Nonparametric multi-dimensional fixed effects panel data models

Daniel J. Henderson[a,b,c], Alexandra Soberon[d], and Juan M. Rodriguez-Poo[d]

[a]Department of Economics, Finance and Legal Studies, University of Alabama, USA
[b]School of Mathematical Sciences, Nankai University, China
[c]Institute for the Study of Labor (IZA), Germany
[d]Department of Economics, Universidad de Cantabria, Spain

28th December 2020

Abstract

Multi-dimensional panel data sets are routinely employed to identify marginal effects in empirical research. Fixed effects estimators are typically used in order to deal with potential correlation between unobserved effects and regressors. Nonparametric estimators for one-way fixed effects models exist, but are cumbersome to employ in practice as they typically require iteration, marginal integration or profile estimation. We develop a nonparametric estimator that works for essentially any dimension fixed effects model, has a closed-form solution and can be estimated in a single-step. A cross-validation bandwidth selection procedure is proposed and asymptotic properties (for either a fixed or large time dimension) are given. Finite sample properties are shown via simulations, as well as with an empirical application which further extends our model to the partially linear setting.

Keywords: Fixed effects, Nonparametric, Multi-dimensional, Panel

JEL Classification: C14, C23

1 Introduction

The growing availability of large data sets has given rise to multi-dimensional panel methods being employed in empirical studies. For example, these methods are being used to study phenomena such as salary data of firms' employees over time (three-dimensional panel), sectoral level trade between countries or regions (three-dimensional panel), air passenger traffic between multiple hubs served by different airlines (four-dimensional panel) and so on (e.g., see Balazsi et al. (2017)). Fixed effects specifications are primarily used to take into account unobserved heterogeneity in these data sets. The introduction of said models enlarges the possibilities of handling large data sets, but theoretical problems arise. Davis (2002) considers estimation of a linear parametric three-dimensional additively separable panel data model with fixed effects. Balazsi et al. (2015) and Ye and Wu (2014) extend this model to allow for additive interaction terms in a three-dimensional model.

In many areas of economics, especially in applied microeconomics, interest primarily hinges on the gradient of the conditional mean (i.e., marginal effects). In the hedonic price literature, the gradient of the conditional mean is used to recover preferences of individuals. For example, Bishop and Timmins (2018) use the gradients from a hedonic price model to determine preferences for clean air. In the risk and uncertainty literature, gradients are required in order to estimate individual attitudes towards risk (e.g., see Chiappori et al. (2009)). These marginal effects are trivial to calculate in linear parametric models as they are simply the estimated slope coefficients.

Economic theory typically does not provide enough information to fully specify a parametric model. Nonparametric and semiparametric regression models can be specified and estimated with much less information. In this paper, we consider the specification and estimation of gradients in multi-dimensional nonparametric fixed effect models.[1] More precisely, we discuss the estimation procedure, the assumptions needed to achieve consistency, derive the asymptotic distribution of the nonparametric estimator (for a fixed or large time dimension) and develop a bandwidth selection procedure. To the best of our knowledge, no other studies in this literature consider estimation of these types of models.

There is a relatively large literature in nonparametric and semiparametric panel data models. Surveys on the topic exist and we suggest the interested reader to consult Ai and Li (2008), Henderson and Parmeter (2015), Li et al. (2015), Parmeter and Racine (2018), Rodriguez-Poo and Soberon (2017) and/or Su and Ullah (2011) and the references within. Even though there is a large literature, nearly all work has focused on models with solely individual effects (i.e., one-way error component models) and are typically cumbersome to estimate under fixed effects. For example, Henderson et al. (2008) require iterative methods and Qian and Wang (2012) use marginal integration. A potentially useful strand of the literature, starting with Su and Ullah (2006), uses profile least-squares (see also Gao and Kunpeng (2013)), but these methods require estimation of the fixed effects. Even in the single-dimensional case, these individual effects are typically considered to be nuisance parameters and the number of individual effects tends towards infinity asymptotically. In addition, for identification reasons, these methods require assumptions on these fixed effects (mean zero), which are not testable. Alternatively, Lee et al. (2019) consider nonparametric estimation of the marginal effect using a local-within transformation to deal with the presence of fixed effects. They show that the resulting estimator satisfies standard properties of the local linear estimator, but is subject to possible random denominator problems. A modification of the proposed weight is required to overcome it. Given the difficulty of these estimators, it is not surprising that these methods have not been extended for a general case multi-dimensional panel data model.

In this paper, we provide a direct estimation procedure for the gradient of the conditional mean in a multi-dimensional nonparametric panel data model, once the unobserved heterogeneity has been deleted using a pairwise transformation. Under rather weak assumptions, we show that our

[1] Although our interest is in the gradient, researchers may also be interested the conditional mean. We discuss how to estimate the unknown function in Section 2.4.

estimator is consistent, asymptotically normal and can be extended to semiparametric models (for essentially any dimension panel), for example, semiparametric partially linear models (see Section 6.1). Further, it is also possible to prove that, in only one-step, our resulting estimator almost achieves optimality.

In order to showcase the applicability of our estimator, we consider an empirical application. Specifically, we extend our model to the partially linear setting to look at the relationship between the price of rental housing and housing vouchers. A heterogeneous result with respect to the ratio of the rent of a rental unit to the US Department of Housing and Urban Development fair market rent has been observed in the literature (Eriksen and Ross (2015)), but has been achieved by arbitrary splits of the sample. Here we avoid these arbitrary splits by adopting a semiparametric approach whereby we obtain an elasticity for each rental unit in the sample and confirm existence of both negative and positive impacts of housing vouchers on the price of rental housing. We find that positive elasticities are concentrated in the Western United States and specifically in areas which are more supply inelastic, but overall, negative elasticities are more prominent in this dataset. This suggests that increasing rents for those who do not receive subsidies are likely localized and not predominant in the U.S.

The remainder of this paper proceeds as follows: Section 2 sets up the econometric model, proposes the pairwise estimator and develops its asymptotic theory. Section 2.4 shows how to recover the unknown function, while Section 3 develops a cross-validation method to select the bandwidths. Section 4 considers extensions to interactive fixed effects. Section 5 uses simulations to determine the finite sample properties of our estimator and Section 6 provides the empirical application. Section 7 concludes. All proofs can be found in the Appendix.

2 Model and estimation procedure

With the aim of estimating the gradient of the conditional mean function (i.e., marginal effects), in this section we outline a pairwise differencing procedure that enables us to deal with the presence of several unobserved effects and, at the same time, obtain a direct nonparametric estimator that almost achieves the optimal rate of convergence for this type of problem (i.e., $\sqrt{N_1 N_2 T |H| H}$ for T large) in a single-step. Further, it leads to some efficiency gains in finite samples since this transformation enables us to consider all time-dependencies within the observations of each individual.[2]

[2] In a previous version of the paper, a fixed effects estimator using a within transformation to deal with the presence of the individual effects was proposed. However, a high-dimensional kernel weight was required to deal with the well-known non-negligible asymptotic bias of this type of differencing estimator. The resulting nonparametric estimator was subject to a large variance and slow rate of convergence. The theoretical and simulation results for our multi-dimensional within estimator are available upon request.

2.1 Multi-dimensional models

Two-dimensional fixed effects panel data models (see Baltagi (2015) and Hsiao (2014)) control for unobserved heterogeneity by introducing a time effect, λ_t, and an individual effect, μ_i. The availability of big datasets which contain information on multiple dimensions require the use of multidimensional panel data methods. These methods allow for analysis of complex big data sets, as they control for several sources of unobserved heterogeneity. For example, three-dimensional panel data models are employed to study phenomena in many economic fields, such as international trade, transportation, labor, housing and migration (see Mátyás (2017) for a recent review).

The most common three-dimensional panel data model is given as

$$Y_{ijt} = X_{ijt}^{\top}\beta + \mu_i + \gamma_j + \lambda_t + \epsilon_{ijt}, \quad i = 1, \dots, N_1, \quad j = 1, \dots, N_2, \quad t = 1, \dots, T, \tag{2.1}$$

where the μ_i, γ_j, and λ_t parameters are the cross-sectional (i and j) and time-specific fixed effects (t), ϵ_{ijt} is the idiosyncratic error term, the X_{ijt} variables are the usual covariates (of dimension d), and β is the $d \times 1$ vector of structural parameters.

In the trade literature, Mátyás et al. (1997) proposed this model and subsequent authors proposed extensions of this model to take into account unobserved bilateral heterogeneity by including bilateral specific (i.e., μ_{ij}, γ_{jt}, λ_{it}) effects (see Egger and Pfaffermayr (2003), Chen and Wall (2005), Baltagi et al. (2003), and Baier and Bergstrand (2007)). In the two-dimensional setting, there are only four types of fixed effects specifications. Nevertheless, as Balazsi et al. (2017) point out, in the three-dimensional models there are (2^6) possible ways to formulate the unobserved fixed effects. The most commonly used specifications in the empirical studies are

$$\begin{aligned}
Y_{ijt} &= X_{ijt}^{\top}\beta + \mu_{ij} + \epsilon_{ijt}, && (2.2)\\
Y_{ijt} &= X_{ijt}^{\top}\beta + \mu_{ij} + \lambda_t + \epsilon_{ijt}, && (2.3)\\
Y_{ijt} &= X_{ijt}^{\top}\beta + \gamma_{jt} + \epsilon_{ijt}, && (2.4)\\
Y_{ijt} &= X_{ijt}^{\top}\beta + \mu_{ij} + \gamma_{jt} + \epsilon_{ijt}, && (2.5)\\
Y_{ijt} &= X_{ijt}^{\top}\beta + \mu_{ij} + \gamma_{jt} + \lambda_{it} + \epsilon_{ijt}. && (2.6)
\end{aligned}$$

Note that these fixed effects specifications are appealing from an empirical point of view as they are implied via economic theory. For example, if Y_{ijt} represents the exports from country i to country j in year t, the gravity literature (Anderson and van Wincoop (2003)) argues that country fixed effects are relevant variables as they represent unobservable multilateral resistance levels. If the multilateral resistance levels are time-varying (i.e., μ_{it} and γ_{jt}), trade theory would support Models 2.5 and 2.6. Another relevant example can be found in the price dispersion literature. If Y_{ijt} is the Gini log-odds ratio for carrier i, for route j, in time period t, Gerardi and Shapiro (2009) find that price dispersion in the airline industry increases with competition. This model requires that the fixed effects are unique to ij pairs (i.e, the γ_{ij} are defined as carrier-route fixed effects) and λ_t is the time effect, and Model 2.3 would be appropriate.

In order to avoid possible misspecification in the functional form, we consider nonparametric estimation of the most general formulation (2.1). In particular, the prime interest of this paper is the local marginal effect (i.e., the first derivative of the nonparametric function). As noted in Lee et al. (2019), estimation of the local behavior of the slope of a regression function (i.e., elasticity or marginal effects) without assuming a pre-specific parametric functional form, is of great interest in many areas of economics (e.g., applied microeconomics or policy evaluation). For example, we may be interested in studying marginal propensity to consume and save in consumer economics, the elasticity of capital and/or labor in production economics, and recovering the preferences of individuals in the hedonic price literature.

For the sake of simplicity, in the following subsection, we develop an estimation technique that leads to consistent estimators of the gradient of the nonparametric especification of (2.1). It is straightforward to extend the procedure to the corresponding nonparametric specifications of models (2.2)-(2.6) (see Section 4) or even higher dimensions of the data.[3], [4]

2.2 Pairwise estimator of the gradient function

We start by assuming that data are available from a three-dimensional panel data model of the form

$$Y_{ijt} = m(X_{ijt}) + \mu_i + \gamma_j + \lambda_t + \epsilon_{ijt}, \quad i = 1, \ldots, N_1, \quad j = 1, \ldots, N_2, \quad t = 1, \ldots, T, \tag{2.7}$$

where index t refers to time periods, indexes i and j denote cross-sectional units, X_{ijt} is a $d \times 1$ vector of explanatory variables, and $m(\cdot)$ is a smooth unknown function. μ_i, γ_j, and λ_t are fixed effects (for example, firm i producing commodity j in time period t), and ϵ_{ijt} are the *i.i.d.* idiosyncratic disturbance terms. μ_i, γ_j and/or λ_t are allowed to be correlated with X_{ijt}.[5],[6]

Our main interest is in the local marginal change of the conditional mean of Y_{ijt} in (2.7) with respect to an element of X_{ijt}, i.e.,

$$D_m(x) = (\partial m(x)/\partial x_1, \ldots, \partial m(x)/\partial x_d)^\top$$

for a given $x = (x_1, \ldots, x_d)^\top$ in a compact support $\mathcal{X} \in I\!R^d$ of X_{ijt}, for some $d \geq 1$.

The large number of dummy variables that characterize these multidimensional panel data models can make the estimation procedure computationally difficult when the dimensionality

[3]For example, a fourth order model can be given as $Y_{ijlt} = X_{ijlt}^\top \beta + \mu_i + \gamma_j + \eta_l + \lambda_t + \epsilon_{ijlt}$, where l represents a third cross-sectional dimension and η_l is its corresponding fixed effect.

[4]See Mundra (2005), Lee et al. (2019), Qian and Wang (2012) or Rodriguez-Poo and Soberon (2015) for non-parametric estimation in a two-dimensional setting.

[5]Given that the fixed effects are additively separable, it can be argued that this is a semiparametric model. We nonetheless follow the literature and refer to this as a nonparametric fixed effects model.

[6]An interesting paper by Freyberger (2018) proposes a nonparametric panel data model with two-dimensional, unobserved (interactive) individual effects that enter non-additively with a fixed time dimension. In the case where the individual effects enter in this particular non-additive structure, this estimator would be preferable. That being said, our primary interest lies in higher-dimensional panel and while it may be feasible to extend this estimator to the higher-dimensional panel data models we have in mind, this is beyond the scope of this paper.

of the data is three or more as we have to deal with the incidental parameter problem (e.g., see Lancaster (2000)). To identify the fixed effects and avoid asymptotically biased estimators, authors usually impose restrictions on the fixed effects. Typically authors either normalize the fixed effects (i.e., set their average to zero) or leave out parameters belonging to the last (or first) individual or time period (e.g., see Balazsi et al. (2015)). Instead, we propose a differencing transformation in (2.7) that enables us to remove all fixed effects simultaneously without having to impose any identification conditions related to the fixed effects in (2.7).

To removed the fixed effects, we first perform a within transformation: $Y_{ijt} - \overline{Y}_t$, where $\overline{Y}_t = (N_1 N_2)^{-1} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} Y_{ijt}$ is the cross-sectional average, to remove the time fixed effects (i.e., λ_t). Second, inspired by Stromberg et al. (2000), Honoré and Powell (2005) and Soberon et al. (2020) (in a different context), a pairwise differencing transformation is proposed to remove the cross-sectional heterogeneities (i.e., μ_i and γ_j), simultaneously.[7] Hence, subtracting from time t of $(Y_{ijt} - \overline{Y}_t)$, time s, for $s \neq t$, we get

$$\widetilde{Y}_{ijts} = (Y_{ijt} - \overline{Y}_t) - (Y_{ijs} - \overline{Y}_s).$$

Applying this transformation to the regression model (2.7) and rearranging terms leads to

$$\begin{aligned} &\widetilde{Y}_{ijts} = m(X_{ijt}) - m(X_{ijs}) - \frac{1}{N_1 N_2} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} [m(X_{ijt}) - m(X_{ijs})] + \widetilde{\epsilon}_{ijts}, \\ &i = 1, \dots, N_1, \quad j = 1, \dots, N_2, \quad t = 1, \dots, T, \quad t < s, \end{aligned} \tag{2.8}$$

where $\widetilde{\epsilon}_{ijts} = (\epsilon_{ijt} - \overline{\epsilon}_t) - (\epsilon_{ijs} - \overline{\epsilon}_s)$.

Assuming that $m(\cdot)$ is sufficiently smooth, we consider a Taylor expansion of each element of $m(\cdot)$ in (2.8) around x obtaining

$$\begin{aligned} &m(X_{ijt}) - m(X_{ijs}) - \frac{1}{N_1 N_2} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} [m(X_{ijt}) - m(X_{ijs})] \\ &= \widetilde{X}_{ijts}^{\top} D_m(x) + R_{m_{ijts}}(x), \end{aligned}$$

where $\widetilde{X}_{ijts} = (X_{ijt} - \overline{X}_t) - (X_{ijs} - \overline{X}_s)$, $D_m(x) = vec(\partial m(x)/\partial x^{\top})$ is a $d \times 1$ vector of gradient functions in (2.8), and $R_{m_{ijts}}(x)$ is the remainder term from the Taylor expansion[8] (see the Appendix).

[7]Alternative transformations are proposed in the literature for this type of multi-dimensional problem. For a fully parametric model, Balazsi et al. (2017) suggest the following transformation

$$\widetilde{Y}_{ijt} = Y_{ijt} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j \cdot} - \overline{Y}_t + 2\overline{Y},$$

where $\overline{Y}_{i\cdot\cdot} = (N_2 T)^{-1} \sum_{jt} Y_{ijt}$ and $\overline{Y}_{\cdot j \cdot} = (N_1 T)^{-1} \sum_{it} Y_{ijt}$. However, its extension to the nonparametric framework is not straightforward as we would have to use a kernel weight that controls the distance among all time periods and cross-sectional units to avoid the non-negligible asymptotic bias pointed out in Mundra (2005) for these types of differencing specifications.

[8]See Fan and Gijbels (1995b) or Ruppert and Wand (1994) for further details.

This suggests that we can estimate the unknown gradient directly by regressing the transformed dependent variable on the right-hand side of this approximation with kernel weights. However, if we consider kernels only around X_{ijt}, the remainder term in the Taylor expansion will not vanish asymptotically since the distance between any of the terms in X_{ijs} and x cannot be controlled by a fixed bandwidth. This phenomena was pointed out in Mundra (2005) and Lee et al. (2019), and was solved in Rodriguez-Poo and Soberon (2015) for a two dimensional panel.

To overcome this issue, we propose to use a kernel weight which controls the distance between any (X_{ijt}, X_{ijs}). Then, for X_{ijt} and X_{ijts} in a neighborhood of x, the unknown vector of gradient functions $D_m(x)$ can be estimated by minimizing the objective function

$$\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\sum_{t=1}^{T-1}\sum_{s=t+1}^{T}\left[\widetilde{Y}_{ijts} - \widetilde{X}_{ijts}^{\top} D_m(x)\right]^2 K_H(X_{ijt}-x)K_H(X_{ijs}-x), \tag{2.9}$$

where H is a $d \times d$ bandwidth matrix that is symmetric and positive definite and each $K(\cdot)$ is a non-negative product kernel function such as, for each u, it holds that

$$K_H(u) = |H|^{-1/2}\prod_{l=1}^{d} k(H^{-1/2}u_l), \quad u = (u_1, \ldots, u_d)^{\top},$$

where $k(\cdot)$ is a univariate kernel function. The above exposition suggests $\widehat{D}_m(\cdot)$ as an estimator for $D_m(\cdot)$.

Assuming that $\sum_{ijts} K_H(X_{ijt}-x)K_H(X_{ijs}-x)\widetilde{X}_{ijts}\widetilde{X}_{ijts}^{\top}$ is non-singular, and letting $\widehat{D}_m(x, H)$ be the minimizer of (2.9), the nonparametric pairwise least-squares estimator of $D_m(x)$ is

$$\begin{aligned}
\widehat{D}_m(x; H) &= \left(\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\sum_{t=1}^{T-1}\sum_{s=t+1}^{T} K_H(X_{ijt}-x)K_H(X_{ijs}-x)\widetilde{X}_{ijts}\widetilde{X}_{ijts}^{\top}\right)^{-1} \\
&\times \sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\sum_{t=1}^{T-1}\sum_{s=t+1}^{T} K_H(X_{ijt}-x)K_H(X_{ijs}-x)\widetilde{X}_{ijts}\widetilde{Y}_{ijts},
\end{aligned} \tag{2.10}$$

where we emphasize that this general form results in a closed-form solution that is based on the product of $2d$ kernels. Note that it is straightforward to redefine $\widetilde{X}_{ijts}$ and $\widetilde{Y}_{ijts}$ for essentially any dimension panel with the necessary transformation (see Section 4). Further, selection of the bandwidth via data-driven techniques will be developed in Section 3.

2.3 Asymptotic development

To simplify notaton, we define $I\!N = N_1N_2$.[9] This notation simplifies the discussion on the asymptotic properties, since it implicitly allows us to write $I\!N \to \infty$ when both N_1 and N_2 diverge to infinity, or when one of these dimensions is fixed and the other grows to infinity. In

[9] Using this notation makes the theory more general for additional dimensions.

practice, it is often the case that only one dimension is large (e.g., firm-commodity linked data when the number of commodities is much larger than the number of firms). Thus, one dimension, say N_1, might be considered fixed, while the other, say N_2, is considered large. However, the notation used in this paper to establish the asymptotic properties of the proposed estimator enables to consider different scenarios:

i) $N_1 \to \infty$ and N_2 fixed.

ii) N_1 fixed and $N_2 \to \infty$.

iii) $N_1 \to \infty$ and $N_2 \to \infty$.

In order to establish the asymptotic properties of $\widehat{D}_m(x; H)$, the following assumptions about the data generating process defined in (2.7), moments, and densities are necessary.

Assumption A1 *X_{ijt} is independent across the subscripts i and j for each fixed t, and strictly stationary over t for each fixed i and j.*

Assumption A2 *For $\kappa = |t - s|$, where $\kappa \in \{1, \ldots, (T-1)\}$, the random errors ϵ_{ijt} are independent across i and j and satisfy $E(\epsilon_{ijt}|X_{ij1}, X_{ij(1+\kappa)}) = 0$ and $E(\epsilon_{ijt}^2|X_{ij1} = x_1, X_{ij(1+\kappa))} = x_2) = \sigma^2(x_1, x_2) \in (0, \infty)$, where $\sigma^2(\cdot)$ is continuous at $(x_1, x_2) \in int(\chi)$.*

Assumption A3 *The density of X_{ijt} satisfies $0 < f_{X_{ijt}}(\cdot) < \infty$ and is twice continuously differentiable in all its arguments with bounded second-order derivatives at any point of its support. For $t < s$, the joint density of distinct elements of (X_{ijt}, X_{ijs}) is bounded and continuously differentiable in all its arguments, at any point of its support.*

Assumption A4 *The map $m(\cdot) : \chi \to I\!R$ is Borel measurable and twice continuously differentiable at x in the interior of χ with bounded derivatives.*

Assumption A5 *$K(u) = \prod_{\ell=1}^{d} k(u_\ell)$ is a product kernel, and the univariate kernel function $k(\cdot)$ is compactly supported and bounded such that $\int k(u)du = 1$, $\int uu^\top k(u)du = \mu_2(K)I_d$, and $\int k^2(u)du = R(K)$, where $\mu_2(K) \neq 0$ and $R(K) \neq 0$ are scalars and I_d is a $d \times d$ identity matrix. All odd-order moments of k vanish, that is, $\int u_1^{\iota_1}, \ldots, u_d^{\iota_d} k(u)du = 0$, for all non-negative integers $\iota_1, \ldots, \iota_d$ such that their sum is odd.*

Assumption A6 *The bandwidth matrix H is symmetric and positive definite, where each element of H tends to zero and $I\!N|H| \to \infty$, as $I\!N \to \infty$, where $|H| \equiv h_1 \cdots h_d$.*

Assumption A7 *There exists some $\delta > 0$ such that $E[|\epsilon_{ij1}|^{(2+\delta)}|X_{ij1} = x_1, X_{ij(1+\kappa)} = x_2]$ and $E[|\epsilon_{ij(1+\kappa)}|^{(2+\delta)}|X_{ij1} = x_1, X_{ij(1+\kappa)} = x_2]$ are continuous at $(x_1, x_2) \in int(\chi)$, for $\kappa = |t - s|$, where $\kappa \in \{1, \ldots, (T-1)\}$.*

Assumptions A1 and A2 are rather standard in panel data analysis. Dependence between the subscripts i and j could be considered, but that is not a usual problem in the vast majority of the empirical studies conducted on multidimensional setting so we leave that case for future research. Assumptions A3 and A4 are basically smoothness and boundedness conditions on the density function and moment functionals. Assumptions A5 and A6 are standard in the nonparametric literature for the kernel function and bandwidth. For the sake of simplicity, a second-order kernel is used so the number of continuous regressions should be less than four (e.g., Cai and Masry (2000) or Qian and Wang (2012)). If $d \geq 4$, we have to use higher order kernels or resort to semi-parametric methods to avoid the well-known "curse of dimensionality". Therefore, we are able to incorporate some prior information coming from economic theory or past theory in the parametric part of the model and, at the same time, we keep maintaining the flexibility in the specification of the model in the nonparametric part. Further, it follows from Assumption A5 that $\int u^a k^b(u)du < \infty$ for $a = 2, 4$ and $b = 1, 2$ in the univariate case. Assumption A6 allows for T either fixed or going to infinity. Thus, in order to establish the asymptotic normality of this estimator when T is fixed, Assumption A7 is necessary to check the Lyapounov condition. Nevertheless, when T is large, the mathematical derivation is more complicated and some additional (stronger) conditions such as the following are needed.

Assumption B1 *The bandwidth matrix H is symmetric and positive definite, where each element of H tends to zero and $\mathbb{N}T|H| \to \infty$, as $\mathbb{N} \to \infty$ and $T \to \infty$, where $|H| = h_1 \cdots h_d$.*

Assumption B2 *For each fixed i and j, $(X_{ijt}, \epsilon_{ijt})$ is a strictly stationary α-mixing process with the mixing coefficient satisfying the condition $\alpha(\widetilde{\kappa}) = O(\widetilde{\kappa}^{-\tau})$, where $\tau = (2+\delta)(1+\delta)/\delta$ and $\widetilde{\kappa} = |\kappa' - \kappa|$, for $\kappa \in \{1, \ldots, (T-1)\}$ and $\kappa < \kappa'$.*

Assumption B3 *$\mathbb{N}T^{(\tau-1)/\tau}|H|^{(2+\delta)/(1+\delta)} \to \infty$ as $\mathbb{N} \to \infty$ and $T \to \infty$.*

We do not need to impose an explicit condition on the limit relation between $\mathbb{N}$ and T, as is standard in the large panel data literature (e.g., Hahn and Newey (2004) and Lee and Phillips (2015)). In particular, in Assumption B1 we can allow for T to be quite small relative to $\mathbb{N}$, so the large T condition that we use here is much weaker than in the standard large panel data regression literature. Further, conditions B2 and B3 are similar to those established for nonlinear time series models. In particular, many stationary time series or Markov chains fulfilling certain (mild) conditions are α-mixing with exponentially decaying coefficients. Then, Assumption B2 contains the α-mixing condition for weakly dependent stochastic processes (see Cai (2003), Cai and Li (2008) and/or Carrasco and Chen (2002) for a deeper discussion).

Before stating our theorems, we introduce the following notation: $\int uu^\top uu^\top k(u)du = \mu_4(K)I_d$, $\mu_2^2(K) = (\mu_2(K))^2$, $\int uu^\top k^2(u)du = R_2(K)I_d$, and $\widetilde{\nu}_v(K) = R_2(K)R^{d-1}(K)/\mu_2^{2d}(K)$. The conditional variance of ϵ_{ijt} satisfies $Var(\epsilon_{ij1}^2|X_{ij1} = x, X_{ij(1+\kappa)} = x) = \sigma^2(x,x)$ and the conditional covariance is $E(\epsilon_{ij1}\epsilon_{ij(1+\kappa)}|X_{ij1}=x, X_{ij(1+\kappa)}=x)=\sigma_{1,(1+\kappa)}(x,x)$. In addition,

$$
\begin{aligned}
\varpi_4(x,H) &= \int u_1u_1^\top H^{1/2}\mathcal{H}_m(x)H^{1/2}u_1u_1^\top K(u_1)du_1,\\
\varpi_2(x,H) &= \int u_1u_{(1+\kappa)}^\top H^{1/2}\mathcal{H}_m(x)H^{1/2}u_{(1+\kappa)}u_1^\top K(u_1)K(u_{(1+\kappa)})du_1du_{(1+\kappa)}
\end{aligned}
$$

are $d\times d$ positive definite and finite matrices, for $\kappa\in\{1,\ldots,(T-1)\}$.

Finally, let $\mathbb{X}=(X_{111},\ldots,X_{N_1N_2T})$ be the observed covariates sample. Under these assumptions, the following results regarding the conditional Mean Squared Error (MSE) and the asymptotic distribution of the nonparametric pairwise least-squares estimator, $\widehat{D}_m(x,H)$, are obtained. For $\kappa=|t-s|$, we let $\mathcal{H}_m(\cdot)$ be the Hessian matrix of $m(\cdot)$ and $D_{f_\kappa}(\cdot)$ be the gradient vector of $f_{X_{ij1},X_{ij(1+\kappa)}}(\cdot,\cdot)$.

Theorem 1

(i) *Under Assumptions A1-A6, as $I\!N$ tends to infinity, for fixed T,*

$$
\begin{aligned}
MSE[\widehat{D}_m(x;H)|\mathbb{X}] \sim & \left(\frac{1}{2\mu_2^d(K)\sum_{\kappa=1}^{T-1}\left(1-\frac{\kappa}{T}\right)f_{X_{ij1},X_{ij(1+\kappa)}}(x,x)}\right)^2\varphi_1(x,h)\varphi_1(x,h)^\top\\
&+\frac{d\widetilde{\nu}_v(K)\sum_{\kappa=1}^{T-1}\left(1-\frac{\kappa}{T}\right)[\sigma^2(x,x)-\sigma_{1,(1+\kappa)}(x,x)]f_{X_{ij1},X_{ij(1+\kappa)}}(x,x)}{I\!NT|H|\left(\sum_{\kappa=1}^{T-1}\left(1-\frac{\kappa}{T}\right)f_{X_{ij1},X_{ij(1+\kappa)}}(x,x)\right)^2}H^{-1},
\end{aligned}
$$

where $\varphi_1(x,H)=\sum_{\kappa=1}^{T-1}\left(1-\frac{\kappa}{T}\right)[\varpi_4(x;H)-\varpi_2(x;H)]D_{f_\kappa}(x)$.

(ii) *Under Assumptions A1-A5 and B1-B2, as $I\!N$ tends to infinity and $T\to\infty$,*

$$
\begin{aligned}
MSE[\widehat{D}_m(x;H)|\mathbb{X}] \sim & \left(\frac{1}{2\mu_2^d(K)\sum_{\kappa=1}^{T-1}f_{X_{ij1},X_{ij(1+\kappa)}}(x,x)}\right)^2\varphi_2(x,H)\varphi_2(x,H)^\top\\
&+\frac{d\widetilde{\nu}_v(K)\sum_{\kappa=1}^{T-1}[\sigma^2(x,x)-\sigma_{1,(1+\kappa)}(x,x)]f_{X_{ij1},X_{ij(1+\kappa)}}(x,x)}{I\!NT|H|\sum_{\kappa=1}^{T-1}f_{X_{ij1},X_{ij(1+\kappa)}}(x,x)}H^{-1},
\end{aligned}
$$

where $\varphi_2(x,H)=\sum_{\kappa=1}^{T-1}[\varpi_4(x;H)-\varpi_2(x;H)]D_{f_\kappa}(x)$.

Theorem 2 *Under Assumptions A1-A7. If $\sqrt{I\!N|H|}H^{3/2}=O(1)$, then as $I\!N$ tends to infinity, for fixed T,*

$$
\sqrt{I\!N|H|}H^{1/2}\left(\widehat{D}_m(x;H)-D_m(x)-B^{(1)}(x;H)\right) \xrightarrow{d} \mathcal{N}(0,V^{(1)}(x)),
$$

where,

$$
\begin{aligned}
B^{(1)}(x;H) &= \frac{1}{2\mu_2^d(K)}\sum_{\kappa=1}^{T-1}\left(1-\frac{\kappa}{T}\right)[\varpi_4(x;H)-\varpi_2(x;H)]D_{f_\kappa}(x)\left(\sum_{\kappa=1}^{T-1}\left(1-\frac{\kappa}{T}\right)f_{X_{ij1},X_{ij(1+\kappa)}}(x,x)\right)^{-1}, \\
V^{(1)}(x) &= \frac{\widetilde{\nu}_v(K)}{T}\sum_{\kappa=1}^{T-1}\left(1-\frac{\kappa}{T}\right)[\sigma^2(x,x)-\sigma_{1,(1+\kappa)}(x,x)]f_{X_{ij1},X_{ij(1+\kappa)}}(x,x) \\
&\times \left(\sum_{\kappa=1}^{T-1}\left(1-\frac{\kappa}{T}\right)f_{X_{ij1},X_{ij(1+\kappa)}}(x,x)\right)^{-2} I_d.
\end{aligned}
$$

Theorem 3 *Under Assumptions A1-A5, and B1-B3. If* $\sqrt{I\!N|H|}H^{3/2} = O(1)$*, then as* $I\!N$ *tends to infinity and* $T \to \infty$*,*

$$
\sqrt{I\!N|H|}H^{1/2}\left(\widehat{D}_m(x;H) - D_m(x) - B^{(2)}(x;H)\right) \quad \xrightarrow{d} \quad \mathcal{N}(0, V^{(2)}(x)),
$$

where,

$$
\begin{aligned}
B^{(2)}(x;H) &= \frac{1}{\mu_2^d(K)}\sum_{\kappa=1}^{T-1}[\varpi_4(x;H)-\varpi_2(x;H)]D_{f_\kappa}(x)\left(\sum_{\kappa=1}^{T-1}f_{X_{ij1},X_{ij(1+\kappa)}}(x,x)\right)^{-1}, \\
V^{(2)}(x) &= \widetilde{\nu}_v(K)\sum_{\kappa=1}^{T-1}[\sigma^2(x,x)-\sigma_{1,(1+\kappa)}(x,x)]f_{X_{ij1},X_{ij(1+\kappa)}}(x,x)\left(\sum_{\kappa=1}^{T-1}f_{X_{ij1},X_{ij(1+\kappa)}}(x,x)\right)^{-2} I_d.
\end{aligned}
$$

The results of Theorems 1-3 are rather standard. In particular, it is shown that, conditionally on the sample, $\widehat{D}_m(x;H)$ is a consistent estimator of $D_m(x)$ and the bias term is of the standard order of magnitude for this type of problems. Meanwhile, the variance will be penalized when both H is large and data are sparse in the neighborhood of x. Our pairwise estimator nearly achieves the optimal rate (i.e., $\sqrt{I\!N|H|H}$) of convergence for this type of problem in a single step.

2.4 Estimating the unknown function

Although the focus of this paper is on the estimation of marginal effects (i.e., $D_m(\cdot)$), we could be also interested in the estimation of $m(\cdot)$. However, in this context of fixed effects, this is not an easy task. Our pairwise transformation deletes the unobserved heterogeneity at the cost of eliminating $m(x)$ from the regression model. In a two-dimensional setting, Lee et al. (2019) and Qian and Wang (2012) propose to use additional normalization assumptions and marginal integration techniques, respectively.

Here we extend the marginal integration technique developed in Qian and Wang (2012) to the multidimensional setting. Using the pairwise transformation in (2.7) to remove the unobserved individual heterogeneities, i.e., μ_i and γ_j, we have

$$
Y_{ijt} - Y_{ijs} = m(X_{ijt}) - m(X_{ijs}) + \lambda_t - \lambda_s + \epsilon_{ijt} - \epsilon_{ijs}, \quad i = 1,\ldots,N_1, \quad j = 1,\ldots,N_2, \quad t = 1,\ldots,t, \quad t < s,
$$

and taking differences across i to eliminate both λ_t and λ_s,

$$\ddot{Y}_{ijts} = g(X_{ijt}, X_{ijs}, X_{i(j-1)t}, X_{i(j-1)s}) + \ddot{\epsilon}_{ijt}, \quad i = 1, \ldots, N_1, \quad j = 1, \ldots, N_2, \quad t = 1, \ldots, t, \quad t < s, \tag{2.11}$$

where $\ddot{Y}_{ijts} = Y_{ijt} - Y_{ijs} - Y_{i(j-1)t} + Y_{i(j-1)s}$, and $\ddot{\epsilon}_{ijts} = \epsilon_{ijt} - \epsilon_{ijs} - \epsilon_{i(j-1)t} + \epsilon_{i(j-1)s}$. We then define $g(\cdot) : I\!R^{4d} \to I\!R^d$ as an additive function that satisfies

$$g(u, v_1, v_2, v_3) = m(u) - m(v_1) - m(v_2) + m(v_3), \quad u, v_1, v_2, v_3 \in I\!R^d.$$

Following Qian and Wang (2012), we propose to estimate $g(u, v_1, v_2, v_3)$ using multivariate kernel smoothing methods, say Nadaraya-Watson (see Nadaraya (1964) or Watson (1964)) or local linear (see Fan and Gijbels (1995b) or Ruppert and Wand (1994), among others). In particular, the local linear estimator of $g(u, v_1, v_2, v_3)$ solves the following criterion function for θ_0,[10]

$$\begin{aligned} \min_{\theta_0,\theta_1,\theta_2,\theta_3,\theta_4} \sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\sum_{t=1}^{T-1}\sum_{s=1+t}^{T} &[\ddot{Y}_{ijts} - \theta_0 - \theta_1^\top(X_{ijt} - u) - \theta_2^\top(X_{ijs} - v_1) - \theta_3^\top(X_{i(j-1)t} - v_2) - \theta_4^\top(X_{i(j-1)s} - v_3)]^2 \\ &\times K_{H_0}(X_{ijt} - u)K_{H_0}(X_{ijs} - v_1)K_{H_0}(X_{i(j-1)t} - v_2)K_{H_0}(X_{i(j-1)s} - v_3), \end{aligned} \tag{2.12}$$

where $K_{H_0}(u) = |H_0|^{-1/2}K(H_0^{-1/2}u)$ and H_0 is a $d \times d$ bandwidth matrix whose elements may be different from those in H.

Let $\ddot{Y} = [\ddot{Y}_{ijts}]$ be a column vector, $\Gamma = [1, (X_{ijt}-u)^\top, (X_{ijs}-v_1)^\top, (X_{i(j-1)t}-v_2)^\top, (X_{i(j-1)s}-v_3)^\top]$ be a $(1+4d)$-column matrix, and let $W_0 = diag\{K_{H_0}(X_{ijt}-u)K_{H_0}(X_{ijs}-v_1)K_{H_0}(X_{i(j-1)t}-v_2)K_{H_0}(X_{i(j-1)s} - v_3)\}$. Assuming that $\Gamma^\top W_0\Gamma$ is invertible, the solution to the minimization problem (2.12) for $\widehat{\theta}_0$ (hence $\widehat{g}(u, v_1, v_2, v_3)$) is

$$\widehat{g}(u, v_1, v_2, v_3; H_0) = \widehat{\theta}_0 = e_1^\top(\Gamma^\top W_0\Gamma)^{-1}\Gamma^\top W_0\ddot{Y}, \tag{2.13}$$

where e_1 is a $(1 + 4d)$-dimensional vector whose first element is 1 and the rest are zeros. Then, we proceed to estimate $m(\cdot)$ by marginally integrating $\widehat{g}(u, v_1, v_2, v_3)$,

$$\widehat{m}(u; H_0) = \int_{\mathcal{X}} \widehat{g}(u, v_1, v_2, v_3; H_0)q(v_1)q(v_2)q(v_3)dv_1dv_2dv_3, \tag{2.14}$$

where $q(v)$ is a pre-determined positive weighting function.

To develop the asymptotic theory, we assume the following:

Assumption C1 *$q(\cdot)$ are defined on the compact support $\mathcal{X} \in I\!R^d$, twice continuously differentiable, and $\int_{\mathcal{X}} m(u)q(u)du = 0$.*

[10]Note that in the above minimization problem, we only use non-overlapped cross-sections of j. More precisely, we use only the differencing between j and $j-1$, but we do not include the difference between other distinct pairs of j_1 and j_2, where $|j_1 - j_2| > 1$. Therefore, we have dropped more than half of the sample related with the individuals j, so an efficiency effect over the resulting estimator is expected. If we included all possible differences, the resulting samples are no longer independent in cross-sections and it would be necessary to resort to U-statistics techniques in order to obtain the main asymptotic properties of the local linear estimator. Therefore, if the estimation of the level function were the primary objective of the paper, the U-statistics solution would be recommended.

Assumption C2 *For each i and j, all joint densities of X_{ijt} are bounded from above and from zero and are continuously differentiable.*

Assumption C3 $H_0^{1/2} = \widetilde{H}_0^{1/2}(N_1 N_2^*)^{-1/(4+d)}$, *where $\widetilde{H}_0^{1/2}$ is a diagonal matrix with positive constant on the diagonal and $N_2^* = N_2$ if N_2 is even, else $N_2^* = (N_2 - 1)/2$.*

Note that Assumption C1 contains some normalization conditions that enables us to identify $q(\cdot)$. If the integral of $q(\cdot)$ is not equal to one, then the unknown function $m(\cdot)$ may be identified up to a multiplicative constant. Note that if we do not impose $\int m(u)q(u)du = 0$, then $m(\cdot)$ may be identified up to an additive constant. Further, conditions C2-C3 are fairly standard assumptions in the marginal integration literature.

Let $f_4(u, v_1, v_2, v_3)$ denote the joint density of $(X_{ij1}, X_{ij1+\kappa}, X_{i(j-1)1}, X_{i(j-1)(1+\kappa)})$, for $\kappa \in \{1, \ldots, (T-1)\}$, and $v^* = v_1 + v_2 - v_3$. The following theorem collects the asymptotic properties results of the estimator $\widehat{m}(u)$ defined in (2.14).

Theorem 4 *Let u be an interior point of supp(f) and let Assumptions A1-A5, A7, and C1-C3 hold. Given a fixed T and as $N_1 \to \infty$ and/or $N_2^* \to \infty$, we have*

$$\sqrt{N_1^* N_2^* |H_0|^{1/2}}\left(\widehat{m}(u; H_0) - m(u) - B(u, H_0)\right) \quad \xrightarrow{d} \quad \mathcal{N}(0, V(u)),$$

where

$$B(u, H_0) = \frac{1}{2}\mu_2^d(K)\left[tr\{H_0 \mathcal{H}_m(u)\} - \int_{\mathcal{X}} tr\{H_0 \mathcal{H}(v^*)\} q(v^*) dv^*\right] + o_p(tr\{H_0\}),$$

$$V(u) = \frac{R^d(K)}{T}\sum_{\kappa=1}^{T-1}\left(1 - \frac{\kappa}{T}\right)\left[\sigma^2(x, x) - \sigma_{1,(1+\kappa)}(x, x)\right]\left(\int_{\mathcal{X}} \frac{f_4(u, v_1, v_2, v_3) q^2(v_1) q^2(v_2) q^2(v_3)}{\left(\sum_{\kappa=1}^{T-1}\left(1 - \frac{\kappa}{T}\right) f_4(u, v_1, v_2, v_3)\right)^2} dv_1 dv_2 dv_3\right).$$

The proof of this theorem is a straightforward extension of the proof for Theorems $2-3$ in Qian and Wang (2012) and hence omitted. Further, the above results could be extended to the particular case in which both $I\!N$ and T are large, but that is beyond the scope of this paper as we are primarily concerned with the gradient and not the level function.

Finally, it is well-known that we may implement the marginal integration in (2.14) by numerical integration methods such as Simpson's or Trapezoidal rules. However, using these methods with multidimensional data can be computational cumbersome since $O(N_1 N_2)$ operations would be required in order to estimate each point in the dimension of interest. Alternatively, similar to Qian and Wang (2012), we propose to use the sample version of (2.14) of the form

$$\widehat{m}_s(u; H_0) = \frac{2}{N_1 N_2^* T^2 (T-1)} \sum_{i=1}^{N_1}\sum_{j=1}^{N_2^*}\sum_{t'=1}^{T}\sum_{t=1}^{T-1}\sum_{s=t+1}^{T} \widehat{g}(u, X_{ijt'}, X_{i(j-1)t}, X_{i(j-1)s}; H_0). \quad (2.15)$$

Note this estimator behaves asymptotically the same as in (2.14) when $g(\cdot)$ is the density of X_{ijt} (e.g., see Qian and Wang (2012) for a deeper explanation in a two-dimensional panel data setting).

3 Bandwidth selection

The bandwidth term plays a crucial role in estimation of the gradient function. Choosing a large bandwidth reduces the variance of the nonparametric estimates, but at the cost of enlarging its bias. In order to solve this trade-off, a large literature has focused on the development of appropriate bandwidth selection techniques for the conditional mean (see Fan and Gijbels (1995a) or Müller et al. (1987), among others). Unfortunately, these types of techniques are scarce when our interest is the gradient function in a panel data framework.

In this section, we propose a data-driven bandwidth technique based on the gradient of an unknown regression function in a panel data context. For the sake of simplicity, we assume $\sigma_{1,(1+\kappa)}(x,x)=0$. In particular, we want to choose $H = diag(h_1,\ldots,h_d)$ optimally in the sense that they minimize the estimation of mean squared error for the first order derivative functions of $m(x)$.[11] Let $\widehat{D}_m(x;H)$ be the pairwise estimator of $D_m(x)=\partial m(x)/\partial x^\top$ obtained in (2.10), we would like to choose H minimizing the following sample analog of the estimation of mean squared error:

$$CV(H) = \frac{1}{N_1 N_2}\binom{T}{2}^{-1}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\sum_{t=1}^{T-1}\sum_{s=t+1}^{T}\left[\widehat{D}_m(X_{ijts};H) - D_m(X_{ijts})\right]^2 M(X_{ijts}), \tag{3.1}$$

where $M(\cdot)$ is a weight function with bounded support that trims out data near the boundary of the support of x.

However, this objective function is infeasible given the oracle function, $D_m(x_{ijts})$, is unknown. To overcome this, authors typically employ the well-known solution proposed in Fan and Gijbels (1995a). However, this procedure may lead to poor bandwidth selection given that pilot smoothing parameters are required.

In order to avoid this drawback, we propose to choose a bandwidth parameter that is asymptotically equivalent to choosing a bandwidth that minimizes the objective function in (3.1). Specifically, following the proposal in Henderson et al. (2015), we construct a feasible objective function by replacing the oracle function by a consistent estimator from the set of local-polynomial estimators. We choose the local-cubic (LC) estimator, given that its bias is of sufficiently smaller-order relative to the local-linear estimator. Instead of considering the whole vector of the derivative function, $D_m(x)$, we consider each partial derivative separately, $D_{m_\ell}(x) = \partial m(x_1,\ldots,x_d)/\partial x_\ell$ for $\ell = 1,\ldots,d$. Without loss of generality we will focus on the case of $\ell = 1$.

The feasible objective function we propose to minimize is

$$CV(H) = \frac{1}{I\!N}\binom{T}{2}^{-1}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\sum_{t=1}^{T-1}\sum_{s=t+1}^{T}\left[\widehat{D}_{m_1}(X_{ijts};H) - \widehat{D}_{m_1,LC}(X_{ijts};H)\right]^2 M(X_{ijts}), \tag{3.2}$$

where $\widehat{D}_{m_1,LC}(X_{ijts};H)$ is the corresponding pairwise local-cubic estimator for $D_{m_1}(x) = \partial m(x_1,\ldots,x_d)/\partial x_1$.

[11]In practice, authors use a diagonal bandwidth matrix.

In order to show the asymptotic equivalence between the local-cubic and the local-linear estimator, we assume the following:

Assumption D1 *The map $m(\cdot)$ is four times continuously differentiable at x in the interior of $\mathcal{X}$ with bounded derivatives.*

Denote $Bias^0(\cdot)$ and $Var^0(\cdot)$ as the leading bias and variance terms of the corresponding estimator. Using Assumption D1 and following a similar procedure as in Theorem 1, as $I\!N \to \infty$ and $T \to \infty$, the conditional leading bias of these two estimators are

$$\begin{aligned}
Bias^0(\widehat{D}_{m_1}(x,H)) &= \left[\frac{(\mu_4(K)-\mu_2(K))}{2\mu_2(K)\sum_{\kappa=1}^{T-1} f_{X_{ij1},X_{ij(1+\kappa)}}(x,x)} m_1''(x)\sum_{\kappa=1}^{T-1} f'_{X_{ij1},X_{ij(1+\kappa)}}(x,x) + \frac{\mu_4(K)}{6\mu_2(K)} m_1'''(x)\right] h_1^2 \\
&+ \frac{\mu_2(K)}{2\sum_{\kappa=1}^{T-1} f_{X_{ij1},X_{ij(1+\kappa)}}(x,x)} \sum_{\ell\neq 1}^{d}\sum_{\kappa=1}^{T-1}\left[\mu_4(K)m_\ell'''(x) - \mu_2^2(K)m_\ell'''(x)\right] D_f(x) h_\ell^2 \\
&+ \mu_2(K)\sum_{\ell\neq 1}^{d}\sum_{\kappa=1}^{T-1}\frac{f_{X_{\ell,ij1},X_{\ell,ij(1+\kappa)}}(x,x)}{f_{X_{1,ij1},X_{1,ij(1+\kappa)}}(x,x)} m_\ell'''(x) h_\ell^2, \\
Bias^0(\widehat{D}_{m_1,LC}(x,H)) &= O_p(\|H\|^2),
\end{aligned}$$

where $\|H\| = \sum_{\ell=1}^{d} h_\ell^2$ and we denote $m_\ell'(x) = \partial m(x)/\partial x_\ell$, $m_1''(x) = \partial^2 m(x)/\partial x_1^2$, $m_\ell''(x) = \partial^2 m(x)/\partial x_t x_\ell$, $m_1'''(x) = \partial^3 m(x)/\partial x_1^3$, and $m_\ell'''(x) = \partial^3 m(x)/\partial x_t \partial x_{t'} \partial x_\ell$.

The conditional leading variances are

$$\begin{aligned}
Var^0(\widehat{D}_m(x,H)) &= \frac{\sigma^2(x,x)R_2^{d-1}(K)R(K)}{I\!NTh_1^4h_2^2\cdots h_d^2\mu_2^d(K)}\left(\sum_\kappa f_{X_{ij1},X_{ij(1+\kappa)}}(x,x)\right)^{-1}(1+o_p(1)), \\
Var^0(\widehat{D}_{m,LC}(x,H)) &= \frac{R^{d-1}(K)\sigma^2(x,x)\varrho_1}{I\!NTh_1^4h_2^2\cdots h_d^2\varrho_2^2} H^{-1}\left(\sum_\kappa f_{X_{ij1},X_{ij(1+\kappa)}}(x,x)\right)^{-1}(1+o_p(1)),
\end{aligned}$$

respectively, where $\varrho_1 = \mu_6^{2d}(K)R_2(K) + \mu_4^{2d}(K)R_6(K) - 2\mu_6^d(K)\mu_4^d(K)R_4(K)$, and $\varrho_2 = \mu_2^d(K)\mu_6^d(K) - \mu_4^{2d}(K)$.

By the same reasoning as in Hall et al. (2007) and Racine and Li (2004), the leading term of $CV(H)$ in (3.1) is given by

$$\begin{aligned}
CV_1^0(H) &= \int \left\{\left[Bias^0(\widehat{D}_{m_1}(x,H))\right]^2 + Var^0\left(\widehat{D}_m(x,H)\right)\right\} f(x)M(x)dx \\
&= \int [B(x)]^2 f(x)M(x)dx + \frac{V_1(K)R^{d-1}(K)}{I\!NTh_1^4h_2^2\cdots h_d^2}\int \mathcal{V}(x)f(x)M(x)dx, \qquad (3.3)
\end{aligned}$$

where $V_1(K) = R_2(K)/\mu_2^{2d}(K)$ and $\mathcal{V}(x) = \sigma^2(x,x)/\sum_\kappa f_{X_{ij1},X_{ij(1+\kappa)}}(x,x)$. Similarly, the leading term of $CV_1(H)$ in (3.2) is

$$\begin{aligned} CV_{1,LC}^0(H) &= \int \left\{ \left[Bias^0(\widehat{D}_{m_1}(x,H)) - Bias^0(\widehat{D}_{m_1,LC}(x,H)) \right]^2 + Var^0 \left(\widehat{D}_{m_1}(x,H) - \widehat{D}_{m_1,LC}(x,H) \right) \right\} \\ &\times \ f(x)M(x)dx \\ &= \int [B(x)]^2 f(x)M(x)dx + \frac{V_{1,3}(K)R^{d-1}(K)}{I\!N T h_1^4 h_2^2 \cdots h_d^2} \int \mathcal{V}(x)f(x)M(x)dx, \end{aligned} \tag{3.4}$$

where

$$V_{1,3}(K) = \frac{R_2(K)}{\mu_2^{2d}(K)} + \frac{\varrho_1}{\varrho_2^2} - \frac{2(R_2(K)\mu_6^d(K) - R_4(K)\mu_4^d(K))}{\mu_2^d(K)\varrho_2}.$$

We have shown that the bias of $\widehat{D}_{1,LC}(x,H)$ is asymptotically negligible and the only element of the local-cubic estimator which appears in the asymptotic expansion of the objective function is the variance of the difference between these two estimators. Furthermore, in the limit, the variance of this difference behaves (up to a constant depending on the kernel) the same as if we had the oracle gradient. Therefore, bandwidths selected replacing the oracle gradient with the local-cubic estimator are asymptotically equivalent to those selected with the unknown oracle gradient (see Henderson et al. (2015) for a deeper discussion in the cross-sectional setting).

Only for expositional simplicity, lets assume that the bandwidths are equal across d and let $h_{0,opt}$ and $h_{0,cubic}$ denote the values of h that minimize (3.3) and (3.4), respectively. In this special case, we have that

$$\begin{aligned} h_{0,opt} &= \left(\frac{V_1(K)R^{d-1}(K) \int \mathcal{V}(x)f(x)M(x)dx}{\int [B(x)]^2 f(x)M(x)dx} \right)^{1/8} I\!N T^{-1/8}, \\ h_{0,cubic} &= \left(\frac{V_{1,3}(K)R^{d-1}(K) \int \mathcal{V}(x)f(x)M(x)dx}{\int [B(x)]^2 f(x)M(x)dx} \right)^{1/8} I\!N T^{-1/8}. \end{aligned}$$

Therefore, $h_{0,cubic} = (V_{1,3}(K))/V_1(K))^{1/(T+6)} h_{0,opt}$. Let $\widetilde{h}_{cubic}$ denote the value of h that minimizes (3.2), $\widetilde{h}$ is corrected by the following expression

$$\widehat{h}_{cubic} = (V_1(K)/V_{1,3}(K))^{1/(T+6)} \widetilde{h}_{cubic}.$$

Finally, under some regularity conditions similar to those in Hall et al. (2007), it is relatively straightforward to show that

$$\widehat{h}_{cubic}/h_{0,opt} \xrightarrow{p} 1.$$

In the case of an Epanechnikov kernel function, it is relatively easy show that $V_1/V_{1,3} = 660/9065$. Note that these bandwidths have been chosen when T is large and $d = 1$ (single dimension X). However, it is relatively easy to extend our results for T fixed and $d > 1$, but we omit these results for sake of brevity.

4 Lower-dimensional and interactive fixed effects models

As we have argued, the proposed estimators are very general and can handle essentially any dimension panel. These are outlined for the parametric case by Balazsi et al. (2015) and we list several cases here for completeness. These include the commonly used one-way and two-way models

$$\begin{aligned} Y_{it} &= m\left(X_{it}\right) + \mu_i + \epsilon_{it}, \\ Y_{it} &= m\left(X_{it}\right) + \mu_i + \lambda_t + \epsilon_{it}, \end{aligned}$$

where, for $t < s$, the corresponding pairwise transformations are

$$\begin{aligned} \widetilde{Y}_{its} &= Y_{it} - Y_{is}, \\ \widetilde{Y}_{its} &= Y_{it} - Y_{is} - \overline{Y}_t + \overline{Y}_s, \end{aligned}$$

where the regressors (X) and errors (ϵ) are transformed similarly.

The two cases above are obvious, so it is probably more practical to consider fixed effects which are unique to (i, j), (i, t) and/or (j, t) as these are popular in empirical research. In particular, the corresponding nonparametric expression of the multidimensional panel data models (2.2)-(2.6) are the following

$$\begin{aligned} Y_{ijt} &= m\left(X_{ijt}\right) + \mu_{ij} + \epsilon_{ijt}, \\ Y_{ijt} &= m\left(X_{ijt}\right) + \mu_{ij} + \lambda_t + \epsilon_{ijt}, \\ Y_{ijt} &= m\left(X_{ijt}\right) + \gamma_{jt} + \epsilon_{ijt}, \\ Y_{ijt} &= m\left(X_{ijt}\right) + \mu_{ij} + \gamma_{jt} + \epsilon_{ijt}, \\ Y_{ijt} &= m\left(X_{ijt}\right) + \mu_{ij} + \gamma_{jt} + \lambda_{it} + \epsilon_{ijt}, \end{aligned}$$

where, for $t < s$, $i < k$, and $j < l$, the corresponding pairwise transformations are

$$\begin{aligned} \widetilde{Y}_{ijts} &= Y_{ijt} - Y_{ijs}, \\ \widetilde{Y}_{ijts} &= Y_{ijt} - Y_{ijs} - \overline{Y}_t + \overline{Y}_s, \\ \widetilde{Y}_{ijts} &= Y_{ijt} - Y_{kjt}, \\ \widetilde{Y}_{ijts} &= Y_{ijt} - Y_{kjt} - \overline{Y}_{ij} + \overline{Y}_{kj}, \\ \widetilde{Y}_{ijts} &= Y_{ijt} - \overline{Y}_{jt} - \overline{Y}_{it} + \overline{Y}_t - Y_{ijs} + \overline{Y}_{js} + \overline{Y}_{is} - \overline{Y}_s, \end{aligned}$$

where the regressors (X) and errors (ϵ) are transformed similarly. Once we have obtained the transformed variables $\widetilde{X}_{ijt}$ and $\widetilde{Y}_{ijt}$, we can similarly apply the estimation procedure developed in Subsection 2.3. Further, it is relatively straightforward to extend these to higher-dimensional panels.[12] We leave these extensions to given applications of our estimators.[13]

5 Finite sample properties

In order to assess the finite-sample performance of the pairwise estimator presented in Section 2 and the GBCV bandwidth selection procedure proposed in Section 3, several simulations will be performed. To cover commonly used situations, we focus on the following two scenarios: (a) two-dimensional (two-way) fixed effects panel data models and (b) three-dimensional (three-way) fixed effects panel data models.

5.1 Two-dimensional model

We begin by considering the following data generating process (DGP)

$$Y_{it} = m(X_{it}) + \mu_i + \lambda_t + \epsilon_{it}, \qquad i = 1, \dots, N, \quad t = 1, \dots, T \tag{5.1}$$

where X_{it} is a random variable generated such that $X_{it} = 0.5X_{i(t-1)} + \xi_{it}$ and ξ_{it} is an $i.i.d.$ random variable normally distributed with zero mean and variance 1. We generate $\mu_i = \vartheta_i + c_0\overline{X}_i$ and $\lambda_t = \vartheta_t + c_0\overline{X}_t$, where ϑ_i and ϑ_t are $U[-1, 1]$ random variables, $c_0 = 0.5$ controls the magnitude of the fixed effects, $\overline{X}_i = T^{-1}\sum_{t=1}^{T} X_{it}$ and $\overline{X}_t = N^{-1}\sum_{i=1}^{N} X_{it}$.

The error term ϵ_{it} is generated as $N(0, 1)$ and two functional forms for $m(\cdot)$ are considered:

DGP1 $m_1(X_{it}) = (1/3)X_{it}^3$,

DGP2 $m_2(X_{it}) = 1.5X_{it}^2/(1 + X_{it}^2)$.

We use 1000 replications, the number of time periods T is either 3, 5, or 10, and the number of cross-sections N is either 100 or 300. Given Assumption A5 (bounded support kernel), we use Epanechnikov kernel functions.[14]

We consider both fixed bandwidths and the GBCV bandwidth selection mechanism for the pairwise estimator proposed in Section 2. Further, we use the Average Mean Squared Error (AMSE) as a measure of the performance of our estimator that is computed as

$$AMSE(\widehat{D}_m(x; h)) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(\widehat{D}_m(x; h) - D_m(x; h))^2.$$

[12]For the four dimensional panel we proposed before, the transformation would be $\widetilde{Y}_{ijls} = Y_{ijlt} - Y_{ijls} - \overline{Y}_t + \overline{Y}_s$, for $t < s$, where $\overline{Y}_t$ and $\overline{Y}_s$ are the corresponding cross-sectional means.

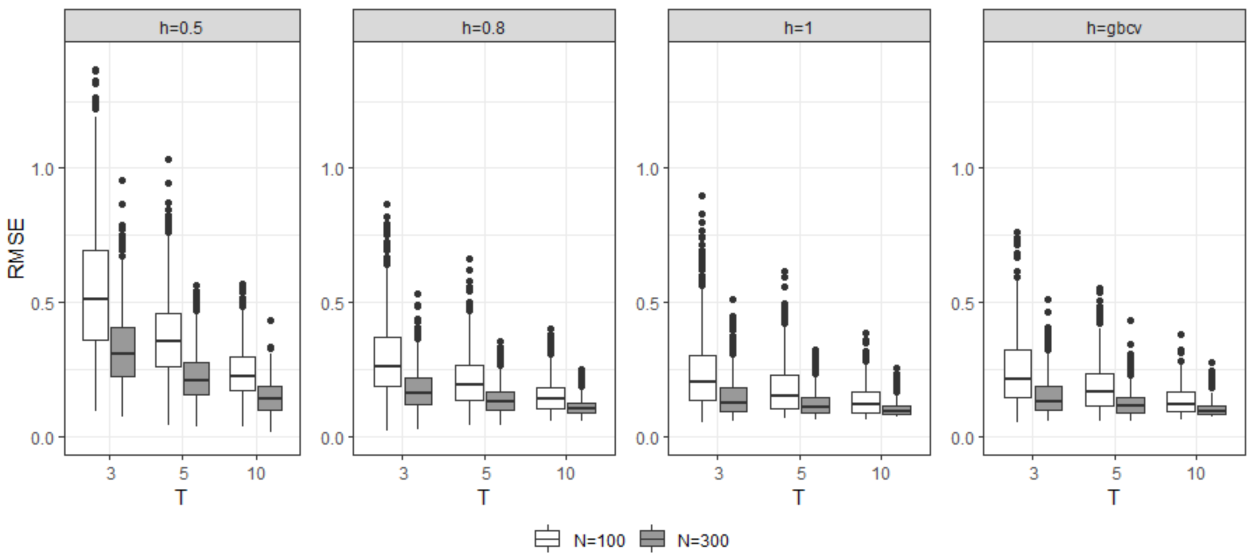
[13]A potentially interesting extension would be to consider factor models. However, this is beyond the scope of this paper.

[14]The use of bounded kernel functions was purely for convenience in the asymptotic development. It is worth noting that we also ran each of the simulations with Gaussian kernels and our estimators performed well. In fact, we found non-trivial improvements as compared to Lee et al. (2019). When using an Epanechnikov kernel function, the performance of the two estimators were nearly indistinguishable. This bodes well for the pairwise estimator as it is obtained under less restrictive assumptions. These results are available upon request.

Figures 1-2 depict boxplots of the 1000 AMSE values of the pairwise gradient estimators for $m_1(\cdot)$ and $m_2(\cdot)$, respectively. In addition, for the sake of comparison, our Supplementary Material contains the estimated bias, standard deviation (Std), and square root (RAMSE) of the AMSE of the pairwise estimator and other nonparametric estimators proposed in the literature to estimate the gradient functions.

Figures 1 and 2 illustrate the consistency of our pairwise estimator for the gradient of the unknown function as all AMSE values converge toward zero as the sample size increases. In addition, the proposed GBCV bandwidth selection mechanism performs well. The resulting AMSE is considerably smaller as the time dimension increases.

Figure 1: Boxplots of the 1000 AMSE values for the gradient estimators of $m_1(\cdot)$ for both fixed and GBCV selected bandwidths in two-dimensional panel data models.



Further, when we assess the finite sample performance of our proposed GBCV bandwidth selection procedure the true unknown gradient is known, so a comparison to the oracle setting is feasible. Our performance criterion is the average squared error (ASE),

$$ASE(\widehat{\beta}_{LL,A}) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\widehat{D}_{m,LL}(X_{it};h) - D_{m,A}(X_{it})\right)^2,$$

where $\widehat{D}_{m,LL}(\cdot)$ is the local-linear pairwise estimator of $D_m(\cdot) = \partial m(\cdot)/\partial z$ and $D_{m,A}(\cdot)$ is one of the estimators from: (i) the local-quadratic estimator, (ii) the local-cubic estimator, and (iii) the true gradient function. Tables 1-2 present percentiles of the estimated ASE for the bandwidths selected by GBCV using local-quadratic, local-cubic, and the infeasible estimator over the 1000 simulations for DGP1 and DGP2, respectively.

The median ASEs in Tables 1-2 provide insight into the general behavior of the proposed bandwidth selection mechanism while the extreme deciles provide insight into the tail performance

Figure 2: Boxplots of the 1000 AMSE values for the gradient estimators of $m_2(\cdot)$ for both fixed and GBCV selected bandwidths in two-dimensional panel data models.

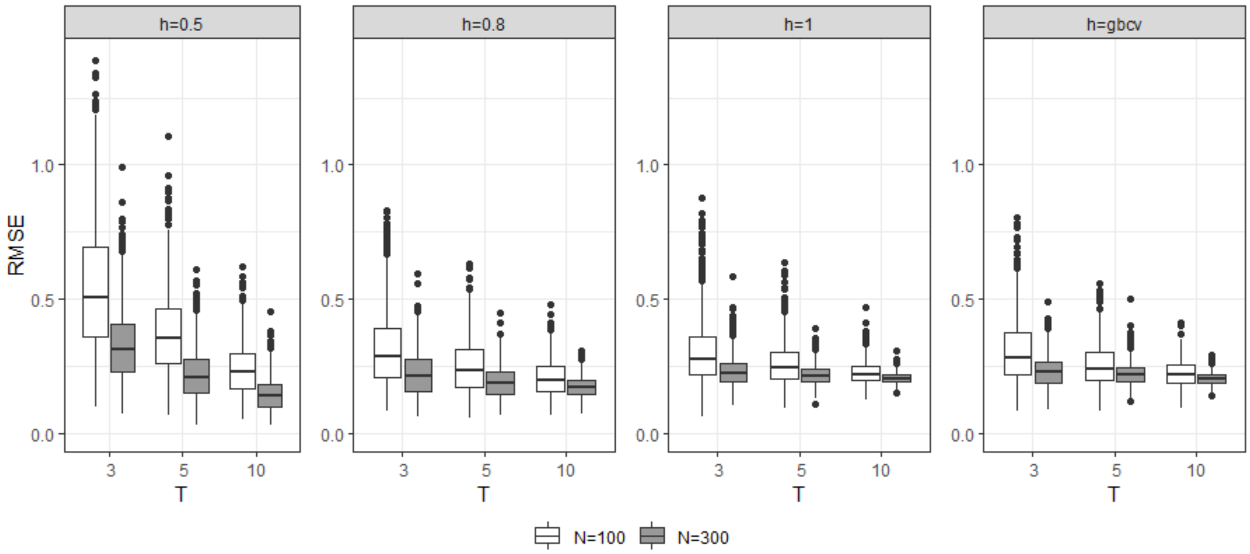


Table 1: ASE for DGP1 for GBCV selected bandwidths over 1000 simulations.

N	Consistent estimator	T=3	T=5	T=10
100	Local-quadratic	[0.192,0.975,3.523]	[0.090,0.495,1.531]	[0.032,0.183,0.727]
	Local-cubic	[0.036,0.212,0.959]	[0.012,0.096,0.362]	[0.006,0.042,0.172]
	Infeasible true $D_m(\cdot)$	[0.011,0.046,0.196]	[0.008,0.028,0.097]	[0.007,0.014,0.053]
200	Local-quadratic	[0.076,0.370,1.175]	[0.033,0.165,0.547]	[0.012,0.085,0.253]
	Local-cubic	[0.010,0.076,0.280]	[0.006,0.032,0.012]	[0.003,0.016,0.048]
	Infeasible true $D_m(\cdot)$	[0.006,0.017,0.061]	[0.006,0.012,0.037]	[0.006,0.009,0.021]

Note: Numbers in brackets are the $10th$*,* $50th$*, and* $90th$ *percentiles of ASE across* 1000 *simulations, respectively.*

Table 2: ASE for DGP2 for GBCV selected bandwidths over 1000 simulations.

N	Consistent estimator	T=3	T=5	T=10
100	Local-quadratic	[0.178,1.065,3.423]	[0.114,0.506,1.601]	[0.036,0.219,0.716]
	Local-cubic	[0.039,0.253,0.923]	[0.017,0.132,0.461]	[0.008,0.063,0.264]
	Infeasible true $D_m(\cdot)$	[0.026,0.081,0.244]	[0.029,0.058,0.131]	[0.029,0.048,0.085]
200	Local-quadratic	[0.082,0.396,1.332]	[0.044,0.205,0.603]	[0.033,0.114,0.349]
	Local-cubic	[0.017,0.101,0.360]	[0.011,0.066,0.214]	[0.008,0.054,0.144]
	Infeasible true $D_m(\cdot)$	[0.027,0.052,0.099]	[0.029,0.047,0.073]	[0.031,0.041,0.056]

Note: Numbers in brackets are the $10th$*,* $50th$*, and* $90th$ *percentiles of ASE across* 1000 *simulations, respectively.*

of a given method across the simulations. The tables show that (at the median) the GBCV local-cubic estimator dominates the local-quadratic for all sample sizes and both DGPs and is very close (in terms of performance) to using the true gradients. Further, the gains from using local-cubic increase as the sample size increases. Here we argue that the local-cubic version of the GBCV delivers bandwidths which behave as though one deployed the infeasible, known gradient of the unknown conditional mean. In other words, our simulations results confirm the theoretical conclusions of Section 3.

5.2 Three-dimensional model

For the three-dimensional setting, $i = 1, \ldots, N_1$, $j = 1, \ldots, N_2$, and $t = 1, \ldots, T$, we consider the following DGPs

$$\begin{aligned} Y_{ijt} &= m(X_{1ijt}) + \mu_i + \gamma_j + \lambda_t + \epsilon_{ijt}, & (5.2) \\ Y_{ijt} &= m(X_{1ijt}, X_{2ijt}) + \mu_i + \gamma_j + \lambda_t + \epsilon_{ijt}, & (5.3) \end{aligned}$$

where X_{1ijt} and X_{2ijt} are random variables generated such that $X_{1ijt} = 0.5X_{1ij(t-1)} + \xi_{1ijt}$ and $X_{2ijt} = 0.5\xi_{2ijt} + 0.5\xi_{3ijt}$, where ξ_{1ijt}, ξ_{2ijt}, and ξ_{3ijt} denote *i.i.d.* random variables normally distributed with zero mean and variance 1. Several functional forms for $m(\cdot)$ are considered:

DGP3 $m_1(X_{1ijt}) = (1/3)X_{1ijt}^3$,

DGP4 $m_2(X_{1ijt}) = 1.5X_{1ijt}^2/(1 + X_{1ijt}^2)$,

DGP5 $m_1(X_{1ijt}, X_{2ijt}) = (1/3)(X_{1ijt} + X_{2ijt})^3$,

DGP6 $m_2(X_{1ijt}, X_{2ijt}) = 1.5(X_{1it} + X_{2ijt})^2/\left[1 + (X_{1ijt} + X_{2ijt})^2\right]$,

where DGP3 and DGP4 are univariate problems whereas DGP5 and DGP6 are multivariate.

We generate $\mu_i = \vartheta_i + c_0\overline{X}_{1i}$, $\gamma_j = \vartheta_j + x\overline{X}_1$, and $\lambda_t = \vartheta_t + c_0\overline{X}_{1t}$, where ϑ_i, ϑ_j, and ϑ_t are $U[-1, 1]$, $c_0 = 0.5$ controls the magnitude of the fixed effects, whereas $\overline{X}_{1i} = (N_2T)^{-1}\sum_{jt} X_{1ijt}$, $\overline{X}_{1t} = (N_1N_2)^{-1}\sum_{ij} X_{1it}$, and $\overline{X}_{1j} = (N_1T)^{-1}\sum_{it} X_{1ijt}$. The number of cross-sections ($I\!N = N_1N_2$) are varied from $(N_1, N_2) = (15, 10)$ to $(N_1, N_2) = (25, 20)$, the number of time periods T is either 3, 5 or 10. We repeat each experiement 1000 times. We again use Epanechnikov kernel functions and consider both fixed and GBCV bandwidths and AMSE is used as a performance measure of the pairwise estimator that is computed as

$$AMSE(\widehat{D}_m(x; h)) = \frac{1}{N_1N_2T}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\sum_{t=1}^{T}\left(\widehat{D}_m(x; h) - D_m(x; h)\right)^2.$$

Figures 3 and 4 give the results for the univariate cases ($d = 1$). The consistency of the pairwise estimator for each bandwidth is corroborated. Again, as expected from our theoretical findings, the AMSEs collapse to zero as the sample size increases in all the cases considered.

Figures 5 and 6 collect the results for the multivariate DGP (5.3). As expected, the curse of dimensionality is evident although the estimator still performs well in this setting. In summary, the results here along with the theoretical results suggest that the pairwise estimator performs well in practice.

Figure 3: Boxplots of the 1000 AMSE values for the gradient estimators of $m_1(\cdot)$ for both fixed and GBCV selected bandwidths in univariate three-dimensional panel data models.

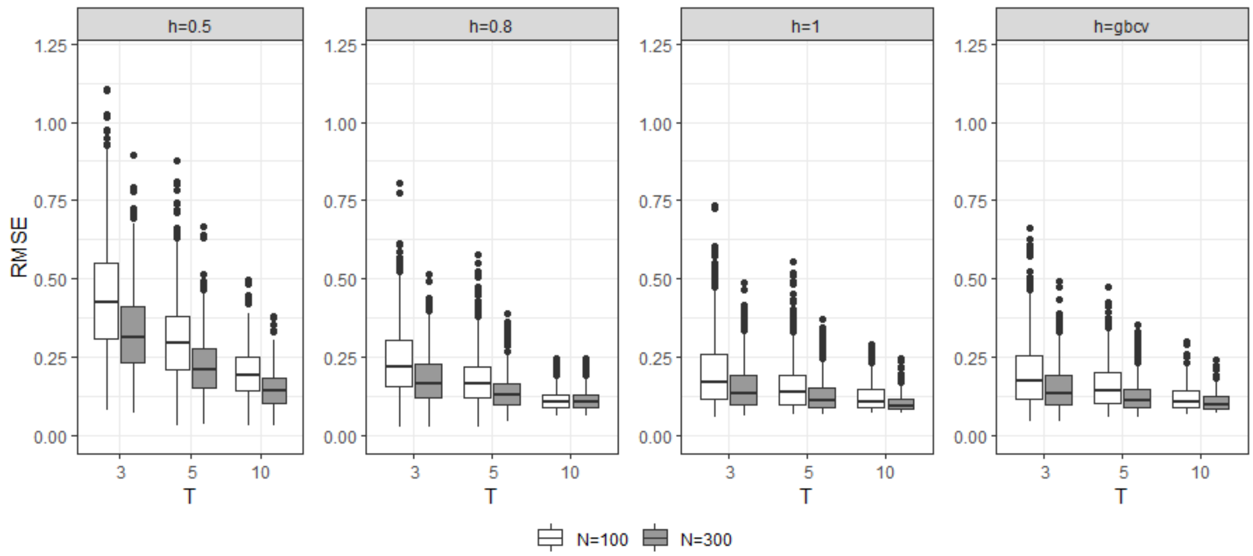
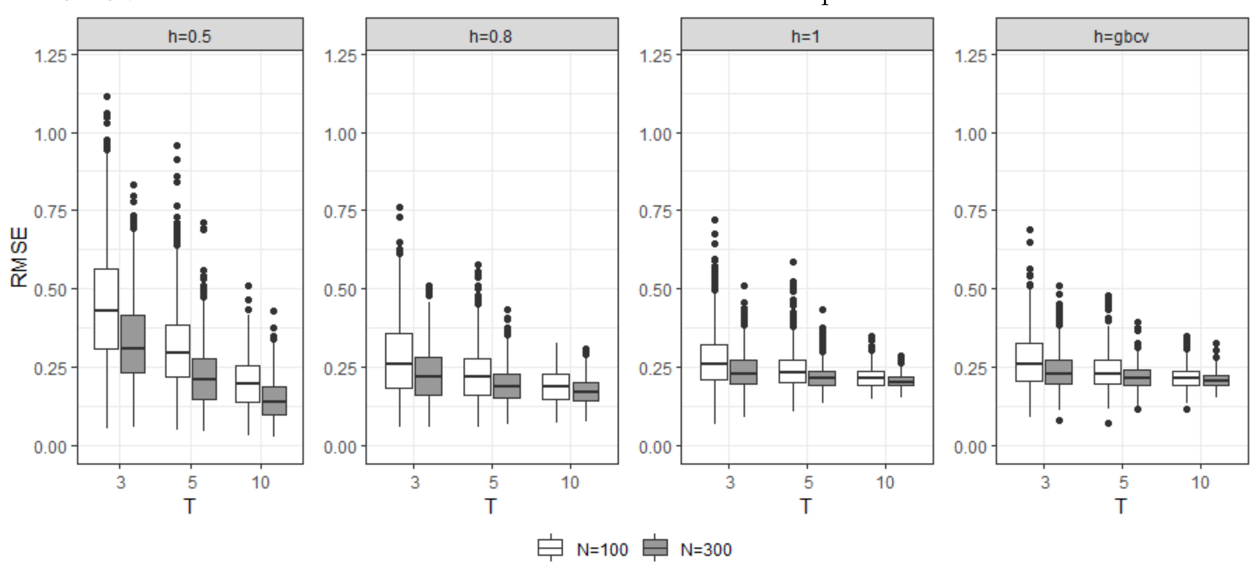
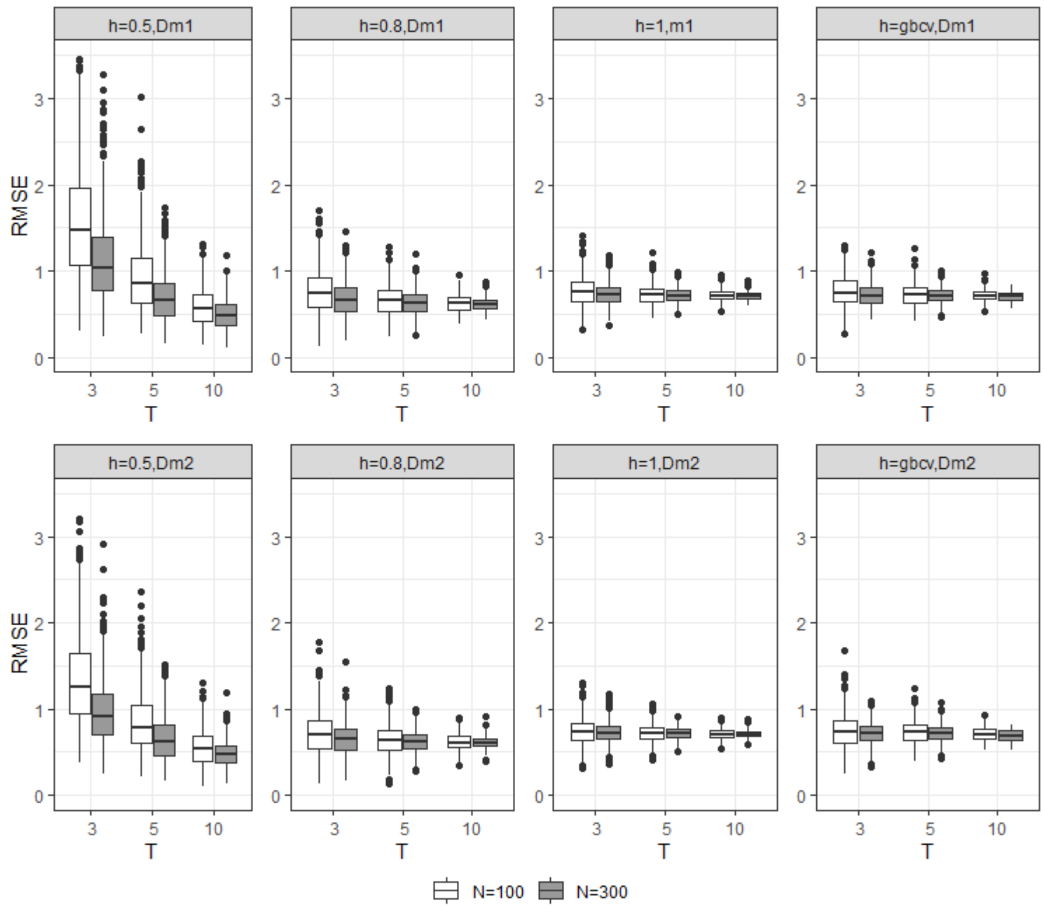


Figure 4: Boxplots of the 1000 AMSE values for the gradient estimators of $m_2(\cdot)$ for both fixed and GBCV selected bandwidths in univariate three-dimensional panel data models.



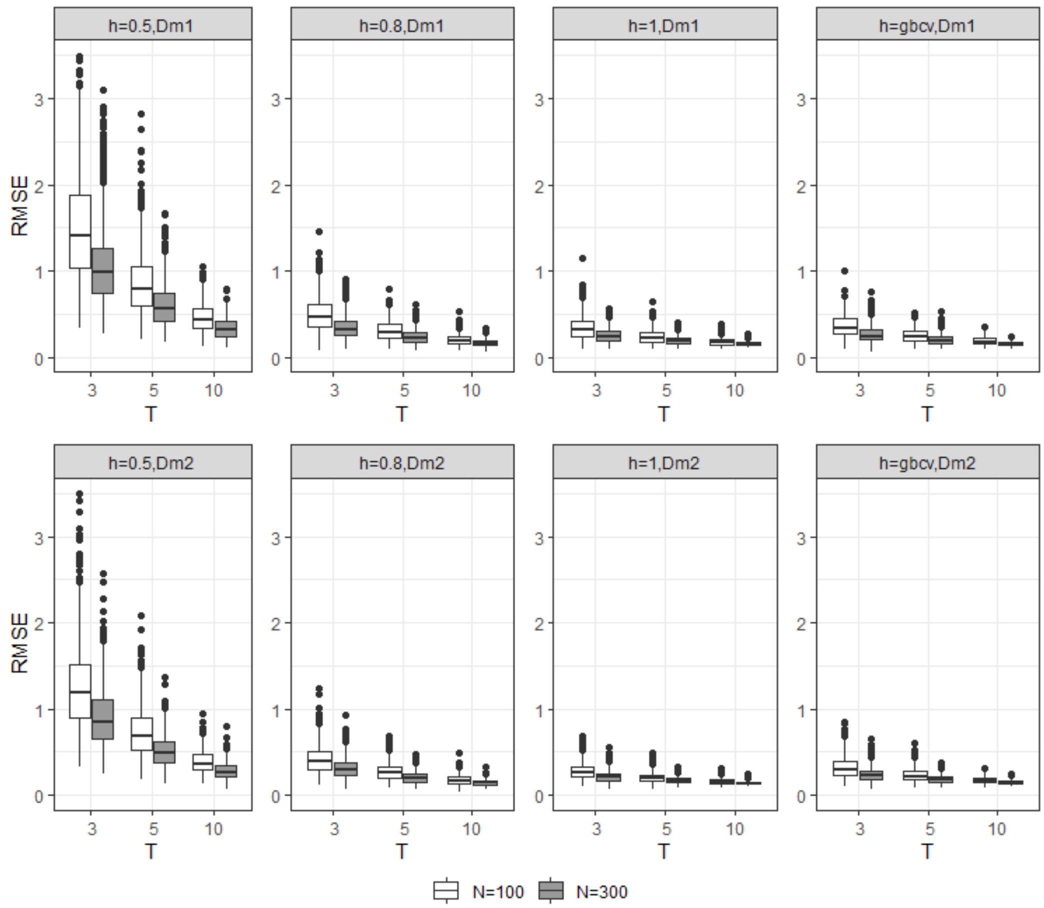
Finally, in order to assess the finite sample performance of the proposed bandwidth selection mechanism in the multidimensional setting, we consider DGP3 and DGP4. Again, our perform-

Figure 5: Boxplots of the 1000 AMSE values for the gradient estimators of $m_1(\cdot)$ for both fixed and GBCV selected bandwidths in multivariate three-dimensional panel data models.



Note: The boxplots at the top depict the results for the gradient estimator of $m_1(\cdot)$ with respect to X_1 (i.e. $D_{m_1}(\cdot)$), whereas the boxplots for the gradient estimator of $m_1(\cdot)$ with respect to X_2 (i.e. $D_{m_2}(\cdot)$) are collected at the bottom.

Figure 6: Boxplots of the 1000 AMSE values for the gradient estimators of $m_2(\cdot)$ for both fixed and GBCV selected bandwidths in multivariate three-dimensional panel data models.



Note: The boxplots at the top depict the results for the gradient estimator of $m_1(\cdot)$ with respect to X_1 (i.e. $D_{m_1}(\cdot)$), whereas the boxplots for the gradient estimator of $m_1(\cdot)$ with respect to X_2 (i.e. $D_{m_2}(\cdot)$) are collected at the bottom.

ance criterion is ASE:

$$ASE(\widehat{D}_{m_{LL},A}) = \frac{1}{N_1 N_2 T} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \sum_{t=1}^{T} \left(\widehat{D}_{m,LL}(X_{ijt}; h) - D_{m,A}(X_{ijt}) \right)^2$$

and the resulting percentiles of estimated ASE are collected in Tables 3 and 4.

Table 3: ASE for DGP3 for GBCV selected bandwidths over 1000 simulations.

N	Consistent estimator	T=3	T=5	T=10
100	Local-quadratic	[0.109,0.650,2.079]	[0.114,0.506,1.601]	[0.027,0.130,0.444]
	Local-cubic	[0.020,0.135,0.506]	[0.017,0.132,0.461]	[0.005,0.032,0.103]
	Infeasible true $D_m(\cdot)$	[0.008,0.030,0.121]	[0.029,0.058,0.131]	[0.006,0.011,0.030]
200	Local-quadratic	[0.075,0.340,1.135]	[0.044,0.205,0.603]	[0.016,0.086,0.232]
	Local-cubic	[0.011,0.079,0.281]	[0.011,0.066,0.214]	[0.003,0.015,0.072]
	Infeasible true $D_m(\cdot)$	[0.007,0.018,0.070]	[0.029,0.048,0.073]	[0.006,0.010,0.027]

Note: Numbers in brackets are the 10*th,* 50*th, and* 90*th percentiles of ASE across* 1000 *simulations, respectively.*

Table 4: ASE for DGP4 for GBCV selected bandwidths over 1000 simulations.

N	Consistent estimator	T=3	T=5	T=10
100	Local-quadratic	[0.154,0.695,2.233]	[0.096,0.369,1.085]	[0.045,0.195,0.575]
	Local-cubic	[0.022,0.150,0.599]	[0.015,0.106,0.396]	[0.007,0.066,0.236]
	Infeasible true $D_m(\cdot)$	[0.026,0.067,0.152]	[0.026,0.052,0.102]	[0.029,0.046,0.069]
200	Local-quadratic	[0.091,0.389,1.216]	[0.048,0.209,0.626]	[0.023,0.108,0.360]
	Local-cubic	[0.013,0.102,0.398]	[0.009,0.069,0.248]	[0.009,0.043,0.142]
	Infeasible true $D_m(\cdot)$	[0.029,0.052,0.104]	[0.029,0.045,0.070]	[0.031,0.042,0.059]

Note: Numbers in brackets are the 10*th,* 50*th, and* 90*th percentiles of ASE across* 1000 *simulations, respectively.*

We find similar conclusions in Tables 3 and 4 as we do in Tables 1 and 2. We conclude that the proposed GBCV bandwidth selection mechanism performs well in finite samples.

6 Application: The price of rental housing and housing vouchers

Beginning with the Housing Act of 1937, the U.S. federal government has operated a large number of programs with the intent to improve the housing of low-income households. While there have been many variants of these programs, they generally fall into three different categories: (1) government ownership and operation of newly built low-income housing, (2) government contracts with private parties to build (or improve existing) and operate low-income housing and (3) subsidies to eligible households for private market housing. The first two are supply side programs and the latter is a demand side program. There has been a substantial increase in demand side programs as of late as it has been argued that demand side programs are generally more cost effective and do not result in the same concentrations of poverty as traditional supply based programs.

All programs have costs. While the substantial price tag (around \$40 billion yearly in means-tested housing programs and another \$6 billion in tax credits to low-income households) is understood, one potential negative consequence is that these subsidies on the demand side may lead to an increase in the price of rental housing for households that do not receive the subsidy. The literature to this point is unclear on this particular issue (Olsen (2003)). For example, Susin (2002) found (for the period 1974-1993) that increased housing vouchers (subsidies) increased the price of rental housing for households who did not receive vouchers, whereas a large social experiment in the 1970's found no statistically significant effect. While these studies are useful, they each suffer from potential endogeneity and hence their results may be suspect.

Eriksen and Ross (2015) use a panel dataset to estimate the effect of a large, arguably exogenous, increase in the number of housing vouchers (between 2000-2002) in order to identify the effect of increasing the supply of housing vouchers on short-term rents. In their full-sample, they find an elasticity which is close to zero (both nominally and statistically). When splitting the sample at arbitrary values of a relevant variable (ratio of the rent of a rental unit in the base year to the US Department of Housing and Urban Development fair market rent for that metropolitan statistical area in the same year), they find statistically significant negative effects for those units which were initially 80% below the 1997 ratio and statistically significant positive elasticities for those units which were between 80% and 120% of that ratio.

Our goal here is to take the data from Eriksen and Ross (2015) to determine if a more flexible approach can yield additional insight into the heterogeneous effects of housing vouchers on rents. Here we hope to avoid splitting the sample arbitrarily by adopting a semiparametric approach whereby we can obtain an elasticity for each unit in each time period.[15] Given the number of attributes of each rental unit (i.e., number of regressors), we extend our econometric methodology to the case of a partially linear model. Specifically, we will model housing vouchers nonparametrically, but will allow the remaining attributes to enter linearly. This extension addresses the curse of dimensionality problem.

Using a balanced sample of housing units, we find both negative and positive elasticities of housing rents with respect to housing vouchers. While we find some evidence of positive and significant elasticities and while those positive and significant elasticities are primarily in the range suggested by Eriksen and Ross (2015), the majority of the evidence suggests primarily negative elasticities throughout the sample. For the positive elasticities, our point estimates suggest that most of the positive elasticities reside in the Western United States and for housing units that are more supply inelastic.

The remainder of this section is as follows: Section 6.1 develops the partially linear version of

[15]Eriksen and Ross (2015) attempt to further model heterogeneity by interacting the log of vouchers with a fifth-order polynomial of the log of the supply elasticity.

our estimator for the case of a single nonparametric regressor and three-dimensional fixed effects. Section 6.2 discusses the data from Eriksen and Ross (2015), and Section 6.3 presents both the parametric and semiparametric results.

6.1 Partially linear model

Given the number of regressors, it seems prudent to try to reduce the dimension of our nonparametric problem. Here we consider the extension of our model to one that is partially linear. We considered the partially linear approach of Robinson (1988), but elected to adopt the profile least-squares approach designed for partially linear varying coefficient models in Fan and Huang (2005). Note that estimation via the partially linear approach of Robinson (1988) produced similar results. Here we will model rents on vouchers nonparametrically, but will allow the relationship between rents and the remaining regressors to be linear.

Our extension of Equation (2.7) to the partially linear case is given as

$$Y_{ijt} = m\left(X_{ijt}\right) + Z_{ijt}^{\top}\beta + \mu_i + \gamma_j + \lambda_t + \epsilon_{ijt}, \tag{6.1}$$

where $i = 1, 2, \ldots, N_1$, $j = 1, 2, \ldots, N_2$, and $t = 1, 2, \ldots, T$. In our setting, X_{ijt} is scalar ($d = 1$) and will be the log of housing vouchers and Z_{ijt} is a vector of $q = 7$ different attributes of the rental property. Our interest will lie in the gradient of $m(\cdot)$, i.e., $D_m(\cdot)$, and the finite parameter vector β. As before, we allow μ_i, γ_j, and/or λ_t to be correlated with X_{ijt}, but also allow them to be correlated with Z_{ijt}. This requires us to transform Y_{ijt} with respect to the cross-sectional and time dimensions.

Similar to Equation (2.8), we define, the pairwise transformed regression model as

$$\widetilde{Y}_{ijts} = \widetilde{m}(X_{ijts}) + \widetilde{Z}_{ijts}^{\top}\beta + \widetilde{\epsilon}_{ijts}, \quad i = 1, \ldots, N_1, \quad j = 1, \ldots, N_2, \quad t = 1, \ldots, T, \quad t < s, \tag{6.2}$$

where $\widetilde{Y}_{ijts} = (Y_{ijt} - \overline{Y}_t) - (Y_{ijs} - \overline{Y}_s)$. $\widetilde{Z}_{ijts}$ and $\widetilde{\epsilon}_{ijts}$ are defined similarly, whereas

$$\widetilde{m}(X_{ijts}) = m(X_{ijt}) - m(X_{ijs}) - \frac{1}{N_1 N_2}\sum_{ij}[m(X_{ijt}) - m(X_{ijs})].$$

Noting that X_{ijt} is scalar in our application, so we can estimate the gradient function, $D_m(\cdot)$, using the following locally weighted linear regression

$$\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\sum_{t=1}^{T-1}\sum_{s=t+1}^{T}\left[\left(\widetilde{Y}_{ijts} - \widetilde{Z}_{ijts}^{\top}\beta\right) - \widetilde{X}_{ijts}D_m\left(x\right)\right]^2 K_h\left(X_{ijt} - x\right)K_h\left(X_{ijs} - x\right), \tag{6.3}$$

where the kernel functions, $K_h(\cdot)$ are as we defined before except that we use the notation h instead of H because in the application, H is a scalar bandwidth term.

Taking the first-order condition of (6.3) with respect to the gradient function $D_m(x)$, for a given β, leads to the infeasible estimator

$$\widetilde{D}_m(x;h) = \frac{\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\sum_{t=1}^{T-1}\sum_{s=t+1}^{T} K_h\left(X_{ijt}-x\right)K_h\left(X_{ijs}-x\right)\widetilde{X}_{ijts}(\widetilde{Y}_{ijts}-\widetilde{Z}_{ijts}^{\top}\beta)}{\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\sum_{t=1}^{T-1}\sum_{s=t+1}^{T} K_h\left(X_{ijt}-x\right)K_h\left(X_{ijs}-x\right)\widetilde{X}_{ijts}^2}. \quad (6.4)$$

In order to overcome this situation and obtain a consistent estimate of the gradient of our unknown function, we follow Fan and Huang (2005). We first derive a closed form solution for the estimator of β and then plug that consistent estimate into (6.4). It is perhaps easier to derive the result in matrix notation and so we rewrite (6.2) as

$$\widetilde{Y} - \widetilde{Z}\beta = \widetilde{M} + \widetilde{\epsilon}.$$

We define $NT_P = N_1N_2T(T-1)/2$. Then, $\widetilde{M} = \left[\widetilde{m}(X_{1121}),\ldots,\widetilde{m}(X_{N_1N_2T(T-1)})\right]^{\top}$, $\widetilde{Y} = (\widetilde{Y}_{1121},\ldots,\widetilde{Y}_{N_1N_2T(T-1)})^{\top}$, and $\widetilde{\epsilon} = (\widetilde{\epsilon}_{1121},\ldots,\widetilde{\epsilon}_{N_1N_2T(T-1)})^{\top}$ are $NT_P\times 1$ vectors, and $\widetilde{Z} = (\widetilde{Z}_{1121},\ldots,\widetilde{Z}_{N_1N_2T(T-1)})^{\top}$ is a $NT_P\times q$ dimensional matrix.

In this framework, the solution to the problem (6.3) in matrix form is given by

$$\widetilde{D}_m(x;h) = (\widetilde{X}^{\top}W_x\widetilde{X})^{-1}\widetilde{X}^{\top}W_x(\widetilde{Y}-\widetilde{Z}\beta),$$

where $\widetilde{X} = (\widetilde{X}_{1121},\ldots,\widetilde{X}_{N_1N_2T(T-1)})$ is a $NT_p\times 1$ vector and W_x is a $NT_P\times NT_P$ block diagonal matrix such as

$$W_x = diag\left\{K_h(X_{112}-x)K_h(X_{111}-x),\ldots,K_h(X_{N_1N_2T}-x)K_h(X_{N_1N_2(T-1)}-x)\right\}.$$

Approximating the elements of $\widetilde{M}$ through a Taylor expansion and following similar steps as in Fan and Huang (2005), the estimator of $\widetilde{M}$ is obtained as

$$\widehat{\widetilde{M}} = S(\widetilde{Y}-\widetilde{Z}^{\top}\beta),$$

where $S = (S_{1121}^{\top},\ldots,S_{N_1N_2T(T-1)}^{\top})^{\top}$ is a $NT_P\times NT_P$ smoothing matrix whose ijt-th element is

$$S_{ijts} = \widetilde{X}_{ijts}(\widetilde{X}^{\top}W_{x_{ijts}}\widetilde{X})^{-1}\widetilde{X}^{\top}W_{x_{ijts}}.$$

Using this result, it is relatively straightforward to obtain the estimator of the finite dimensional parameter as

$$\widehat{\beta} = \left[\widetilde{Z}^{\top}\left(I_{NT_P}-S\right)^{\top}\left(I_{NT_P}-S\right)\widetilde{Z}\right]^{-1}\widetilde{Z}^{\top}\left(I_{NT_P}-S\right)^{\top}\left(I_{NT_P}-S\right)\widetilde{Y}, \quad (6.5)$$

where I_{NT_P} is an identity matrix of dimension NT_P. With this estimate of β in hand, we obtain the feasible estimator of the gradient function as

$$\widehat{D}_m(x;h) = \frac{\sum_{ijts} K_h\left(X_{ijt}-x\right) K_h\left(X_{ijs}-x\right) \widetilde{X}_{ijts}(\widetilde{Y}_{ijts} - \widetilde{Z}_{ijts}^{\top}\widehat{\beta})}{\sum_{ijts} K_h\left(X_{ijt}-x\right) K_h\left(X_{ijs}-x\right) \widetilde{X}_{ijts}^2}. \tag{6.6}$$

Note that in practice we must fix the value of x in order to obtain the pairwise estimators of $\widetilde{\beta}$ and/or $\widehat{\beta}$. In both cases, we choose the median value from the data vector, but note that the choice of this value did not significantly impact the estimated value of β when we restricted our choices to be within the interquartile range of the data.

Following a similar proof scheme as in Fan and Huang (2005) and Cai et al. (2019), it is relatively straightforward to show that $\widehat{\beta}$ is a consistent estimator of β. Therefore, the asymptotic properties obtained in Section 2.3 are valid for $\widehat{D}_m(x;h)$ in a partially linear framework.

6.2 Data

The data come directly from Eriksen and Ross (2015) and we will only discuss them briefly. Their primary source is the public-use version of the American Housing Survey (AHS). The U.S. Census Bureau conducts the AHS survey every two years with the goal of being nationally representative of all housing units in the U.S. Their original unbalanced sample for the years 1997, 1999, 2001 and 2003 includes $8,388$ rental housing units in 135 Metropolitan Statistical Areas (MSAs). Excluding rental units which were publicly-owned or rent restricted resulted in $24,721$ unit-year observations. We took the $2,713$ rental units which were observed in each of the four time periods as our sample to avoid issues relating to attrition. This resulted in $10,852$ unit-year observations.

The descriptive statistics for both samples can be found in Table 5. The first value is the sample mean of the reported variable and the value below it is the sample standard deviation of that variable. The first variable listed is our dependent variable, the log of the reported *Rent* and utility cost, which is \$1,013 per month in the full sample and \$988 per month in the balanced sample. The log of $Vouchers$ is our main regressor of interest and is defined as the estimated number of vouchers at the metropolitan level. Eriksen and Ross (2015) argue and provide evidence that the large increase in the number of vouchers between 2000 and 2002 was unrelated to past changes in rents and deem that this variable is likely exogenous. The log of per capita $Income$, the log of *Population* and log rental $Vacancy$ rates all come from the U.S. Census. The binary variables, *Rodents*, $Washer$, *Cracks* and *Sewage* are used to control for the presence of rodents, whether there was a washer or dryer, large cracks in the walls and if the sewage system had broken down in the previous year. Each set of regressions will also include unit, MSA and year fixed effects.[16]

[16]It is arguable that the MSA fixed effect is redundant given that rental units do not change MSAs over time. We also attempted to model a two-way fixed effects specification and found that the results are qualitatively similar. These are available upon request.

Table 5: Descriptive statistics for the full and balanced samples: Sample means for each variable (the first five are measured in logs) with the sample standard deviation underneath

	Full Sample	Balanced Sample
$Rent$	6.5568	5.5608
	0.5011	0.4845
$Vouchers$	9.0374	9.1516
	1.3970	1.3367
$Income$	10.5686	10.566
	0.1632	0.1690
$Population$	14.6675	14.5973
	1.0466	1.0212
$Vacancy$	2.0268	2.0064
	0.3690	0.3659
$Rodents$	0.1360	0.1413
	0.3428	0.3483
$Washer$	0.3864	0.3681
	0.4869	0.4823
$Cracks$	0.0778	0.0771
	0.2678	0.2668
$Sewage$	0.0198	0.0185
	0.1392	0.1348
N_1	8,388	2,713
N_2	135	135
T	4	4
Obs	24,721	10,852

6.3 Results

Here we present the results from our application. We first look at the parameter estimates from a pairwise parametric procedure (fully linear model) and then move to the analogous semiparametric results. Given that our semiparametric approach treats $Vouchers$ nonparametrically, it allows us to look at the distribution of the estimated elasticities of housing rents with respect to housing vouchers.

We note here that we were able to successfully replicate the results of Eriksen and Ross (2015) using both their Stata code as well as with our own code in `R`.

6.3.1 Parametric

The first column of numbers in Table 6 reports the parameter estimates from the pairwise parametric estimator for the balanced sample (analogous to Table 4 in Eriksen and Ross (2015)). The number below each parameter estimate is its corresponding standard error. It is important to first point out that while a majority of the point estimates from the full sample are similar to that of the balanced sample, there are some changes. The coefficient on the significant variables $Income$ and $Population$ are about 25% larger and 40% smaller in magnitude than in the balanced sample, respectively. Of the remaining variables, $Washer$ is significant, while $Vacancy$, $Rodents$, $Cracks$ and $Sewage$ are insignificant in each case.

6.3.2 Semiparametric

We begin our discussion of the semiparametric results by looking at the bandwidths calculated via the estimation method discussed in Section 3. The cross-validated bandwidth for the pairwise estimator is 0.3887. Note that this result is far below a few standard deviations (see Hall et al. (2007)) of the log of vouchers (1.3367) and hence we believe that substantial nonlinearities exists in the relationship between $Rent$ and $Vouchers$.

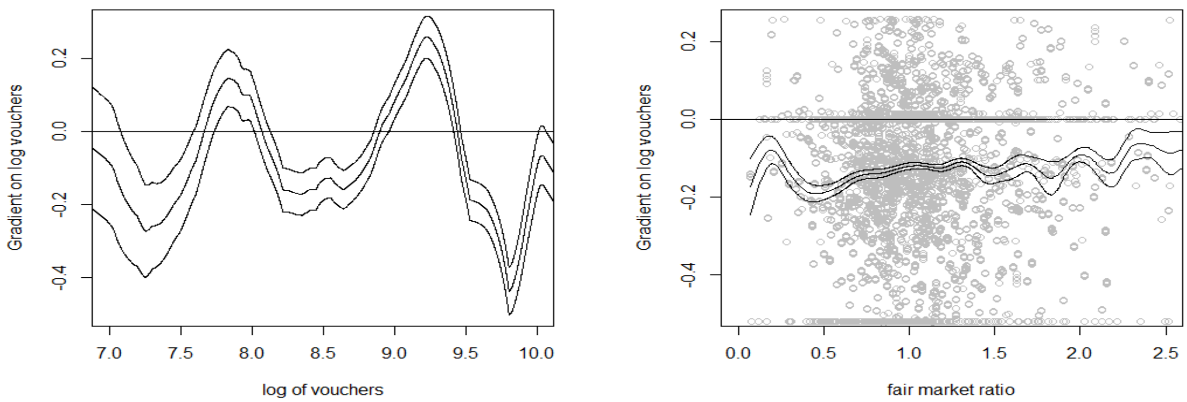
The parameter estimates (and corresponding standard errors) for the finite dimensional parameter vector β for the semiparametric estimator is given in the second column in Table 6. The parameter estimates for the semiparametric pairwise estimators are very close to their parametric counterparts. The most notable differences between the coefficient estimates are the coefficients on $Population$ which are more than twice the size of their parametric counterparts.

Given that we treat $Vouchers$ nonparametrically, it is not included in the finite dimensional parameter vector. We get unique elasticities for each rental unit in each year. Instead of giving an average or median estimate for each estimator, we plot the estimated gradient versus the log of vouchers over a grid of points in panels (a) and (b) of Figure 7 along with 95% pointwise confidence bounds obtained via the bootstrap procedure outlined in Härdle et al. (2004, p. 119).

The bootstrap confidence bounds show portions which are both positive and negative and cases of significance and insignificance of each. This may explain why the (global) parametric (mean) estimates are close to zero. The semiparametric estimates show significant nonlinearities

Table 6: Parameter estimates (with standard errors listed beneath) from both the parametric and semiparametric procedures: $Vouchers$ is treated nonparametrically in the semiparametric case and is not included in the finite dimensional parameter vector

	Parametric	Semiparametric
$Vouchers$	0.0091	
	0.0159	
$Income$	1.0682	1.1306
	0.1087	0.1077
$Population$	0.2245	0.5158
	0.0956	0.0630
$Vacancy$	-0.0093	-0.0105
	0.0131	0.0131
$Rodents$	0.0146	0.0159
	0.0078	0.0078
$Washer$	0.0341	0.0337
	0.0095	0.0095
$Cracks$	-0.0006	0.0002
	0.0091	0.0092
$Sewage$	0.0072	0.0089
	0.0172	0.0172
Unit fixed effects	Yes	Yes
MSA fixed effects	Yes	Yes
Year fixed effects	Yes	Yes
N_1	2,713	2,713
N_2	135	135
T	4	4
Obs	10,852	10,852



(a) Pairwise gradient vs. log vouchers

(b) Pairwise gradient vs. fair market ratio

Figure 7: Gradient estimate with respect to the log of vouchers versus either the log of vouchers or fair market ratio. In the right panel the scatter plot is given as well as a separate nonparametric regression fit to the see the relationship between the two variables. All plots include 95% pointwise confidence intervals.

present in the relationship. That being said, there is no clear picture of what is driving these results.

Given this puzzle in panel (a), we can further analyze the results by exploiting clues in Eriksen and Ross (2015). In their Table 5, they arbitrarily split the sample into different groups: rents less than 80% of the ratio of the 1997 rent divided by the US Department of Housing and Urban Development fair market rent before the voucher expansion and those between 80 and 120 percent. The semiparametric results generate unique elasticities for each rental unit in each time period and so we can simply look at the relationship between each elasticity and the ratio in question. Panel (d) of Figure 7 does just that. In each panel, we present a scatterplot of the two values (estimated elasticities of housing vouchers and the fair market rent ratio), along with a nonparametric fit (and relevant 95% pointwise confidence bounds). While we see positive elasticities in the range Eriksen and Ross (2015) allude to, the vast majority of the elasticities are negative. In fact, the nonparametric regression and confidence bounds are substantially below zero.

To study these differences further, we looked more closely at observations with positive elasticities and those with negative elasticities. We found the main differences between these groups to be with respect to region. Therefore, in Figure 8, we plotted the analogous scatter plots and nonparametric regression fits for each region: Midwest, North, South and West. The Midwest and North appear to have many elasticities both above and below zero while the South is primarily below zero and the West has a large portion above zero. In fact, the West is the only region whereby the nonparametric fits have portions which are significantly above zero.

Larger western cities tend to have the ocean, lots of local restrictions, and mountains. Southern cities are generally flat and seemingly can grow forever. This begs the question of whether or not this result is at least partially driven by vacancy rates as areas that are more supply inelastic (supply constrained) cannot build as quickly and easily when demand increases. Therefore, in Figure 9, we give plots analogous to those in Figure 8, but for low and high vacancy rates as well as for supply inelastic and supply elastic areas.[17] The results for both low vacancy and supply inelastic areas are quite similar. We see a majority of positive elasticies and substantial portions greater than zero. In fact, for low vacancy, the entire curve is above zero. This is stronger than the result shown in Eriksen and Ross (2015). For high vacancy and supply elastic markets, there are some observations both above and below zero, but the nonparametric fits tend to be below zero for all fair market ratios.

Overall, we showed that while many of the results from the parametric model held true, the heterogeneity of the elasticities was masked by assuming a single coefficient on housing vouchers. We found a large percentage of negative coefficients whereas the parametric model was unable to find this result without splitting the sample arbitrarily. We were able to find evidence that positive elasticities were more concentrated in the West where supply elasticities were more inelastic. Future research should look into what attributes are related to the sign of these elasticities as well as potential heterogeneity in other attributes.

7 Conclusion

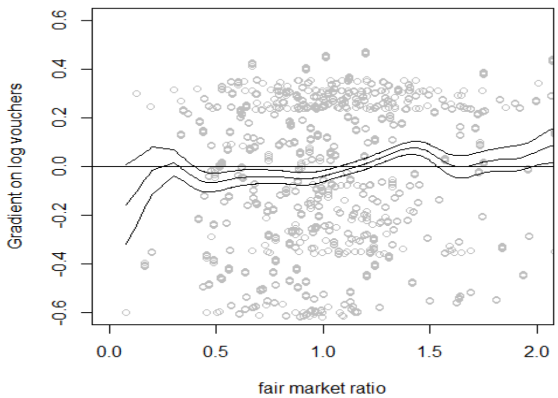
In this paper, we developed a nonparametric gradient estimator for multi-dimensional panel data models. This estimator can handle fixed effects of essentially any dimension as well as interactive effects. We developed the asymptotic results for our estimators and showed the finite sample properties via simulation. We further suggest a cross-validation method for data driven bandwidth selection. Finally, we extended our model to the partially linear setting and included an empirical application that related U.S. rental prices to housing vouchers and showed that negative elasticities tended to be in areas which were more supply elastic whereas the positive elasticities tended to be in areas which were more supply inelastic.

We believe these general results can be applied to many economic problems, and believe that in a given application, the estimator can easily be modified to handle the unobserved heterogeneity present in the model.

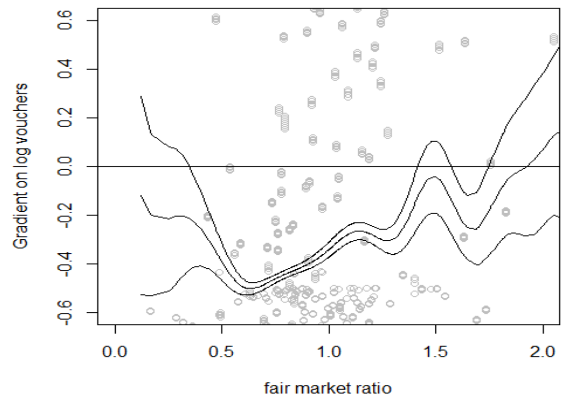
Acknowledgments

Financial support comes from the Programa Estatal de Generación de Conocimiento y Fortalecimiento Científico y Tecnológico del Sistema de I+D+i y del Programa Estatl de I+D+i Orientada a los Retos de la Sociedad/Spanish Ministry of Science and Innovation (Ref. PID2019-105986GB-C22). This work is part of the Research Project APIE 1/2015-17: "New methods for the empirical

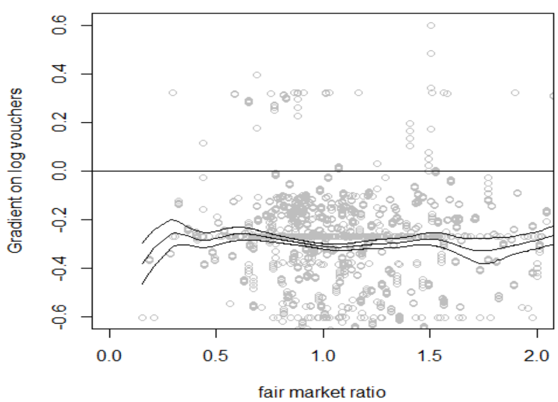
[17]The supply elastic and inelastic regions are only available for populations in excess of 500,000. This only leaves us with 94 of the 135 MSA areas. See Eriksen and Ross (2015) for more details.



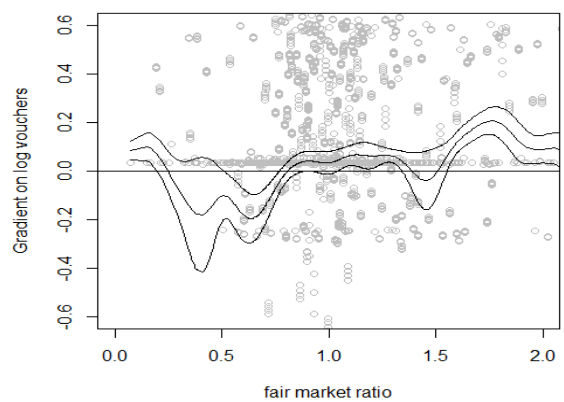
(a) Midwest



(b) North

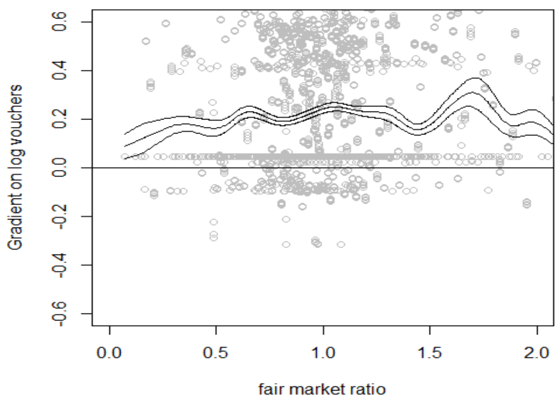


(c) South

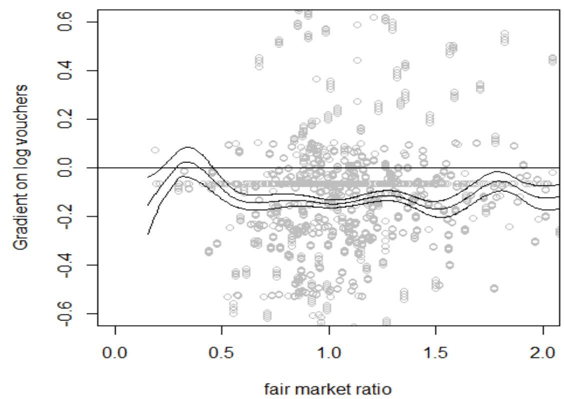


(d) West

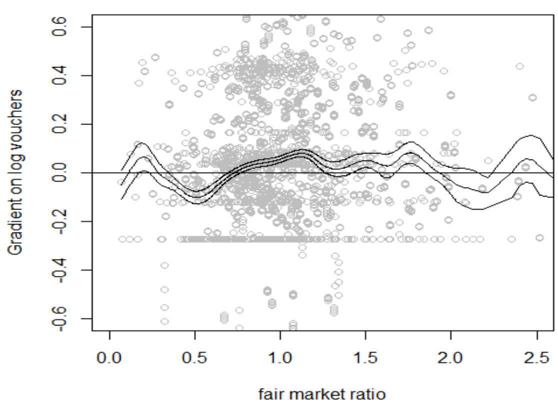
Figure 8: Pairwise gradient versus the ratio of the rent of a rental unit in the base year to the US Department of Housing and Urban Development fair market rent (fair market ratio) for different regions (Midwest, North, South and West). Each panel includes a scatter plot as well as a separate nonparametric regression fit to the see the relationship between the two variables. All plots include 95% pointwise confidence intervals.



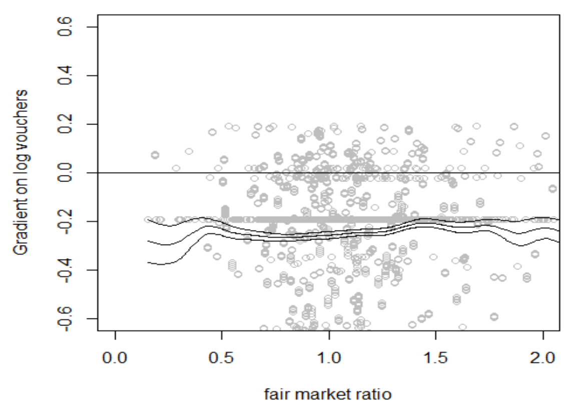
(a) Low vacancy



(b) High vacancy



(c) Supply inelastic



(d) Supply elastic

Figure 9: Pairwise gradient versus the ratio of the rent of a rental unit in the base year to the US Department of Housing and Urban Development fair market rent (fair market ratio) for low vacancy rates, high vacancy rates, supply inelastic and supply elastic regions (note that supply elasticities are only available for populations in excess of 500,000 - 93 of the 135 MSA areas). Each panel includes a scatter plot as well as a separate nonparametric regression fit to the see the relationship between the two variables. All plots include 95% pointwise confidence intervals.

analysis of financial markets" of the Santander Financial Institute (SANFI) of UCEIF Foundation resolved by the University of Cantabria and funded with sponsorship from Banco Santander.

Appendix

Before proceeding to the asymptotic analysis of the proposed estimator, we first establish the definition of a strongly mixing sequence. Let $\{\vartheta_t\}$ be a strictly stationary process and $\mathcal{F}_{t'}^{t}$ denote the sigma algebra generated by $(\vartheta_{t'}, \ldots, \vartheta_t)$ for $t' \leq t$. A process $\{\vartheta_t\}$ is said to be strongly mixing or α-mixing if

$$\alpha(\tau) = \sup_{t' \in \mathcal{N}} \{|P(A \bigcap B) - P(A)P(B)| : A \in \mathcal{F}_{-\infty}^{t'}, B \in \mathcal{F}_{t'+\tau}^{\infty}\} \to 0$$

as $\tau \to \infty$.

Proof of Theorem 1: In order to show the conditional MSE of our nonparametric estimator, some notation is needed. $D_m(x)$ is the first-order derivative d-dimensional vector of $m(\cdot)$ such that $D_m(x) = vec(\partial m(x)/\partial x^\top)$, $\mathcal{H}_m(x)$ is the $d \times d$ Hessian matrix of $m(\cdot)$, i.e., $\mathcal{H}_m(x) = \partial m(x)/\partial x \partial x^\top$, and $n \equiv N_1 N_2 T$. Further, we denote

$$\begin{aligned} \varpi_4(x, H) &= \int u_1 u_1^\top H^{1/2} \mathcal{H}_m(x) H^{1/2} u_1 u_1^\top K(u_1) du_1, \\ \varpi_2(x, H) &= \int u_1 u_{(1+\kappa)}^\top H^{1/2} \mathcal{H}_m(x) H^{1/2} u_{1(1+\kappa)} u_1^\top K(u_1) K(u_{1+\kappa})) du_1 du_{(1+\kappa)} \end{aligned}$$

as $d \times d$ positive and finite matrices.

Let $\mathbb{X} = (X_{111}, \ldots, X_{N_1 N_2 T})$ be the observed covariates sample. The conditional MSE of $\widehat{D}_m(x; H)$ is of the form

$$MSE[\widehat{D}_m(x; H)|\mathbb{X}] = tr(Var[\widehat{D}_m(x; H)|\mathbb{X}]) + Bias[\widehat{D}_m(x; H)|\mathbb{X}]^\top Bias[\widehat{D}_m(x; H)|\mathbb{X}], \tag{7.1}$$

where

$$\begin{aligned} Bias[\widehat{D}_m(x; H)|\mathbb{X}] &= E[\widehat{D}_m(x; H)|\mathbb{X}] - D_m(x) \\ Var[\widehat{D}_m(x; H)|\mathbb{X}] &= E\left[(\widehat{D}_m(x; H) - E[\widehat{D}_m(x; H)|\mathbb{X}])(\widehat{D}_m(x; H) - E[\widehat{D}_m(x; H)|\mathbb{X}])^\top |\mathbb{X}\right]. \end{aligned}$$

For the sake of simplicity, let us denote

$$K_{ijt} = K\left(H^{-1/2}(X_{ijt} - x)\right) \quad \text{and} \quad K_{ijs} = K\left(H^{-1/2}(X_{ijs} - x)\right).$$

By Assumption A2, the conditional expectation of (2.10) is

$$E[H^{1/2} \widehat{D}_m(x; H)|\mathbb{X}] = \left(\frac{1}{N_1 N_2 |H|} \sum_{ijts} K_{ijt} K_{ijs} \widetilde{X}_{h,ijts} \widetilde{X}_{h,ijts}^\top\right)^{-1} \frac{1}{N_1 N_2 |H|} \sum_{ijts} K_{ijt} K_{ijs} \widetilde{X}_{h,ijts}^\top \widetilde{m}(X_{ijts}), \tag{7.2}$$

where $\widetilde{X}_{h,ijts} = H^{-1/2}\widetilde{X}_{ijts}$ and

$$\widetilde{m}(X_{ijts}) = m(X_{ijt}) - m(X_{ijs}) - \frac{1}{N_1N_2}\sum_{ij}\left[m(X_{ijt}) - m(X_{ijs})\right].$$

Approximating $\widetilde{m}(X_{ijts})$ using the multivariate Taylor's theorem and rearranging terms,

$$\widetilde{m}(X_{ijts}) \simeq \widetilde{X}_{h,ijts}^{\top}H^{1/2}D_m(x) + \frac{1}{2}\widetilde{Q}_{ijts}(x) + \widetilde{R}_{ijts}(x), \tag{7.3}$$

where

$$\begin{aligned}
\widetilde{Q}_{ijts}(x) &= (X_{ijt}-x)^{\top}\mathcal{H}_m(x)(X_{ijt}-x) - (X_{ijs}-x)^{\top}\mathcal{H}_m(x)(X_{ijs}-x) \\
&- \frac{1}{I\!N}\sum_{ij}\Big[(X_{ijt}-x)^{\top}\mathcal{H}_m(x)(X_{ijt}-x) - (X_{ijs}-x)^{\top}\mathcal{H}_m(x)(X_{ijs}-x)\Big], \\
\widetilde{R}_{ijts}(x) &= (X_{ijt}-x)^{\top}\mathcal{R}_m(X_{ijt};x)(X_{ijt}-x) - (X_{ijs}-x)^{\top}\mathcal{R}_m(X_{ijs};x)(X_{ijs}-x) \\
&- \frac{1}{I\!N}\sum_{ij}\Big[(X_{ijt}-x)^{\top}\mathcal{R}_m(X_{ijt};x)(X_{ijt}-x) - (X_{ijs}-x)^{\top}\mathcal{R}_m(X_{ijs};x)(X_{ijs}-x)\Big],
\end{aligned}$$

where $I\!N = N_1N_2$.

Also, $\mathcal{R}_m(X_{ijt};x)$ is a residual term such as

$$\mathcal{R}_m(X_{ijt};x) = \int_0^1\left[\frac{\partial^2 m}{\partial x\partial x^{\top}}(x+\varphi(X_{ijt}-x)) - \frac{\partial^2 m(x)}{\partial x\partial x^{\top}}\right](1-\varphi)d\varphi \tag{7.4}$$

and φ is a weight function. A similar definition is given for $\mathcal{R}_m(X_{ijs};x)$.

Replacing (7.3) in (7.2), we obtain

$$H^{1/2}E[\widehat{D}_m(x;H)|\mathbb{X}] - H^{1/2}D_m(x) = \Psi_n^{-1}B_n + \Psi_n^{-1}R_n, \tag{7.5}$$

and using the expressions (2.10) and (7.2),

$$H^{1/2}\widehat{D}_m(x;H) - H^{1/2}E[\widehat{D}_m(x;H)|\mathbb{X}] = \Psi_n^{-1}U_n, \tag{7.6}$$

where

$$
\begin{aligned}
\Psi_n &= \binom{T}{2}^{-1} \frac{1}{I\!N|H|} \sum_{ijts} \widetilde{X}_{h,ijts} \widetilde{X}_{h,ijts}^{\top} K_{ijt} K_{ijs}, \\
B_n &= \binom{T}{2}^{-1} \frac{1}{2I\!N|H|} \sum_{ijts} \widetilde{X}_{h,ijts} \widetilde{Q}_{ijts}(x) K_{ijt} K_{ijs}, \\
R_n &= \binom{T}{2}^{-1} \frac{1}{I\!N|H|} \sum_{ijts} \widetilde{X}_{h,ijts} \widetilde{R}_{ijts}(x) K_{ijt} K_{ijs}, \\
U_n &= \binom{T}{2}^{-1} \frac{1}{I\!N|H|} \sum_{ijts} \widetilde{X}_{h,ijts} \widetilde{\epsilon}_{ijts} K_{ijt} K_{ijs}.
\end{aligned}
$$

In order to prove the results of this theorem, the following Lemmas are necessary which will be proved below.

Lemma 1 *Under Assumptions A1-A6,*

(i) $\Psi_n = \frac{4\mu_2^d(K)}{(T-1)} \sum_{\kappa=1}^{T-1} \left(1 - \frac{\kappa}{T}\right) f_{X_{ij1},X_{ij(1+\kappa)}}(x,x) I_d (1 + o_p(1))$,

(ii) $B_n = \frac{2}{(T-1)} \sum_{\kappa=1}^{T-1} \left(1 - \frac{\kappa}{T}\right) [\varpi_4(x,H) - \varpi_2(x,H)] H^{1/2} D_{f_\kappa}(x) + o_p(H^{3/2})$,

(iii) $R_n = o_p(H^{3/2})$.

Lemma 2

(i) *Under Assumptions A1-A6, as $I\!N \to \infty$ and fixed T,*

$$
\begin{aligned}
& I\!NT(T-1)|H| Var(U_n) \\
& \quad \xrightarrow{p} \quad 16 R_2^d(K) R^d(K) \sum_{\kappa=1}^{T-1} \left(1 - \frac{\kappa}{T}\right) (\sigma^2(x,x) - \sigma_{1,(1+\kappa)}(x,x)) f_{X_{ij1},X_{ij(1+\kappa)}}(x,x) I_d.
\end{aligned}
$$

(ii) *Under Assumptions A1-A5 and B1-B2, as $I\!N \to \infty$ and $T \to \infty$,*

$$
\begin{aligned}
& I\!NT(T-1)^2|H| Var(U_n) \\
& \quad \xrightarrow{p} \quad 16 R_2^d(K) R^d(K) \sum_{\kappa=1}^{T-1} (\sigma^2(x,x) - \sigma_{1,(1+\kappa)}(x,x)) f_{X_{ij1},X_{ij(1+\kappa)}}(x,x) I_d.
\end{aligned}
$$

Proof of Lemma 1: In order to prove the assertion (i), $\widetilde{X}_{ijts}$ can be rewritten as $\widetilde{X}_{ijts} = (X_{ijt} - X_{ijs}) - I\!N^{-1} \sum_{ij} (X_{ijt} - X_{ijs})$. Under Assumption A1, as $I\!N$ tends to infinity,

$\text{plim}\frac{1}{I\!N}\sum_{ij}(X_{ijt}-X_{ijs})=0$. Using this result and Assumption A1, we get

$$
\begin{aligned}
E(\Psi_n) &= \frac{4}{T(T-1)|H|}\sum_{ts}E[H^{-1/2}(X_{ijt}-X_{ijs})(X_{ijt}-X_{ijs})^\top H^{-1/2}K_{ijt}K_{ijs}]+o_p(1)\\
&= \frac{4}{T(T-1)|H|}\sum_{\kappa}(T-\kappa)E[H^{-1/2}(X_{ij1}-x)(X_{ij1}-x)^\top H^{-1/2}K_{ij1}K_{ij(1+\kappa)}]\\
&- \frac{4}{T(T-1)|H|}\sum_{\kappa}(T-\kappa)E[H^{-1/2}(X_{ij1}-x)(X_{ij(1+\kappa)}-x)^\top H^{-1/2}K_{ij1}K_{ij(1+\kappa)}]+o_p(1)\\
&= \frac{4}{T(T-1)}\sum_{\kappa}(T-\kappa)\int u_1u_1^\top f(X_{ij1}=x+H^{1/2}u_1,X_{ij(1+\kappa)}=x+H^{1/2}u_{(1+\kappa)})K(u_1)K(u_1)du_1du_{(1+\kappa)}\\
&- \frac{4}{T(T-1)}\sum_{\kappa}(T-\kappa)\int u_1u_\kappa^\top f(X_{ij1}=x+H^{1/2}u_1,X_{ij(1+\kappa)}=x+H^{1/2}u_{(1+\kappa)})K(u_1)K(u_{(1+\kappa)})du_1du_{(1+\kappa)}\\
&+ o_p(1)\\
&= \frac{4\mu_2^d(K)}{(T-1)}\sum_{\kappa=1}^{T-1}\left(1-\frac{\kappa}{T}\right)f_{X_{ij1},X_{ij(1+\kappa)}}(x,x)I_d+o_p(1), \qquad (7.7)
\end{aligned}
$$

for $\kappa=|t-s|$, where $\kappa\in\{1,\dots,(T-1)\}$.

To conclude the proof of Lemma 1(i), it is necessary to turn to a law of large numbers. Then, $Var(\Psi_n)\to 0$ must be proved, as $I\!N\to\infty$ with fixed T. With this aim, we have

$$
\begin{aligned}
& Var(\Psi_n)\\
= \; & \frac{1}{I\!NT^2|H|^2}\sum_{\kappa}(T-\kappa)Var\left(H^{-1/2}\widetilde{X}_{ij1(1+\kappa)}\widetilde{X}_{ij1(1+\kappa)}^\top H^{-1/2}K_{ij1}K_{ij(1+\kappa)}\right)\\
+ \; & \frac{1}{I\!NT^2|H|^2}\sum_{\kappa}\sum_{\kappa<\kappa'}Cov\left(H^{-1/2}\widetilde{X}_{ij1(1+\kappa)}\widetilde{X}_{ij1(1+\kappa)}^\top H^{-1/2}K_{ij1}K_{ij(1+\kappa)},H^{-1/2}\widetilde{X}_{ij1(1+\kappa')}\widetilde{X}_{ij(1+\kappa')}^\top H^{-1/2}K_{ij1}K_{ij(1+\kappa')}\right),
\end{aligned}
$$

where, under Assumption A4, it is straightforward to show

$$
\frac{1}{I\!NT^2|H|^2}\sum_{\kappa}(T-\kappa)Var\left(H^{-1/2}\widetilde{X}_{ij1(1+\kappa)}\widetilde{X}_{ij1(1+\kappa)}^\top H^{-1/2}K_{ij1}K_{ij(1+\kappa)}\right)\le\frac{C}{I\!N|H|}+o_p\left(\frac{1}{I\!N|H|}\right),
$$

and

$$
\begin{aligned}
&\frac{1}{I\!NT^2|H|^2}\sum_{\kappa}\sum_{\kappa<\kappa'}Cov\left(H^{-1/2}\widetilde{X}_{ij1(1+\kappa)}\widetilde{X}_{ij1(1+\kappa)}^\top H^{-1/2}K_{ij1}K_{ij(1+\kappa)},H^{-1/2}\widetilde{X}_{ij1(1+\kappa')}\widetilde{X}_{ij1(1+\kappa')}^\top H^{-1/2}K_{ij1}K_{ij(1+\kappa')}\right)\\
&\le\frac{C'}{I\!N|H|}+o_p\left(\frac{1}{I\!N|H|}\right).
\end{aligned}
$$

Therefore, as $I\!N|H|\to\infty$ with fixed T, it is shown that this variance term tends to zero, so Lemma 1(i) holds.

Focusing now on the behaviour of B_n, under Assumption A1 and the law of large numbers,

B_n is asymptotically equal to

$$
\begin{aligned}
E(B_n) &= \frac{2}{T(T-1)|H|}\sum_{\kappa}(T-\kappa)E[H^{-1/2}(X_{ij1}-x)(X_{ij1}-x)^\top \mathcal{H}_m(x)(X_{ij1}-x)K_{ij1}K_{ij(1+\kappa)}] \\
&- \frac{2}{2T(T-1)|H|}\sum_{\kappa}(T-\kappa)E[H^{-1/2}(X_{ij1}-x)(X_{ij(1+\kappa)}-x)^\top \mathcal{H}_m(x)(X_{ij(1+\kappa)}-x)K_{ij1}K_{ij(1+\kappa)}] \\
&+ o_p(1) \\
&= \frac{2}{T(T-1)}\sum_{\kappa=1}^{T-1}(T-\kappa)\varpi_4(x,H)H^{1/2}D_{f_\kappa}(x) - \frac{2}{T(T-1)}\sum_{\kappa=1}^{T-1}(T-\kappa)\varpi_2(x,H)H^{1/2}D_{f_\kappa}(x) \\
&+ o_p(H^{3/2}).
\end{aligned} \tag{7.8}
$$

Using similar arguments as above, it can be shown that any component of the variance of B_n converges to zero as $H \to 0$ and $I\!N|H| \to \infty$.

Finally, in order to show that the nonparametric estimator of the gradient vector $\widehat{D}_m(x;H)$ is asymptotically unbiased, it is necessary to show that

$$R_n = o_p(H^{3/2}) \tag{7.9}$$

as $I\!N$ tends to infinity with fixed T. To this end, under Assumption A1 and using the law of large numbers,

$$
\begin{aligned}
E(R_n) &= \frac{2}{T(T-1)|H|}\sum_{\kappa=1}(T-\kappa)E\left[H^{-1/2}(X_{ij1}-x)(X_{ij1}-x)^\top \mathcal{R}_m(X_{ij1};x)(X_{ij1}-x)K_{ij1}K_{ij(1+\kappa)}\right] \\
&- \frac{2}{T(T-1)|H|}\sum_{\kappa}(T-\kappa)\,E\left[H^{-1/2}(X_{ij1}-x)(X_{ij(1+\kappa)}-x)^\top \mathcal{R}_m(X_{ij(1+\kappa)};x)\right. \\
&\times \left.(X_{ij(1+\kappa)}-x)K_{ij1}K_{ij(1+\kappa)}\right] + o_p(1).
\end{aligned}
$$

Summing and subtracting the following expression

$$\frac{2}{T(T-1)|H|}\sum_{\kappa}(T-\kappa)H^{-1/2}(X_{ij1}-x)(X_{ij(1+\kappa)}-x)^\top \mathcal{R}_m(X_{ij1};x)(X_{ij(1+\kappa)}-x)K_{ij1}K_{ij(1+\kappa)}$$

and, after rearranging terms, R_n can be split up into the following terms

$$E(R_n) = E(I\!I_1(x)) + E(I\!I_2(x)) + o_p(1),$$

where

$$
\begin{aligned}
E(\mathbb{I}_1(x)) &= \frac{2}{T(T-1)|H|}\sum_{\kappa}(T-\kappa)E[H^{-1/2}(X_{ij1}-x)(X_{ij1}-x)^{\top}\mathcal{R}_m(X_{ij1};x)(X_{ij1}-x)K_{ij1}K_{ij(1+\kappa)}] \\
&- \frac{2}{T(T-1)|H|}\sum_{\kappa}(T-\kappa)E[H^{-1/2}(X_{ij1}-x)(X_{ij(1+\kappa)}-x)^{\top}\mathcal{R}_m(X_{ij1};x)(X_{ij(1+\kappa)}-x)K_{ij1}K_{ij(1+\kappa)}], \\
E(\mathbb{I}_2(x)) &= \frac{2}{T(T-1)|H|}\sum_{\kappa}(T-\kappa))E[H^{-1/2}(X_{ij1}-x)(X_{ij(1+\kappa)}-x)^{\top}\left(\mathcal{R}_m(X_{ij1};x)-R_m(X_{ij(1+\kappa)};x)\right) \\
&\times (X_{ij(1+\kappa)}-x)K_{ij1}K_{ij(1+\kappa)}].
\end{aligned}
$$

Analyzing $E(\mathbb{I}_1(x))$ and $E(\mathbb{I}_2(x))$ separately, we follow a similar procedure as above. Under Assumption A3,

$$
\begin{aligned}
E(\mathbb{I}_1(x)) &= \frac{2}{T(T-1)}\sum_{\kappa}(T-\kappa)\int u_1u_1^{\top}H^{1/2}\mathcal{R}_m(x+H^{1/2}u_1;x)H^{1/2}u_1u_1^{\top}H^{1/2}D_{f_\kappa}(x)K(u_1)K(u_{(1+\kappa)})du_1du_{(1+\kappa)} \\
&- \frac{2}{T(T-1)}\sum_{\kappa}(T-\kappa)\int u_1u_{(1+\kappa)}^{\top}H^{1/2}\mathcal{R}_m(x+H^{1/2}u_1;x)H^{1/2}u_{(1+\kappa)}u_1^{\top}H^{1/2}D_{f_\kappa}(x)K(u_1) \\
&\times K(u_{(1+\kappa)})du_1du_{(1+\kappa)}.
\end{aligned}
$$

By definition (7.4) and Assumption A4,

$$
R_m(x+H^{1/2}u_1;x) \leq \int_0^1 \zeta(\varphi\|H^{1/2}u_1\|)(1-\varphi)d\varphi,
$$

where $\zeta(\cdot)$ is the modulus of continuity of $\partial^2 m(x)/\partial x\partial x^{\top}$. Hence, assuming that $\zeta(\varphi\|H^{1/2}u_1\|) \to 0$ as $\mathbb{N} \to \infty$,

$$
\begin{aligned}
E|\mathbb{I}_1(x)| &\leq \frac{C_1}{T(T-1)}\sum_{\kappa}|T-\kappa|\int\int_0^1 |u_1u_1^{\top}H^{1/2}||\zeta(\varphi\|H^{1/2}u_1\|)||H^{1/2}u_1u_1^{\top}H^{1/2}|D_{f_\kappa}(x)|K(u_1)| \\
&\times |K(u_{(1+\kappa)})||1-\varphi|d\varphi du_1du_{(1+\kappa)} \\
&+ \frac{C_2}{T(T-1)}\sum_{\kappa}|T-\kappa|\int\int_0^1 |u_1u_{(1+\kappa)}^{\top}H^{1/2}||\zeta(\varphi\|H^{1/2}u_1\|)||H^{1/2}u_{(1+\kappa)}u_1^{\top}H^{1/2}|D_{f_\kappa}(x) \\
&\times |K(u_1)||K(u_{(1+\kappa)})||1-\varphi|d\varphi du_1du_{(1+\kappa)}
\end{aligned}
$$

and $E(\mathbb{I}_1(x)) = o_p(H^{3/2})$ follows by dominated convergence. Similarly,

$$
\begin{aligned}
E|\mathbb{I}_2(x)| &\leq \frac{C_3}{T(T-1)}\sum_{\kappa}|T-\kappa|\int\int_0^1 |u_1u_{(1+\kappa)}^{\top}H^{1/2}||\zeta(\varphi(H^{1/2}u_1))-\zeta(\varphi(H^{1/2}u_{(1+\kappa)}))| \\
&\times |H^{1/2}u_{(1+\kappa)}u_1^{\top}H^{1/2}|D_{f_\kappa}(x)|K(u_1)||K(u_{(1+\kappa)})|d\varphi du_1du_{(1+\kappa)}.
\end{aligned}
$$

Proceeding as before, by dominated convergence, $E(\mathbb{I}_2(x)) = o_p(H^{3/2})$ and expression (7.11) holds.

$\blacksquare$

Using the results of Lemma 1 in (7.5), by the Slutsky theorem and after rearranging terms, the conditional bias of $\widehat{D}_m(x;H)$ when T is fixed is

$$E[\widehat{D}_m(x;H)|\mathbb{X}] - D_m(x) = \frac{1}{2\mu_2^d(K)\sum_{\kappa=1}^{T-1}\left(1-\frac{\kappa}{T}\right) f_{X_{ij1},X_{ij(1+\kappa)}}(x,x)} \sum_{\kappa=1}^{T-1}\left(1-\frac{\kappa}{T}\right)[\varpi_4(x,H) - \varpi_2(x,H)] D_{f_\kappa}(x) + o_p(H) + o_p\left(\frac{1}{\sqrt{I\!N|H|}}\right). \quad (7.10)$$

Similarly, when $T \to \infty$, we obtain

$$E[\widehat{D}_m(x;H)|\mathbb{X}] - D_m(x) = \frac{1}{2\mu_2^d(K)\sum_{\kappa=1}^{T-1} f_{X_{ij1},X_{ij(1+\kappa)}}(x,x)} \sum_{\kappa=1}^{T-1}[\varpi_4(x,H) - \varpi_2(x,H)] D_{f_\kappa}(x) + o_p(H) + o_p\left(\frac{1}{\sqrt{I\!N|H|}}\right). \quad (7.11)$$

Proof of Lemma 2: Proceeding as in (7.7), under Assumptions A1-A2, it can be written

$$\begin{aligned}
& I\!N(T-1)^2|H|Var(U_n) \\
&= \frac{4}{T^2|H|}\sum_{\kappa=1}^{T-1}(T-\kappa)Var(H^{-1/2}X_{ij1(1+\kappa)}\epsilon_{ij1(1+\kappa)}K_{ij1}K_{ij(1+\kappa)}) \\
&+ \frac{4}{T^2|H|}\sum_{\kappa=1}^{T-1}\sum_{\kappa'\neq\kappa}^{T-1}(T-\kappa)(T-\kappa')E[H^{-1/2}X_{ij1(1+\kappa)}X_{ij1(1+\kappa')}^\top H^{-1/2}\epsilon_{ij1(1+\kappa)}\epsilon_{ij1(1+\kappa')}K_{ij1}^2K_{ij(1+\kappa)}K_{ij(1+\kappa')}] \\
&+ o_p(1) \\
&= I\!I_3(x) + I\!I_4(x) + o_p(1), \qquad (7.12)
\end{aligned}$$

where $X_{ij1(1+\kappa)} = X_{ij1} - X_{ij(1+\kappa)}$ and $X_{ij1(1+\kappa')} = X_{ij1} - X_{ij(1+\kappa')}$. Similar definition are considered for $\epsilon_{ij1(1+\kappa)}$ and $\epsilon_{ij1(1+\kappa')}$.

Analyzing each of these terms separately and under strict stationarity, by the law of iterated expectations,

$$
\begin{aligned}
I\!I_3(x) &= \frac{4}{T^2|H|}\sum_{\kappa}(T-\kappa)E[E(\tilde{\epsilon}^2_{ij1(1+\kappa)}|X_{ij1},X_{ij(1+\kappa)})H^{-1/2}(X_{ij1}-x)(X_{ij1}-x)^\top H^{-1/2}K^2_{ij1}K^2_{ij(1+\kappa)}] \\
&- \frac{4}{T^2|H|}\sum_{\kappa}(T-\kappa)E[E(\tilde{\epsilon}^2_{ij1(1+\kappa)}|X_{ij1},X_{ij(1+\kappa)})H^{-1/2}(X_{ij1}-x)(X_{ij(1+\kappa)}-x)^\top H^{-1/2}K^2_{ij1}K^2_{ij(1+\kappa)}] \\
&= \frac{8}{T^2}\sum_{\kappa}(T-\kappa)\{\sigma^2(x,x)-\sigma_{1,(1+\kappa)}(x,x)\}f_{X_{ij1},X_{ij(1+\kappa)}}(x,x)\int u_1u_1^\top K^2(u_1)K^2(u_{(1+\kappa)})du_1du_{(1+\kappa)} \\
&+ \frac{8}{T^2}\sum_{\kappa}(T-\kappa)\{\sigma^2(x,x)-\sigma_{1,(1+\kappa)}(x,x)\}f_{X_{ij1},X_{ij(1+\kappa)}}(x,x)\int u_{(1+\kappa)}u_{(1+\kappa)}^\top K^2(u_1)K^2(u_{(1+\kappa)})du_1du_{(1+\kappa)} \\
&+ o_p(1) \\
&= \frac{16R_2^d(K)R^d(K)}{T^2}\sum_{\kappa}(T-\kappa)\{\sigma^2(x,x)-\sigma_{1,(1+\kappa)}(x,x)\}f_{X_{ij1},X_{ij(1+\kappa)}}(x,x)I_d(1+o_p(1)), \qquad (7.13)
\end{aligned}
$$

where

$$\sigma^2(x,x)=E(\epsilon^2_{ij1}|X_{ij1}=x,X_{ij(1+\kappa)}=x) \quad \text{and} \quad \sigma_{1,(1+\kappa)}(x,x)=E(\epsilon_{ij1}\epsilon_{ij(1+\kappa)}|X_{ij1}=x,X_{ij(1+\kappa)}=x).$$

Considering now the behavior of $I\!I_4(x)$, we analyze two different cases: (i) T fixed and (ii) $T\to\infty$.

Case (i): For any $t\geq 1$, by Assumptions A1-A2 and the law of iterated expectations,

$$
\begin{aligned}
I\!I_4(x) &= \frac{4}{T^2|H|}\sum_{\kappa}\sum_{\kappa<\kappa'}(T-\kappa)(T-\kappa')E[E(\epsilon_{ij1(1+\kappa)}\epsilon_{ij1(1+\kappa')}|X_{ij1},X_{ij(1+\kappa)},X_{ij(1+\kappa')})H^{-1/2}(X_{ij}-X_{ij(1+\kappa)}) \\
&\times (X_{ij1}-X_{ij(1+\kappa')})^\top H^{-1/2}K^2_{ij1}K_{ij(1+\kappa)}K_{ij(1+\kappa')}] \\
&= \frac{4R_2^d(K)|H|^{1/2}}{T^2}\sum_{\kappa}\sum_{\kappa<\kappa'}(T-\kappa)(T-\kappa')\sigma(x,x,x)f_{X_{ij1},X_{ij(1+\kappa)},X_{ij(1+\kappa')}}(x,x,x)I_d+o_p(|H|^{1/2}), \qquad (7.14)
\end{aligned}
$$

where $\sigma(x,x,x)=E(\epsilon_{ij1(1+\kappa)}\epsilon_{ij1(1+\kappa')}|X_{ij1}=x,X_{ij(1+\kappa)}=x,X_{ij(1+\kappa')}=x)$.

Replacing (7.13)-(7.14) in (7.12) and after rearranging terms,

$$
\begin{aligned}
Var(U_n) &= \frac{16R_2^d(K)R^d(K)}{I\!NT(T-1)^2|H|}\sum_{\kappa=1}^{T-1}\left(1-\frac{\kappa}{T}\right)\{\sigma^2(x,x)-\sigma_{1,(1+\kappa)}(x,x)\}f_{X_{ij1},X_{ij(1+\kappa)}}(x,x)I_d \\
&\times (1+o_p(1)). \qquad (7.15)
\end{aligned}
$$

Case (ii): Following a similar proof scheme as in Cai and Li (2008), we split $I\!I_4(x)$ into two terms obtaining

$$
\begin{aligned}
I\!I_4(x) &= \frac{1}{T^2|H|}\sum_{\kappa=1}^{a_T}\sum_{\kappa<\kappa'}^{a_T}(T-\kappa)(T-\kappa')E[H^{-1/2}X_{ij1(1+\kappa)}X^\top_{ij1(1+\kappa')}H^{-1/2}\epsilon_{ij1(1+\kappa)}\epsilon_{ij1(1+\kappa')}K^2_{ij1}K_{ij(1+\kappa)}K_{ij(1+\kappa')}] \\
&+ \frac{1}{T^2|H|}\sum_{\kappa>a_T}^{T-1}\sum_{\kappa<\kappa'}^{T-1}(T-\kappa)(T-\kappa')E[H^{-1/2}X_{ij1(1+\kappa)}X^\top_{ij1(1+\kappa')}H^{-1/2}\epsilon_{ij1(1+\kappa)}\epsilon_{ij1(1+\kappa')}K^2_{ij1}K_{ij(1+\kappa)}K_{ij(1+\kappa')}] \\
&= I\!I_4^{(1)}(x)+I\!I_4^{(2)}(x), \qquad (7.16)
\end{aligned}
$$

where a_T is a sequence of positive integers such that $a_T h^d \to 0$. First, we plan to show that $I\!I_4^{(1)}(x) \to 0$. Using the results in (7.14) it is straightforward to show

$$\begin{aligned} |I\!I_4^{(1)}(x)| &\leq \frac{1}{T^2|H|}\sum_{\kappa=1}^{a_T}\sum_{\kappa<\kappa'}^{a_T}(T-\kappa)(T-\kappa')\left|E[\epsilon_{ij1(1+\kappa)}\epsilon_{ij1(1+\kappa')}H^{-1/2}X_{ij1(1+\kappa)}X_{ij1(1+\kappa')}^{\top}H^{-1/2}K_{ij1}^2K_{ij(1+\kappa)}K_{ij(2+\kappa')}]\right| \\ &\leq \frac{C|H|^{1/2}}{T^2}\sum_{\kappa=1}^{a_T}\sum_{\kappa'\neq\kappa}^{a_T}(T-\kappa)(T-\kappa') \leq Ca_T|H|^{1/2} = o_p(1), \end{aligned} \tag{7.17}$$

given that

$$\left||H|^{-1}E[\epsilon_{ij1(1+\kappa)}\epsilon_{ij1(1+\kappa')}H^{-1/2}X_{ij1(1+\kappa)}X_{ij1(1+\kappa')}^{\top}H^{-1/2}K_{ij1}^2K_{ij(1+\kappa)}K_{ij(2+\kappa')}]\right| = O_p(|H|^{1/2}).$$

Focusing now on the behavior of the leading term of $I_4^{(2)}(x)$ and using the Davydov's inequality, for $\widetilde{\kappa} = \kappa' - \kappa$, when $\kappa' \neq \kappa$,

$$\begin{aligned} &\left|(|H|^2H)^{-1}Cov[X_{ij1(1+\kappa)}\epsilon_{ij1(1+\kappa)}K_{ij1}K_{ij(1+\kappa)}, X_{ij1(1+\kappa')}\epsilon_{ij1(1+\kappa')}K_{ij1}K_{ij(1+\kappa')}]\right| \\ &\leq [\alpha(\widetilde{\kappa})]^{\delta/(2+\delta)}\|(|H|H^{1/2})^{-1}X_{ij1(1+\kappa)}\epsilon_{ij1(1+\kappa)}K_{ij1}K_{ij(1+\kappa)}\|_{(2+\delta)} \\ &\times\|(|H|H^{1/2})^{-1}X_{ij1(1+\kappa)}\epsilon_{ij1(1+\kappa1)}K_{ij1}K_{ij(1+\kappa1)}\|_{(2+\delta)}, \end{aligned} \tag{7.18}$$

Conditioning on $(X_{ij1}, X_{ij(1+\kappa)})$, using Assumption A2 and the c_T inequality,

$$\begin{aligned} &E|(|H|H^{1/2})^{-1}\epsilon_{ij1(1+\kappa)}X_{ij1(1+\kappa)}K_{ij1}K_{ij(1+\kappa)}|^{(2+\delta)} \\ &\leq CE|(|H|H^{1/2})^{-1}\epsilon_{ij1}X_{ij1(1+\kappa)}K_{ij1}K_{ij(1+\kappa)}|^{(2+\delta)} + CE|(|H|H^{1/2})^{-1}\epsilon_{ij(1+\kappa)}X_{ij1(1+\kappa)}K_{ij1}K_{ij(1+\kappa)}|^{(2+\delta)}. \end{aligned}$$

Analyzing each term separately and using the law of iterated expectations,

$$\begin{aligned} &E|(|H|H^{1/2})^{-1}\epsilon_{ij1}X_{ij(1+\kappa)}K_{ij1}K_{ij(1+\kappa)}|^{(2+\delta)} \\ &\leq |H|^{-(1+\delta)}E(|\epsilon_{ij1}|^{(2+\delta)}|X_{ij1} = x, X_{ij(1+\kappa)} = x)f_{X_{ij1},X_{ij(1+\kappa)}}(x,x)\int\{|u_1|^{(2+\delta)} + |u_{(1+\kappa)}|^{(2+\delta)}\} \\ &\times K^{(2+\delta)}(u_1)K^{(2+\delta)}(u_{(1+\kappa)})du_1du_{(1+\kappa)} \\ &\leq C|H|^{-(1+\delta)} = O_p(|H|^{-(1+\delta)}). \end{aligned}$$

Similarly, we see that $E|(|H|H^{1/2})^{-1}\epsilon_{ij(1+\kappa)}X_{ij1(1+\kappa)}K_{ij1}K_{ij(1+\kappa)}|^{(2+\delta)} = O_p(|H|^{-(1+\delta)})$. Using these results in (7.18), we can write

$$\begin{aligned} &\left|(|H|^2H)^{-1}Cov[X_{ij1(1+\kappa)}\epsilon_{ij1(1+\kappa)}K_{ij1}K_{ij(1+\kappa)}, X_{ij1(1+\kappa')}\epsilon_{ij1(1+\kappa')}K_{ij1}K_{ij(1+\kappa')}]\right| \\ &\leq C[\alpha(\widetilde{\kappa})]^{\delta/(2+\delta)}|H|^{-2(1+\delta)/(2+\delta)} = O_p\left(\alpha(\widetilde{\kappa})^{\delta/(2+\delta)}|H|^{-2(1+\delta)/(2+\delta)}\right). \end{aligned}$$

Therefore, by Assumption B2 and choosing a_T such that $|H|^{1/2}a_T^2 = O_p(1)$, so that the requirement $a_T|H|^{1/2} \to 0$ holds, the (ℓ, ℓ')th element of $\mathbb{I}_4^{(2)}(x)$ becomes

$$|\mathbb{I}_{4(\ell,\ell')}^{(2)}(x)| \leq C|H|^{-\delta/(2+\delta)} \sum_{\kappa > a_T} \sum_{\kappa < \kappa'} [\alpha(\widetilde{\kappa})]^{\delta/(2+\delta)} = O_p(|H|^{-\delta/(2+\delta)} a_T^{-\delta}) \to 0. \tag{7.19}$$

Using (7.17) and (7.19) it is shown that $\mathbb{I}_4(x) \to 0$ as $T \to \infty$, so Lemma 2(ii) is proved.

■

Considering now the conditional covariance matrix in (7.1), we use the results in Lemmas 1 and 2. Then, by the Slutsky theorem, as $\mathbb{N} \to \infty$, for fixed T,

$$Var[\widehat{D}_m(x;H)|\mathbb{X}] = H^{-1} \frac{R_2^d(K) R^d(K) \sum_{\kappa}^{T-1} \left(1 - \frac{\kappa}{T}\right) \{\sigma^2(x,x) - \sigma_{1,(1+\kappa)}(x,x)\} f_{X_{ij1}, X_{ij(1+\kappa)}}(x,x)}{\mathbb{N} T |H| \mu_2^{2d}(K) \left(\sum_{\kappa=1}^{T-1} \left(1 - \frac{\kappa}{T}\right) f_{X_{ij1}, X_{ij(1+\kappa)}}(x,x)\right)^2} \times (1 + o_p(1)). \tag{7.20}$$

Similarly, as $\mathbb{N} \to \infty$ and $T \to \infty$, the conditional covariance matrix of the pairwise least-squares nonparametric estimator is of the form

$$Var[\widehat{D}_m(x;H)|\mathbb{X}] = H^{-1} \frac{R_2^d(K) R^d(K) \sum_{\kappa}^{T-1} \{\sigma^2(x,x) - \sigma_{1,(1+\kappa)}(x,x)\} f_{X_{ij1}, X_{ij(1+\kappa)}}(x,x)}{\mathbb{N} T |H| \mu_2^{2d}(K) \left(\sum_{\kappa=1}^{T-1} f_{X_{ij1}, X_{ij(1+\kappa)}}(x,x)\right)^2} (1 + o_p(1)).$$

Finally, replacing (7.10)-(7.11) and (7.20)-(7.21) in (7.1), the proof of this theorem is done.

■

Proof of Theorem 2. With the aim of obtaining the asymptotic distribution of the nonparametric estimator for $D_m(x)$, the Lyapunov condition has to be checked. For any unit vector $\imath_d \in \mathbb{R}^d$, we define $\phi_{ij}^* = \binom{T}{2}^{-1} \sum_{t=1}^{T-1} \sum_{s=1+t}^{T} \phi_{ijts}$, where $\phi_{ijts} = |H|^{-1/2} \imath_d^\top X_{h,ijts} \epsilon_{ijts} K_{ijt} K_{ijts}$. As T is fixed, by Assumption A1 we get that ϕ_{ij}^* are independent random variables. Then, by the law of iterated expectations, the expression to analyze is

$$\sqrt{\mathbb{N}|H|H^3} \imath_d^\top U_n = \frac{1}{\sqrt{\mathbb{N}}} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \phi_{ij}^* + o_p(\sqrt{\mathbb{N}|H|H^3}). \tag{7.21}$$

Under Assumptions A1-A2, it is easy to show that $E(\phi_{ij}^*) = 0$, whereas the variance term is equal to

$$\begin{aligned} Var(\phi_{ij}^*) &= \binom{T}{2}^{-2} \sum_{\kappa}^{T-1} (T-\kappa) Var(\phi_{ij1(1+\kappa)}) + \binom{T}{2}^{-2} \sum_{\kappa}^{T-2} \sum_{\kappa<\kappa'}^{T-1} (T-\kappa)(T-\kappa') Cov(\phi_{ij1(1+\kappa)}, \phi_{ij1(1+\kappa')}) \\ &= \frac{16 d R_2^d(K) R^d(K)}{T^2 (T-1)^2} \sum_{\kappa=1}^{T-1} (T-\kappa) \{\sigma^2(x,x) - \sigma_{1,(1+\kappa)}(x,x)\} f_{X_{ij1}, X_{ij(1+\kappa)}}(x,x) (1 + o_p(1)), \end{aligned} \tag{7.22}$$

given that, previously it was shown,

$$\begin{aligned}
Var(\phi_{ijts}) &= \frac{16dR_2^d(K)R^d(K)}{T^2(T-1)^2}\sum_{\kappa=1}^{T-1}(T-\kappa)\{\sigma^2(x,x)-\sigma_{1,(1+\kappa)}(x,x)\}f_{X_{ij1},X_{ij(1+\kappa)}}(x,x) \\
&\times (1+o_p(1)), \\
Cov(\phi_{ij1(1+\kappa)},\phi_{ij1(1+\kappa')}) &= \frac{4dR_2^d(K)|H|^{1/2}}{T^2(T-1)^2}\sum_{\kappa}\sum_{\kappa<\kappa}(T-\kappa)(T-\kappa')\sigma(x,x,x)f_{X_{ij1},X_{ij(1+\kappa)},X_{ij(1+\kappa')}}(x,x,x) \\
&+ o_p(|H|^{1/2}).
\end{aligned}$$

By Minkowski's inequality and Assumption A7, using a similar procedure as in (7.19), it can be shown $E|\phi^*_{ij}|^{(2+\delta)} \le C|H|^{-\delta/2}$, for some $\delta > 0$. Then, by Assumption A6, it is proved that

$$\frac{1}{(I\!N)^{(2+\delta)}}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}E|\phi^*_{ij}|^{(2+\delta)} \le C(I\!N|H|)^{-\delta/2} \to 0, \tag{7.23}$$

so the Lyapounov's condition holds and the proof of the theorem is closed.

∎

Proof of Theorem 3. As $T \to \infty$, ϕ_{ijts} is a stationary α-mixing process. In this framework, the common approach to prove asymptotic normality is to employ Doob's small-block and large-block technique. See Ibragimmov and Linnik (1971), Cai and Li (2008), or Cai et al. (2015) among others, for further details. For the sake of comparison, we follow a similar notation as in Cai and Li (2008). We partition the time observations $\{1,\ldots,T\}$ into $(2q_T+1)$ subsets with large block of size r_T and small block of size $s_T < T$, with $r_T + s_T < T$. Set $q_T = \lfloor T/(r_T+s_T)\rfloor$, where $\lfloor x\rfloor$ denotes the integer part of x, the expression to analyze is

$$\begin{aligned}
\sqrt{I\!NT|H|H^{3/2}}(T-1)\imath_d^\top U_n &= \frac{2}{\sqrt{I\!NT}}\left(\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\sum_{r=0}^{q_T-1}\eta_{ijr}+\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\sum_{r=0}^{q_T-1}\xi_{ijr}+\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}\zeta_{ijq_T}\right) \\
&\equiv \frac{1}{\sqrt{I\!NT}}\left(Q_{n,1}+Q_{n,2}+Q_{n,3}\right),
\end{aligned} \tag{7.24}$$

where

$$\eta_{ijr} = \sum_{t=l(r_T+s_T)+1}^{r(r_T+s_T)+r_T}\phi_{ijt}, \quad \xi_{ijr} = \sum_{r(r_T+s_T)+r_T+1}^{(r+1)(r_T+s_T)}\phi_{ijt}, \quad \text{and} \quad \zeta_{ijr} = \sum_{q_T(r_T+s_T)+1}^{T}\phi_{ijt}.$$

In order to prove the asymptotic normality of $\sqrt{I\!NT|H|^{T/2}}\imath_d^\top U_n$, we follow a similar proof scheme as in Cai and Li (2008). Choosing $r_T = \lfloor T^{1/\tau}\rfloor$ and $s_T = \lfloor T^{1/(\tau+1)}\rfloor$, where $\tau = (2+$

$\delta)(1+\delta)/\delta$. By Assumption B3, it can be shown

$$\frac{s_T}{r_T} = \frac{\lfloor T^{1/(\tau+1)} \rfloor}{\lfloor T^{1/\tau} \rfloor} \to 0, \quad \frac{r_T}{s_T} = \frac{1}{T}\lfloor T^{1/\tau} \rfloor \to 0, \quad \text{and} \quad q_T \alpha(s_T) \le C T^{-1/(\tau+1)\tau} \to 0. \tag{7.25}$$

Using the results of Lemma 2 and (7.25), it is straightforward to show

$$n^{-1} E[Q_{n,2}]^2 \xrightarrow{p} 0 \quad \text{and} \quad n^{-1} E[Q_{n,3}]^2 \xrightarrow{p} 0 \tag{7.26}$$

which imply that the sums over small and residuals blocks, $Q_{n,2}/\sqrt{n}$ and $Q_{n,3}/\sqrt{n}$, are asymptotically negligible in probability.

In order to show that $\eta_{i_1 j_1 r}$ in $Q_{n,1}/\sqrt{n}$ is asymptotically independent, we resort to Lemma 1.1 in Volkonskii and Rozanov (1995) obtaining

$$\left| E\left[exp\left((ijt \sum_{r=0}^{q_T-1} \eta_{i_1 j_1 r} \right) - \Pi_{r=0}^{q-1} E\left[exp\left(ijt \eta_{i_1 j_1 r} \right) \right] \right] \right| \le 16 q_T \alpha(s_T), \tag{7.27}$$

which goes to zero as $T \to \infty$ using the conditions in (7.25). Therefore, standard Lindeberg-Feller conditions for the asymptotic normality of the independent setup of $Q_{n,1}/\sqrt{n}$ have to be checked. In particular, by our stationary condition, Lemma 2, and (7.25), it is straightforward to show

$$\frac{1}{n} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \sum_{r=0}^{q_T-1} E(\eta_{ijr}^2) = \frac{q_T r_T}{T} \frac{1}{r_T} Var\left(\sum_{t=1}^{r_T} \phi_{ijt} \right) \to V^{(2)}(x). \tag{7.28}$$

Further, using Theorem 4.1 in Shao and Yu (1996) and Assumption B3,

$$E\left[\eta_{ij1}^2 I\{|\eta_{ij1}| \ge \epsilon V^{(2)}(x)\sqrt{I\!N}\} \right] \le C(I\!N)^{-\delta/2} r_T^{1+\delta/2} \|\phi_{ij1}\|_{2+2\delta}^{2+\delta}.$$

Following a similar proof scheme as in (7.23), it is easy to show $E(|\phi_{ij1}|^{(2+2\delta)}) \le C h^{-Td\delta}$. This latter result together with (7.28) implies

$$E[\eta_{ij1}^2 I\{|\eta_{ij1}| \ge \epsilon V^{(2)}(x)\sqrt{I\!NT}\}] \le C\left(I\!N T^{(\tau+1)/\tau} |H|^{T(2+\delta)/2(1+\delta)} \right)^{-\delta/2}$$

and by definition of r_T,

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \sum_{r=0}^{q_T-1} E[\eta_{ij1}^2 I\{|\eta_{ij1}| \ge \epsilon V^{(2)}(x)\sqrt{n}\}] \\ &= O_p\left((I\!N T^{(\tau-1)/\tau} |H^{T(2+\delta)/2(1+\delta)}|)^{-\delta/2} \right) \end{aligned} \tag{7.29}$$

Finally, from the proof of Theorem 18.4.1 in Ibragimmov and Linnik (1971), we get that a combination of (7.27)-(7.29) implies that $Q_{n,1}/\sqrt{I\!NT} \to N(0, V^{(2)}(x))$. Therefore, using this result together with (7.26), the asymptotic normality of $\sqrt{INT|H|^{T/2}}\imath_d^\top U_n$ is proved by applying the Slutsky's theorem.

∎

References

Ai, C. and Q. Li (2008). Semi-parametric and non-parametric methods in panel data models. In L. Matyas and P. Sevestre (Eds.), *The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory and Practice*, pp. 451–478. Springer.

Anderson, J. E. and E. van Wincoop (2003). Gravity with gravitas: A solution to the border puzzle. *American Economic Review 93*, 170–192.

Baier, S. and J. Bergstrand (2007). Do free trade agreements actually increase members international trade? *Journal of International Economics 71*, 72–95.

Balazsi, L., L. Matyas, and T. Wansbeek (2015). The estimation of multi-dimensional fixed effects panel data models. *Econometric Reviews 37*, 212–227.

Balazsi, L., L. Matyas, and T. Wansbeek (2017). Fixed effects models. In L. Matyas (Ed.), *The Econometrics of Multi-dimensional Panels: Theory and Application*, pp. 1–34. Springer.

Baltagi, B. (2015). *The Oxford Handbook of Panel Data*. Oxford University Press.

Baltagi, B., P. Egger, and M. Pfaffermayr (2003). A generalized design for bilateral trade flow models. *Economics Letters 80*, 391–397.

Bishop, K. C. and C. Timmins (2018). Using panel data to easily estimate hedonic demand functions. *Journal of the Association of Environmental and Resource Economists 5*, 517–543.

Cai, Z. (2003). Nonparametric estimation equations for time series data. *Statistics and Probability Letters 62*, 379–390.

Cai, Z., Y. Fang, M. Ling, and J. Su (2019). Inferences for a partially varying coefficient model with endogenous regressors. *Journal of Business & Economic Statistics 37*, 158–170.

Cai, Z. and Q. Li (2008). Nonparametric estimation of varying coefficient dynamic panel data models. *Econometric Theory 24*, 1321–1342.

Cai, Z. and E. Masry (2000). Nonparametric estimation of additive nonlinear ARX time series: Local linear fitting and projections. *Econometric Theory 16*, 465–501.

Cai, Z., Y. Ren, and L. Sun (2015). Pricing kernel estimation: a local estimating equation approach. *Econometric Theory 31*, 560–580.

Carrasco, M. and X. Chen (2002). Mixing and moment properties of various garch and stochastic volatility models. *Econometric Theory 18*, 17–39.

Chen, I. and H. Wall (2005). Controlling for heterogeneity in gravity models of trade and integration. *Federal Reserve Bank of St. Louir Review 87*, 49–63.

Chiappori, P. A., A. Gandhi, B. Salanié, and F. Salanié (2009). Identifying preferences under risk from discrete choices. *American Economic Review 99*, 356–362.

Davis, P. (2002). Estimating multi-way error components models with unbalanced data structures. *Journal of Econometrics 106*, 67–95.

Egger, P. and M. Pfaffermayr (2003). The proper panel econometric specification of the gravity equation: a three-way model with bilateral interaction effects. *Empirical Economics 28*, 571–580.

Eriksen, M. D. and A. Ross (2015). Housing vouchers and the price of rental housing. *American Economic Journal: Economic Policy 7*, 154–176.

Fan, J. and I. Gijbels (1995a). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *Journal of the Royal Statistical Society, Series B 57*, 371–394.

Fan, J. and I. Gijbels (1995b). *Local Polynomial Modelling and its Applications.* Chapman & Hall.

Fan, J. and T. Huang (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli 11*, 1031–1057.

Freyberger, J. (2018). Non-parametric panel data models with interactive fixed effects. *Review of Economic Studies 85*, 1824–1851.

Gao, Y. and L. Kunpeng (2013). Nonparametric estimation of fixed effects panel data models. *Journal of Nonparametric Statistics 25*, 679–693.

Gerardi, K. and A. Shapiro (2009). Does competition reduce price dispersion? New evidence from the airline industry. *Journal of Political Economy 117*, 1–37.

Hahn, J. and W. Newey (2004). Jacknife and analytical bias reduction for nonlinear panel models. *Journal of Econometrics 72*, 1295–1319.

Hall, P., Q. Li, and J. Racine (2007). Nonparametric estimation of regression functions in the presence of irrelevant regressors. *Review of Economic and Statistics 89*, 784–789.

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

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

Henderson, D. J., Q. Li, C. Parmeter, and S. Yao (2015). Gradient-based smoothing parameter selection for nonparametric regression estimation. *Journal of Econometrics 184*, 233–241.

Henderson, D. J. and C. F. Parmeter (2015). *Applied Nonparametric Econometrics*. New York: Cambridge University Press.

Honoré, B. E. and J. L. Powell (2005). Pairwise difference estimation of nonlinear models. In D. W. K. Andrews and J. H. Stock (Eds.), *Identification and Inference for Econometric Models. Essays in Honor of Thomas Rothenberg*, pp. 520–553. Cambridge University Press, Cambridge.

Hsiao, C. (2014). *Analysis of Panel Data* (3rd ed.). Cambridge University Press.

Ibragimmov, I. and Y. Linnik (1971). *Indepdendent and Stationary Sequences of Random Variables*. Walters-Noordhoff.

Lancaster, T. (2000). The incidental parameter problem since 1948. *Journal of Econometrics 95*, 391–413.

Lee, Y., D. Mukherjee, and A. Ullah (2019). Nonparametric estimation of the marginal effect in fixed-effect panel data models. *Journal of Multivariate Analysis 171*, 53–67.

Lee, Y. and P. Phillips (2015). Model selection in the presece of incidental parameters. *Journal of Econometrics 188*, 474–489.

Li, Q., Y. Sun, and Y. Y. Zhang (2015). Nonparameric panel data regression models. In B. H. Baltagi (Ed.), *The Oxford Handbook of Panel Data*, pp. 285–324. Oxford University Press.

Mátyás, L. (Ed.) (2017). *The Econometrics of Multi-dimensional Panels*. Springer.

Mátyás, L., K. László, and M. Harris (1997). Proper econometic specification of the gravity model. *The World Economy 20*, 363–368.

Müller, H.-G., U. Stadmüller, and T. Schmitt (1987). Bandwidth choice and confidence intervals for derivatives of noisy data. *Biometrika 74*, 743–749.

Mundra, K. (2005). Nonparametric slope estimation for fixed-effects panel data models. Working Paper, San Diego State University.

Nadaraya, E. (1964). On estimating regression. *Theory of Probability and its Applications 9*, 141–142.

Olsen, E. (2003). Housing programs for low-income households. In R. Moffitt (Ed.), *Means-tested transfer programs in the United States*, pp. 365–441. Chicago University Press.

Parmeter, C. F. and J. S. Racine (2018). Nonparametric estimation and inference for panel data models. Working Paper, University of Miami.

Qian, J. and L. Wang (2012). Estimating semiparametric panel data models by marginal integration. *Journal of Econometrics 167*, 483–493.

Racine, J. and Q. Li (2004). Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics 119*, 99–130.

Robinson, P. (1988). Root-n consistent semiparametric regression. *Econometrica 56*, 931–954.

Rodriguez-Poo, J. and A. Soberon (2015). Nonparametric estimation of fixed effects panel data varying coefficient models. *Journal of Multivariate Analysis 133*, 95–122.

Rodriguez-Poo, J. and A. Soberon (2017). Nonparametric and semiparametric panel data models: recent developments. *Journal of Economic Surveys 31*, 923–960.

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

Shao, Q. and H. Yu (1996). Weak convergence of weighted empirical processes of dependent sequences. *Annals of Probability 24*, 2098–2127.

Soberon, A., W. Stute, and J. Rodriguez-Poo (2020). Testing for distributional features in varying coefficient panel data models. *Econometric Reviews 39*(3), 277–298.

Stromberg, A. J., O. Hossjer, and D. M. Hawkins (2000). The least trimmed differences regression estimator and alternatives. *Journal of the American Statistical Association 95*(451), 853–864.

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

Su, L. and A. Ullah (2011). Nonparametric and semiparametric panel econometric models: Estimation and testing. In A. Ullah and D. Giles (Eds.), *Handbook of Empirical Economics and Finance*, pp. 455–497. New York: Taylor and Francis Group.

Susin, S. (2002). Rent vouchers and the price of low-income housing. *Journal of Public Economics 83*, 109–152.

Volkonskii, V. and Y. Rozanov (1995). Some limit theorems for random functions, part i. *Theory of Probability and its Applications 4*, 178–197.

Watson, N. (1964). Smooth regression analysis. *Sankhya, Series A 26*, 359–372.

Ye, X. and X. Wu (2014). Three-dimensional panel data models with interactive effects: estimation and simulation. *Economics Letters 123*, 62–65.