A. Kneip, W. Köhler, 1991, A flexible and fast method for automatic smoothing, Journal of the American Statistical Association, 86, 643-652

Gasser, T., H.-G. Müller, 1984, Estimating regression functions and their derivatives by the kernel method, Scandinavian Journal of Statistics, 11, 171-185.

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.

Härdle, W., J.S. Marron, 1991, Bootstrap simultaneous error bars for nonparametric regression, Annals of Statistics, 19(2), 778-796.

Härdle, W., E. Mammen 1993, Comparing nonparametric versus parametric regression fits, Annals of Statistics, 21, 1926-1947.

Hengartner, N. W. and Sperlich, S., 2005, Rate optimal estimation with the integration method in the presence of many covariates, Journal of Multivariate Analysis, 95, 246-272

Jones, M.C., S.J. Davies, B.U. Park, 1994, Versions of kernel-type regression estimators, Journal of the American Statistical Association, 89(427), 825-832.

Kim, W., Linton, O.B., and Hengartner, N.W., 1999, A Computationally Efficient Oracle Estimator for Additive Nonparametric Regression with Bootstrap Confidence Intervals, Journal of Computational and Graphical Statistics, 8, 278-297.

Kneip, A., L. Simar, 1996, A general framework for frontier estimation with panel data, Journal of Productivity Analysis, 7, 187-212.

Lee, Y., Mukherjee, D., 2008, New nonparametric estimation of the marginal effects in fixed effects panel models: an application on the environmental Kuznets curve, Working paper.

Li, K., 1987, Asymptotic Optimality for Cp, CL, Cross-Validation and Generalized Cross-Validation, The Annals of Statistics, 15, 958-975

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

Linton, O.B., 1997, Efficient estimation of additive nonparametric regression models, Biometrika, 84, 469-473

Linton, O.B., W. Härdle, 1996, Estimation of additive regression model with known links, Biometrika, 83(3), 529-540.

Linton, O.B, Jacho-Chavez, D. 2009, On Internally Corrected and Symmetrized Kernel Estimators for Nonparametric Regression, forthcoming in TEST, Springer.

Linton, O.B., J.P. Nielsen, 1995, A kernel method of estimating structured nonparametric regression based on marginal integration, Biometrika, 82(1), 93-100.

Masry, E., Tjøstheim, D., 1997 Additive nonlinear ARX time series and projection estimates, Econometric Theory, 13, 214C252

Mallows, C.L., 1973, Some Comments on Cp, Technometrics, 15, 661-675.

Mammen, E., Stove, B., and Tjostheim, D., 2009, Nonparametric additive models for panels of time series, Econometric Theory, 25, 442-481.

Millimet, D.L., J.A. List, T. Stengos, 2003, The environmental Kuznets curve: real progress or misspecified models, The Review of Economics and Statistics, 85(4), 1038-1047.

Nadaraya, E.A., 1964, On estimating regression, Theory of Probability and its Applications, 9, 141-142.

Newey, W. K., 1994, Kernel Estimation of Partial Means and a General Variance Estimator, Econometric Theory, 10, 233-253

Robinson, P.M., 1988, Root-N-consistent semiparametric regression, Econometrica, 56, 931-954.

Ruckstuhl, A. F., Welsh, A. H., Carroll, R. J., 2000, Nonparametric function estimation of the relationship between two repeatedly measured variables, Statistica Sinica, 10, 51-71.

Ruppert, D., M.P. Wand, 1994, Multivariate locally weighted least squares regression, Annals of Statistics, 22(3), 1346-1370.

Severance-Lossin, E. and Sperlich, S. (1999), Estimation of derivatives for additive separable models, Statistics 33: 241-265.

Severini, T.A. and Staniswalis, J.G. 1994, Quasi-likelihood Estimation in Semiparametric Models, Journal of the American Statistical Association, 89, 501-511

Stone, C.J., 1977, Consistent nonparametric regression (with discussion), Annals of Statistics, 5, 595-645.

Su, L., Ullah, A., 2006, Profile likelihood estimation of partially linear panel data models with fixed effects, Economics Letters, 92, 75-81

Su, L., Ullah, A., 2007, More efficient estimation of nonparametric panel data models with random effects, Economics Letters, 96, 375-380.

Tjøstheim, D., Auestad, B.H., 1994, Nonparametric Identification of Nonlinear Time Series: Projections, Journal of the American Statistical Association, 89, 1398-1409

Wand, M.P., M.C. Jones, 1993, Comparison of smoothing parameterizations in bivariate kernel density estimation, Journal of the American Statistical Association, 88(422), 520-528.

Watson, G.S., 1964, Smooth regression analysis, Sankhyā, Series A, 26, 359-372.

Appendix I

Proof for Theorem 2: Theorem 1 establishes root-N consistency for $\hat{\beta}$. In the following we may treat $R_{it} = \Delta Y_{it} - \Delta Z_{it}'\beta$ as known. And we have

$$\hat{f}(u) - f(u) = \int \left[\hat{m}(u,v) - m(u,v)\right] q(v)dv,$$

where the integration is taken on $C \subset \mathbb{R}^d$. Throughout the proof we suppress the domain for notational simplicity. Let $\Upsilon = [e_{it}]$ and $M = [m(X_{it}, X_{i,t-1})]$. By standard argument in multivariate kernel regression asymptotics, and the assumptions that (i) (X_{it}, e_{it}) is i.i.d. across i, (ii) $(X_{it}$ is stationary over t with T fixed, (iii) p_2 is continuously partially differentiable, (iv) f (hence m) is at least twice continuously partially differentiable, and other conditions on the kernel and the associated bandwidth matrix, we have

$$\begin{aligned}\hat{m}(u,v) - m(u,v) &= \iota'(\Gamma'W\Gamma)^{-1}\Gamma'W\Upsilon \\ &\quad + \iota'\left[(\Gamma'W\Gamma)^{-1}\Gamma'W(M - \Gamma\begin{pmatrix} m(u,v) \\ \mathcal{D}_m(u,v)\end{pmatrix})\right] + o_p(\operatorname{tr}(H^2)).\end{aligned}$$

Let U_1 and U_2 denote the first and the second terms on the right, respectively. U_2 gives us the desired asymptotic bias, which is an integration of,

$$\begin{aligned}N^{2/(4+d)}\frac{1}{2}\mu_2(k)\operatorname{tr}\left((I_2 \otimes H^2)\mathcal{H}_m(u,v)\right) &= N^{2/(4+d)}\frac{1}{2}\mu_2(k)\left[\operatorname{tr}\left(H^2\mathcal{H}_f(u)\right) - \operatorname{tr}\left(H^2\mathcal{H}_f(v)\right)\right] \\ &= \frac{1}{2}\mu_2(k)\left[\operatorname{tr}\left(H_0^2\mathcal{H}_f(u)\right) - \operatorname{tr}\left(H_0^2\mathcal{H}_f(v)\right)\right].\end{aligned}$$

We now examine U_1. Let $n = NT$. We first write

$$U_1 = \iota' \left(\frac{1}{n} \Gamma' W \Gamma \right)^{-1} \left(\frac{1}{n} \Gamma' W \Upsilon \right).$$

Use the fact that X_{it} is stationary and T is fixed, we use standard arguments to obtain

$$\iota' \left(\frac{1}{n} \Gamma' W \Gamma \right)^{-1} = \left(\begin{array}{ccc} p_2^{-1} + o_p(1) & -p_2^{-2} \frac{\partial p_2}{\partial u'} + o_p(1) & -p_2^{-2} \frac{\partial p_2}{\partial v'} + o_p(1) \end{array} \right).$$

Note that $o_p(1)$ is uniform, for which we require $N|H|^2 \to \infty$, which means $d < 4$. Hence

$$U_1 = (A_1(u,v) + A_2(u,v) + A_3(u,v))(1 + o_p(1)),$$

where

$$\begin{aligned} A_1(u,v) &= p_2^{-1}(u,v) \frac{1}{n} \sum_{i=1}^{N} \sum_{t=1}^{T} K_H(X_{it} - u) K_H(X_{i,t-1} - v) e_{it}, \\ A_2(u,v) &= -\left(\frac{\partial p_2}{\partial u'} p_2^{-2} \right)(u,v) \frac{1}{n} \sum_{i=1}^{N} \sum_{t=1}^{T} K_H(X_{it} - u) K_H(X_{i,t-1} - v)(X_{it} - u) e_{it} \\ A_3(u,v) &= -\left(\frac{\partial p_2}{\partial v'} p_2^{-2} \right)(u,v) \frac{1}{n} \sum_{i=1}^{N} \sum_{t=1}^{T} K_H(X_{it} - u) K_H(X_{i,t-1} - v)(X_{i,t-1} - v) e_{it}. \end{aligned}$$

In the following, we show that

$$N^{2/(4+d)} \int A_1(u,v) q(v) dv \to_d N(0, V(u)),$$

and that $N^{2/(4+d)} \int A_2(u,v) q(v) dv$ and $N^{2/(4+d)} \int A_3(u,v) q(v) dv$ are negligible asymptotically.

We first examine $N^{2/(4+d)} \int A_1(u,v)q(v)dv$,

$$\begin{aligned}
N^{2/(4+d)} \int A_1(u,v)q(v)dv &= N^{2/(4+d)} \frac{1}{n} \sum_{i,t} e_{it} K_H(u - X_{it}) \int \frac{K_H(v - X_{i,t-1})}{p_2(u,v)} q(v)dv \\
&= N^{2/(4+d)} \frac{1}{n} \sum_{i,t} e_{it} K_H(u - X_{it}) \left(\frac{q(X_{i,t-1})}{p_2(u, X_{i,t-1})} + o_p(tr(H)) \right) \\
&= \frac{1}{\sqrt{N}} \sum_i \xi_{i,N} + o_p(1),
\end{aligned}$$

where $\xi_{i,N} = N^{-d/(2(4+d))} T^{-1} \sum_t e_{it} K_H(u - X_{it}) q(X_{i,t-1}) p_2^{-1}(u, X_{i,t-1})$. Note that the integration on the right is a convolution involving the function K_H which reduces to a generalized delta function in the limit.

$(\xi_{i,N})$ is a triangular array and it can be observed that for each N, $(\xi_{i,N})$ are i.i.d. with zero mean. Next we calculate the second moment. We write

$$\mathbb{E} \left(\xi_i \right)^2 = N^{-d/(4+d)} W_1 + N^{-d/(4+d)} W_2,$$

where

$$\begin{aligned}
W_1 &= \frac{1}{T^2} \mathbb{E} \sum_t e_{it}^2 K_H^2(u - X_{it}) q^2(X_{i,t-1}) p_2^{-2}(u, X_{i,t-1}) \\
W_2 &= \frac{1}{T^2} \mathbb{E} \sum_{s \neq t} e_{it} e_{is} K_H(u - X_{it}) K_H(u - X_{is}) q(X_{i,t-1}) q(X_{i,s-1}) p_2^{-1}(u, X_{i,t-1}) p_2^{-1}(u, X_{i,s-1})
\end{aligned}$$

For W_1, we have

$$\begin{aligned}
W_1 &= \frac{1}{T^2} \sum_t \sigma_t^2 \int \frac{1}{|H|^2} K(H^{-1}(u - v)) q^2(y) p^{-2}(u,y) p_2(v,y) dv dy \\
&= \frac{\bar{\sigma}^2}{T|H|} \int K^2(w) q^2(y) p^{-2}(u,y) p_2(u + Hw, y) dw dy \\
&= \frac{\bar{\sigma}^2 \varphi^d(k)}{T|H|} \int q^2(y) p^{-1}(u,y) dy (1 + o_p(1)).
\end{aligned}$$

Let $p_{2,t,s}(x,y,r,v)$ be the joint density of $(X_{1t}, X_{1,t-1}, X_{1s}, X_{1,s-1})$, and denote $\gamma_{t,s} =$

$\mathbb{E}e_{1t}e_{1s}$ and $\bar{\gamma} = T^{-2}\sum_{s\neq t}\gamma_{t,s}$. It is obvious that $\bar{\gamma} < \infty$. We have for W_2,

$$\begin{aligned}
W_2 &= \frac{1}{T^2}\sum_{s\neq t}\gamma_{t,s}\int K_H(u-x)K_H(u-r)\frac{q(y)q(v)}{p_2(u,y)p_2(u,v)}p_{2,t,s}(x,y,r,v)dxdydrdz \\
&= \bar{\gamma}|H|^{-2}\int K(H^{-1}(u-x))K(H^{-1}(u-r))\frac{q(y)q(v)}{p_2(u,y)p_2(u,v)}p_{2,t,s}(x,y,r,v)dxdydrdz \\
&= \bar{\gamma}\int K(w)K(z)q(y)q(v)p_2^{-1}(u,y)p_2^{-1}(u,v)p_{2,t,s}(u+Hw,y,u+Hz,v)dwdydzdv \\
&= \bar{\gamma}\int q(y)q(v)p_2^{-1}(u,y)p_2^{-1}(u,v)p_{2,t,s}(u,y,u,v)dydv(1+o_p(1)) < \infty
\end{aligned}$$

Hence

$$\begin{aligned}
\mathbb{E}\xi_i^2 &= N^{-d/(4+d)}\frac{\bar{\sigma}^2\varphi^d(k)}{T|H|}\int q^2(y)p^{-1}(u,y)dy(1+o_p(1)) + O_p(N^{-d/(4+d)}) \\
&= \frac{\bar{\sigma}^2\varphi^d(k)}{T|H_0|}\int q^2(y)p^{-1}(u,y)dy + o_p(1).
\end{aligned}$$

To see the asymptotic order of $N^{2/(4+d)}\int A_2(u,v)q(v)dv$, we only need to examine

$$\begin{aligned}
& N^{-d/(4+d)}\mathbb{E}e_{it}^2K_H^2(X_{it}-u)p_2^{-4}(u,X_{i,t-1})\left(\frac{\partial p_2}{\partial u'}(u,X_{i,t-1})(X_{it}-u)\right)^2 q^2(X_{i,t-1}) \\
= \ & N^{-d/(4+d)}\frac{\sigma_t^2}{|H|^2}\int K^2(H^{-1}(v-u))p_2^{-4}(u,y)\left(\frac{\partial p_2}{\partial u'}(u,y)HH^{-1}(v-u)\right)^2 q^2(y)p_2(v,y)dvdy \\
= \ & N^{-d/(4+d)}\frac{\sigma_t^2}{|H|}\int K^2(w)p_2^{-4}\left(\frac{\partial p_2}{\partial u'}(u,y)H_0w\right)^2 q^2(y)p_2(u+Hw,y)dwdy\cdot N^{-2/(4+d)} \\
= \ & O_p(N^{-2/(4+d)}).
\end{aligned}$$

Hence $N^{2/(4+d)}\int A_2(u,v)q(v)dv$ is asymptotically negligible. Similarly we can check that $N^{2/(4+d)}\int A_3(u,v)q(v)dv$ is also negligible asymptotically.

To apply the Liapounov's CLT to $\frac{1}{\sqrt{N}}\sum_i \xi_{i,N}$, we need to check whether the following holds for some constant $\epsilon > 0$,

$$\sum_{i=1}^{N}\mathbb{E}|\xi_{i,N}/\sqrt{N}|^{2+\epsilon} \to 0, \text{ as } N \to \infty. \tag{25}$$

The above is equivalent to

$$N^{-\epsilon/2}\mathbb{E}|\xi_{1,N}|^{2+\epsilon} \to 0, \text{ as } N \to \infty.$$

Observe that $|\cdot|^{2+\epsilon}$ is a convex function, hence

$$|\xi_{i,N}|^{2+\epsilon} \leq N^{-d(2+\epsilon)/(2(4+d))}\frac{1}{T}\sum_{t=1}^{T}\left|e_{it}K_H(u - X_{it})q(X_{i,t-1})p_2^{-1}(u, X_{i,t-1})\right|^{2+\epsilon}.$$

And by the positiveness of k, q, and p_2, and the independence of e_{it} from (X_{it}), we have

$$\begin{aligned}
& \mathbb{E}\left|e_{it}K_H(u - X_{it})q(X_{i,t-1})p_2^{-1}(u, X_{i,t-1})\right|^{2+\epsilon} \\
= \; & \mathbb{E}\left|e_{it}\right|^{2+\epsilon}\mathbb{E}\left|K_H(u - X_{it})q(X_{i,t-1})p_2^{-1}(u, X_{i,t-1})\right|^{2+\epsilon} \\
= \; & \mathbb{E}\left|e_{it}\right|^{2+\epsilon}|H|^{-(2+\epsilon)}\int K^{(2+\epsilon)}(H^{-1}(u-x))q^{2+\epsilon}(y)p_2^{-(2+\epsilon)}(u,y)p_2(x,y)dxdy \\
= \; & \mathbb{E}\left|e_{it}\right|^{2+\epsilon}|H|^{-(1+\epsilon)}\int K^{(2+\epsilon)}(w)q^{2+\epsilon}(y)p_2^{-(2+\epsilon)}(u,y)p_2(u+Hw,y)dwdy \\
= \; & |H|^{-(1+\epsilon)}\mathbb{E}\left|e_{it}\right|^{2+\epsilon}\int K^{(2+\epsilon)}(w)dw\int q^{2+\epsilon}(y)p_2^{-(1+\epsilon)}(u,y)dy(1+o_p(1)) \\
= \; & O(N^{(1+\epsilon)/(4+d)}),
\end{aligned}$$

since both k and q are bounded, p_2 is bounded from zero, and $\mathbb{E}\left|e_{it}\right|^{2+\epsilon} \leq \left(\mathbb{E}|e_{it}|^{4+2\epsilon}\right)^{1/2} < \infty$. Hence

$$N^{-\epsilon/2}\mathbb{E}|\xi_{1,N}|^{2+\epsilon} = O(N^{-(1+\epsilon)(d-1)/(4+d)-2\epsilon/(4+d)}) = o(1).$$

Hence the condition in (25) is verified. Now we apply the Liapunov's CLT to the triangular array $(\xi_{i,N})$ and obtain the desired asymptotic distribution. □