Estimation of spatial autoregressive panel data models with nonparametric endogenous effect

Zixin Yang
Peking University

Xiaojun Song
Peking University

Jihai Yu
Peking University

November 11, 2024

Abstract

This paper proposes a sieve generalized method of moments (GMM) method for the estimation of spatial autoregressive panel data models with nonparametric endogenous effect. The new estimator incorporates both linear moments based on the orthogonality of the exogenous regressors with the model disturbances and quadratic moments based on the properties of idiosyncratic errors. We establish the consistency and asymptotic normality of the sieve GMM estimator and show that it is more efficient than the sieve instrumental variable estimator due to additional quadratic moments. We also put forward two new test statistics for testing the linearity of the endogenous effect. Both test statistics are shown to be asymptotic normal under the null and a sequence of local alternatives after proper standardization. Monte Carlo simulations show that the proposed estimators and tests perform well in finite samples. We also apply our method to estimate the environmental Kuznets curve in China and the knowledge spillover effect among 61 countries.

JEL Classification: C14, C31, C36

Keywords: *Generalized method of moments, Spatial autoregressive models, Sieve estimation, Social interaction*

1 Introduction

The spatial autoregressive (SAR) model introduced by Cliff and Ord (1973) has received considerable attention in various fields of economics, as it provides a convenient framework to model the interaction of economic agents. Panel data model with spatial interaction is also of great interest, as it allows researchers to control for both unobserved heterogeneity and spatial correlation; see, e.g., Elhorst (2003), Lee and Yu (2010, 2012), Baltagi et al. (2013) and Kuersteiner and Prucha (2020). Among static spatial panel data models, the most commonly used one is

$$y_{it} = \lambda_0 \sum_{j=1}^{n} w_{ij} y_{jt} + x_{it}'\beta_0 + c_{i0} + \epsilon_{it}, \quad i = 1, \ldots, n,\ t = 1, \ldots, T, \tag{1}$$

where y_{it} is the outcome variable of individual i at time t, x_{it} is the vector of explanatory variables, c_{i0} is the unobserved individual effect, and ϵ_{it} is the error component. The term $\sum_{j=1}^{n} w_{ij} y_{jt}$ captures the endogenous effect, that is, the effect of the other's outcome on one's own outcome, where w_{ij} is a known spatial weight between individuals i and j. Applications of the spatial panel data model (1) include commodity tax competition among US states (Egger et al., 2005), technological spillover through spatial externalities (Ertur and Koch, 2007), and cross-sectional dependence in trade flows (LeSage and Fischer, 2020), to name a few.

Model (1) assumes that the spatial interaction effect is linearly proportional to others' outcome. This implies that if an individual i interacts with j, the impact of a one-point increase in j's outcome on i's outcome is always equal to $\lambda_0 w_{ij}$. Such an assumption simplifies the analysis, but can be violated in many empirical applications. For example, in the context of peer effects in education, the magnitude of peer achievement spillovers experienced by student i could vary by student i's achievement as well as his peers' level of achievement, see, e.g., Hoxby and Weingarth (2005) and Fruehwirth (2013). When choosing study effort, an individual does not place equal weights to all his peers but care mostly about the agents who make a relative high effort (Boucher et al., 2024). Another context where nonlinear interaction effect emerges is criminal activity. As remarked by Belhaj et al. (2014), an individual may be susceptible to criminal influence once his friends are really committed to criminal activity, but such susceptibility may then vanish as friends get close to full-time crime. In this case, the marginal impact of peers' actions on own marginal utility diminishes as the level of peers' actions increases, which leads to a nonlinear response function. If such a nonlinearity is ignored, the estimated endogenous effects may be biased and the resulting policy implications can be misleading. In the literature, there has been a growing interest in estimation and inference of SAR models with nonlinear endogenuous effect. Examples

include functional coefficient model by Sun and Malikov (2018), trending time-varying coefficient model by Chang et al. (2024), among others. However, these models are all designed to capture endogenous effects varying with exogeneous variables or purely with time. When the marginal impact of peer's actions is related to the endogenous outcome, the related methods are not directly applicable.

To account for the potential nonlinearity of the spatial interaction effect, we extend model (1) to the following nonlinear SAR panel data model:

$$y_{it} = \sum_{j=1}^{n} w_{ij} h_0(y_{jt}) + x_{it}'\beta_0 + c_{i0} + \epsilon_{it}, \quad i = 1, \ldots, n,\ t = 1, \ldots, T, \tag{2}$$

where h_0 is an unknown nonparametric function.[1] The linear SAR model (1) is just a special case of this model with $h_0(y_{jt}) = \lambda_0 y_{jt}$. In the context of network interaction, the nonlinear model (2) can be regarded as the best response function of a game where agents have quadratic nonlinear payoffs.[2] Here, h_0 is a nonparametric function of the endogenous variable y_{jt}. The endogeneity of y_{jt} can be controlled using x_{jt} as instrumental variables. Then, model (2) can be viewed as a type of nonparametric instrumental variable (NPIV) regression model, the estimation of which has been investigated in Ai and Chen (2003), Newey and Powell (2003) and Blundell et al. (2007), among others, under independent data. Recently, Hoshino (2022) extends the estimation methods in the above works to the SAR model under cross-sectional data setting and develops a sieve two-stage least squares (2SLS) estimator for β_0 and h_0. However, Hoshino (2022) requires that at least one element of β_0 be away from zero. If this does not hold, the sieve 2SLS estimator may be inconsistent, and the identification of h_0 fails.

In this article, we propose a sieve generalized method of moments (GMM) estimator for the nonlinear SAR model (2). The developed estimator incorporates not only linear moments based on the orthogonality condition between the IVs and the model disturbances but also quadratic moments based on the properties of the idiosyncratic errors. Our contribution to the literature is threefold. First, we extend the nonlinear SAR model in Hoshino (2022) to the panel data setting, where both nonlinear endogenous effect and unobserved individual heterogeneity are considered. Second, we utilize a diverging number of quadratic orthogo-

[1] To account for nonlinear interaction effect, one may also consider the specification $h_0(\sum_{j=1}^{n} w_{ij} y_{jt})$, see, e.g., Belhaj et al. (2014). We leave this for further study.

[2] Consider a population of n agents who need to choose an action to maximize their utilities at some time t. The action of agent i is denoted by a continuous variable y_{it}. Utility function of agent i is assumed to be $V_i(y_{it}, Y_{-i,t}) = y_{it}(\mu_i(Y_{-i,t}) + x_{it}'\beta_0 + c_{i0} + \epsilon_{it}) - \frac{1}{2}y_{it}^2$, where $Y_{-i,t} = (y_{1t}, \ldots, y_{i-1,t}, y_{i+1,t}, \ldots, y_{nt})$ and $\mu_i(\cdot)$ is an aggregate function (see Boucher and Fortin (2016) and Boucher et al. (2024) for explanation of the components in V_i). Then, if $\mu_i(Y_{-i,t}) = \sum_{j=1}^{n} w_{ij} h_0(y_{jt})$, the first order condition directly implies model (2). If $\mu_i(Y_{-i,t}) = h_0(\sum_{j=1}^{n} w_{ij} y_{jt})$, the first order condition implies the model in footnote 1.

nality conditions in identifying and estimating the nonlinear SAR model (2). In absence of valid instruments, quadratic moment conditions are shown to have non-trivial identification power, and can improve the identification of our model. In terms of estimation, the sieve GMM estimator with both linear and quadratic moment conditions would have efficiency gains compared to the sieve 2SLS estimator in Hoshino (2022). Finally, we propose two new test statistics, Lagrangian multiplier (LM) and distance metric (DM) statistics to test the linearity of endogenous effect in our model. Compared with the Wald-type statistic proposed by Hoshino (2022), the LM test utilizes only the restricted estimator under the null model and its performance is less sensitive to the degree of ill-posedness. Both LM and DM test statistics are in a quadratic form of our moment function, which is a diverging dimensional vector of linear-quadratic forms. We provide a new central limit theorem for this type of quantity, which could be of interest in broader applications. The numerical simulation shows that the proposed estimator has satisfactory finite sample performance and the new tests are robust to the choice of smoothing parameters in nonparametric estimation.

The rest of the paper is organized as follows. Section 2 develops the sieve GMM estimation procedure and discusses the identification of our model via both linear and quadratic moment conditions. Section 3 establishes the convergence rate and asymptotic normality of the proposed GMM estimator. In Section 4, we propose two new test statistics for testing the linearity of endogenous effect. The asymptotic distributions of the test statistics are established under both the null hypothesis and a sequence of local alternatives. We evaluate the finite sample performance of the proposed estimator and test statistics in Section 5. In Section 6, we apply our method to investigate the environmental Kuznets curve (EKC) in China and the knowledge spillover among 61 countries. The last section concludes. All mathematical proofs are relegated to the Supplementary Material.

Throughout the paper, we use $i = 1, \ldots, n$ to denote an individual and $t = 1, \ldots, T$ to denote time. All asymptotic theories are established with n going to infinity. The total time period T can be fixed or tend to infinity as $n \to \infty$. For natural numbers l and m, I_m denotes an $m \times m$ identity matrix, $\mathbf{0}_m$ denotes an $m \times 1$ vector of zeros and $\mathbf{0}_{l\times m}$ denotes an $l \times m$ matrix of zeros. For a matrix A, let $\|A\| = [\mathrm{tr}(AA')]^{1/2}$, where $\mathrm{tr}(\cdot)$ is the trace of a matrix. Let $\|A\|_1$ and $\|A\|_\infty$ be its maximum absolute column and row sums, respectively. When A is a square matrix, we use $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ to denote its smallest and largest eigenvalues, respectively. Further, A^- denotes a symmetric generalized inverse of A, and $A^{1/2}$ denotes the square root matrix of A if it is symmetric nonnegative definite. For a general matrix A, we denote $\sigma^2_{\min}(A) \equiv \lambda_{\min}(A'A)$ and $\sigma^2_{\max}(A) \equiv \lambda_{\max}(A'A)$. Finally, if $\{a_{nT}\}$ and $\{b_{nT}\}$ are two sequences of nonnegative numbers, $a_{nT} \lesssim b_{nT}$ means there exists a finite positive C such that $a_{nT} \leq C b_{nT}$ for all n sufficiently large, and $a_{nT} \asymp b_{nT}$ means $a_{nT} \lesssim b_{nT}$ and $b_{nT} \lesssim a_{nT}$.

2 Models and moment conditions

2.1 Model specification

We consider the nonlinear SAR panel data model (2). The matrix form of the model can be written as

$$Y_{nt} = W_n \mathbf{h}_0(Y_{nt}) + X_{nt}\beta_0 + \mathbf{c}_{n0} + \mathcal{E}_{nt}, \quad t = 1, \ldots, T, \tag{3}$$

where $Y_{nt} = (y_{1t}, \ldots, y_{nt})'$ and $\mathcal{E}_{nt} = (\epsilon_{1t}, \ldots, \epsilon_{nt})'$ are $n \times 1$ random vectors, $y_{it} \in \mathcal{R}_{\mathcal{Y}}$ is the outcome variable of interest, ϵ_{it}'s are i.i.d. across i and t with zero mean and variance σ_0^2, and $\mathbf{h}_0(Y_{nt}) = (h_0(y_{1t}), \ldots, h_0(y_{nt}))'$ with $h_0(\cdot)$ being an unknown nonparametric function defined on $\mathcal{R}_{\mathcal{Y}}$. The $W_n = (w_{ij})$ is an $n \times n$ nonstochastic spatial weights matrix that generates the spatial dependence among cross-sectional units, $X_{nt} = (x_{1t}, \ldots, x_{nt})'$ is an $n \times d_X$ matrix of exogenous time-varying regressors, and $\mathbf{c}_{n0}$ is an $n \times 1$ vector of fixed effects. The spatial weights matrix W_n can be time-varying as long as it is exogenously given.

Following Jenish and Prucha (2009, 2012), we assume that individuals $\{1, 2, \ldots, n\}$ are located on a possibly unevenly spaced lattice $\mathcal{D}_n \subset \mathcal{D} \subset \mathbb{R}^d$ with $d \geq 1$. Let $l : \{1, 2, \ldots, n\} \to \mathcal{D}_n \subset \mathcal{D}$ be a mapping from individual i to its location $l(i) \in \mathcal{D} \subseteq \mathbb{R}^d$, and $\rho(i, j)$ be the distance between individual i and j. We maintain the following assumption on $\mathcal{D}$.

Assumption 2.1. *The lattice $\mathcal{D} \subset \mathbb{R}^d$, $d \geq 1$ is infinitely countable. The locations $\{l(i) : 1 \leq i \leq n\}$ are time-invariant and all elements in $\mathcal{D}$ are located at distances of at least 1 from each other, i.e., $\rho(i, j) \geq 1$ for $i \neq j$.*

The assumption of a minimum distance ensures the growth of the cross-sectional sample size n via expansion of the sample region $\mathcal{D}_n \subset \mathcal{D}$, i.e. increasing domain asymptotics. The space $\mathcal{D}$ can be a geographical space, a space of economic characteristics, or a mixture of both. Correspondingly, the distance may refer to physical and/or economic distance induced from any norm on $\mathbb{R}^d$ (Xu and Lee, 2015). Additionally, we assume that each individual does not necessarily interact with all other $n - 1$ units but only those located within a limited distance from him. Specifically, define the δ-neighborhood of individual i as $\mathcal{N}_{n,i}(\delta) = \{j \in \{1, \ldots, n\} \backslash \{i\} : \rho(i, j) \leq \delta\}$.

Assumption 2.2. *(i) For all $i = 1, \ldots, n$ and $n \geq 1$, there exists a fixed threshold distance $\bar{\Delta}$ such that $w_{ij} = 0$ for any $j \notin \mathcal{N}_{n,i}(\bar{\Delta})$. (ii) The spatial weights matrix W_n is exogenously given and satisfies $\sup_n(\|W_n\|_1 + \|W_n\|_\infty) < \infty$.*

The endogenous outcomes y_{it}'s are simultaneously determined by model (3), and thus $\{Y_{nt}\}$ can be viewed as an "equilibrium" of the system of nonlinear simultaneous equations. The following assumption guarantees that there exists a unique solution to model (3) for each t, see Hoshino (2022) for detailed justification.

Assumption 2.3. *(i) $\mathcal{R}_\mathcal{Y}$ is a closed subset of* $\mathbb{R}$.[3] *(ii) h_0 is Lipschitz continuous on $\mathcal{R}_\mathcal{Y}$ with Lipschitz constant $\mathcal{K}$, i.e., for any $y, y' \in \mathcal{R}_\mathcal{Y}$, $|h_0(y) - h_0(y')| \leq \mathcal{K}|y - y'|$, and $\mathcal{K}\|W_n\|_\infty < 1$.*

If one adopts the linear SAR model where $h_0(y) = \lambda_0 y$ and W_n is row-normalized, Assumption 2.3(ii) reduces to $|\lambda_0| < 1$. Assumptions 2.1-2.3 play an important role in ensuring the data follow a near-epoch dependent (NED) process (see Definition A.1 for the definition of NED process), and are maintained throughout the paper.

2.2 Moment conditions

For estimation of β_0 and h_0, we take the first difference of (3) to eliminate the individual effects $\mathbf{c}_{n0}$. For any vector/matrix Z_{nt}, let $\Delta Z_{nt} \equiv Z_{nt} - Z_{n,t-1}$. Then

$$\Delta Y_{nt} = W_n(\mathbf{h}_0(Y_{nt}) - \mathbf{h}_0(Y_{n,t-1})) + \Delta X_{nt}\beta_0 + \Delta \mathcal{E}_{nt}, \quad t = 2, \ldots, T. \tag{4}$$

Assume that the true parameters $\theta_0 = (\beta_0, h_0) \in \Theta = \mathcal{B} \times \mathcal{H}$, where $\mathcal{B}$ is a compact subset of $\mathbb{R}^{d_X}$ and $\mathcal{H}$ is a weighted Hölder ball[4] with radius c_1 and smoothness $\mu_1 > 1/2$, denoted as $\Lambda^{\mu_1}_{c_1}(\mathcal{R}_\mathcal{Y}, \omega_1)$. Then we can approximate h_0 in a finite-dimensional sieve space (e.g., polynomials, splines, wavelets). Specifically, let $\{k_j(\cdot) : j = 1, 2, \ldots\}$ be a sequence of sieve basis functions on $\mathcal{R}_\mathcal{Y}$, and $k^J(y) = (k_1(y), \ldots, k_J(y))'$ for some $J \equiv J_n$ (or $J = J_{nT}$ if $(n, T) \to \infty$). Further, let $\mathcal{H}_J$ be the sieve space of dimension J, that is, $\mathcal{H}_J \equiv \{h_J(\cdot) = k^J(\cdot)'\gamma$ for all γ satisfying $\|h_J(\cdot)[1 + |\cdot|^2]^{-\omega_1/2}\|_{\Lambda^{\mu_1}} \leq c_1\}$. For sufficiently large J, we can find a vector $\gamma_0 = (\gamma_{01}, \ldots, \gamma_{0,J})'$ such that $h_0(\cdot)$ can be well approximated by $h_{0,J}(\cdot) \equiv k^J(\cdot)'\gamma_0 \in \mathcal{H}_J$. Then, Equation (4) could be rewritten as

$$\Delta Y_{nt} = W_n \Delta K_{nt}\gamma_0 + \Delta X_{nt}\beta_0 + \Delta U_{nt}, \quad t = 2, \ldots, T, \tag{5}$$

where $K_{nt} = (k^J(y_{1t}), \ldots, k^J(y_{nt}))'$ and $U_{nt} = \mathcal{E}_{nt} + W_n(\mathbf{h}_0(Y_{nt}) - K_{nt}\gamma_0)$. Equation (5) implies that our model can be approximated by a standard regression model with multiple endogenous regressors $\Delta \bar{k}_{it} \equiv \sum_{j=1}^n w_{ij}(k^J(y_{jt}) - k^J(y_{j,t-1}))$. Let $z_{it} = (z_{it}^{(1)}, \ldots, z_{it}^{(d_Z)})' \in \mathcal{R}_Z$ be a $d_Z \times 1$ vector of continuous exogenous variables, for example, a subvector of x_{it}. Further,

[3] A subset A of $\mathbb{R}$ is open if and only if for each $a \in A$, there exists $\delta > 0$ such that $(a - \delta, a + \delta) \subset A$. A subset S of $\mathbb{R}$ is closed if and only if its complement $\bar{S} = \mathbb{R} \backslash S$, is open. As such, there exist unbounded sets like $[a, +\infty)$ and $\mathbb{R}$ that are closed in $\mathbb{R}$. This is in contrast to the concept of compact subsets in $\mathbb{R}$, which means both closedness and boundedness.

[4] For any real-valued $\mu > 0$, let $\underline{\mu}$ be the largest integer smaller than μ. Denote $\Lambda^\mu(\mathcal{R}_\mathcal{Y})$ as a set of functions: $h : \mathcal{R}_\mathcal{Y} \to \mathbb{R}$ which is $\underline{\mu}$-times continuously differentiable, and the $\underline{\mu}$th derivative, $\nabla^{\underline{\mu}} h(y)$, is Hölder continuous with exponent $\mu - \underline{\mu}$. The space $\Lambda^\mu(\mathcal{R}_\mathcal{Y})$ becomes a Banach space under the Hölder norm: $\|h\|_{\Lambda^\mu} = \sup_{\alpha \leq \underline{\mu}} \sup_{y \in \mathcal{R}_\mathcal{Y}} |\nabla^\alpha h(y)| + \sup_{y \neq y'} |\nabla^{\underline{\mu}} h(y) - \nabla^{\underline{\mu}} h(y')|/|y - y'|^{\mu - \underline{\mu}}$. Let $\Lambda^\mu(\mathcal{R}_\mathcal{Y}, \omega)$ denote the weighted Hölder space of functions $h : \mathcal{R}_\mathcal{Y} \to \mathbb{R}$ such that $h(\cdot)[1 + |\cdot|^2]^{-\omega/2}$ is in $\Lambda^\mu(\mathcal{R}_\mathcal{Y})$. We call $\Lambda^\mu_c(\mathcal{R}_\mathcal{Y}, \omega) \equiv \{h \in \Lambda^\mu(\mathcal{R}_\mathcal{Y}, \omega) : \|h(\cdot)[1 + |\cdot|^2]^{-\omega/2}\|_{\Lambda^\mu} \leq c < \infty\}$ a weighted Hölder ball with radius c and smoothness μ. See Ai and Chen (2003) and Chen et al. (2005) for details.

let $\{q_l(\cdot) : l = 1, 2, \ldots\}$ be a sequence of basis functions on $\mathcal{R}_Z$ and $q^L(z) = (q_1(z), \ldots, q_L(z))'$ for some $L \equiv L_n$ (or $L = L_{nT}$ if $(n, T) \to \infty$). Then, $\sum_{j=1}^n w_{ij}(q^L(z_{jt}) - q^L(z_{j,t-1}))$ would serve as valid instruments for $\Delta\bar{k}_{it}$ if z_{it} is correlated with y_{it}.[5]

Given a sequence of valid instruments, we can construct linear moment conditions based on the instrument orthogonality and obtain a sieve IV estimator for θ_0. As noted above, an instrument matrix for $(W_n\Delta K_{nt}, \Delta X_{nt})$ in (5) can take the form $\Delta B_{nt} \equiv (W_n\Delta Q_{nt}, \Delta X_{nt})$, where $Q_{nt} = (q^L(z_{1t}), \ldots, q^L(z_{nt}))'$, which satisfies that $\mathrm{E}(\Delta B_{nt}'\Delta\mathcal{E}_{nt}) = \mathbf{0}$ for $t = 2, \ldots, T$. For $\theta = (\beta, h) \in \Theta$, let $\Delta U_{nt}(\theta) = \Delta Y_{nt} - W_n\Delta\mathbf{h}(Y_{nt}) - \Delta X_{nt}\beta$. Then the moment functions corresponding to the orthogonality between ΔB_{nt} and $\Delta\mathcal{E}_{nt}$ are $\Delta\mathbf{B}_{n,T-1}'\Delta\mathbf{U}_{n,T-1}(\theta)$, where $\Delta\mathbf{B}_{n,T-1} = (\Delta B_{n2}', \ldots, \Delta B_{nT}')'$ and $\Delta\mathbf{U}_{n,T-1}(\theta) = (\Delta U_{n2}'(\theta), \ldots, \Delta U_{nT}'(\theta))'$. The sieve IV estimator of θ_0 can be obtained by minimizing the distance from $\Delta\mathbf{B}_{n,T-1}'\Delta\mathbf{U}_{n,T-1}(\theta)$ to zero over the sieve space $\Theta_n \equiv \mathcal{B} \times \mathcal{H}_J$.

The instruments, if all generated from x-variables, are valid only if $\beta_0 \neq \mathbf{0}$. Otherwise, the identification fails and the resulting IV estimator is inconsistent. If the exogenous variables are relevant but weak in predicting the outcome, ΔB_{nt} is also not going to be a good instrument in that the corresponding sieve IV estimator suffers from severe ill-posed problem and the variance of the estimated sieve coefficients could be high (Newey, 2013). In the literature of spatial econometrics, several studies have shown that quadratic moments based on the covariance of model disturbances can capture spatial correlation and increase the efficiency of the estimates, see, e.g., Lee (2007), Lee and Yu (2014) and Sun and Malikov (2018). Combining linear moments with quadratic moments leads to consistent estimates of the unknown parameters regardless of whether the exogenous covariates are relevant in predicting the outcome or not (Malikov and Sun, 2017). We expect similar results to hold for our nonlinear SAR panel data model.

Note that for an $n \times n$ matrix P_{ln}, $\mathrm{E}(\Delta\mathcal{E}_{nt}'P_{ln}\Delta\mathcal{E}_{nt}) = 2\sigma_0^2\mathrm{tr}(P_{ln})$, thus the vector $P_{ln}\Delta\mathcal{E}_{nt}$ can be uncorrelated with $\Delta\mathcal{E}_{nt}$ for any matrix P_{ln} satisfying $\mathrm{tr}(P_{ln}) = 0$, while it is correlated with $W_n\Delta K_{nt}$ in general. Denote $\mathbf{P}_{ln,T-1} \equiv I_{T-1} \otimes P_{ln}$ where $\mathrm{tr}(P_{ln}) = 0$. This motivates us to use the following quadratic moments $\Delta\mathbf{U}_{n,T-1}'(\theta)\mathbf{P}_{ln,T-1}\Delta\mathbf{U}_{n,T-1}(\theta)$ for $l = 1, \ldots, m$ with $m = m_n$ (or $m = m_{nT}$ if $(n, T) \to \infty$). Let $N \equiv n(T-1)$. With the selected IVs and matrices P_{ln}'s, the set of moment functions for the GMM estimation is

$$g_{nT}(\theta) = \frac{1}{N}\begin{pmatrix} \Delta\mathbf{U}_{n,T-1}'(\theta)\mathbf{P}_{1n,T-1}\Delta\mathbf{U}_{n,T-1}(\theta) \\ \vdots \\ \Delta\mathbf{U}_{n,T-1}'(\theta)\mathbf{P}_{mn,T-1}\Delta\mathbf{U}_{n,T-1}(\theta) \\ \Delta\mathbf{B}_{n,T-1}'\Delta\mathbf{U}_{n,T-1}(\theta) \end{pmatrix}. \tag{6}$$

[5]Given $q^L(z_{it})$ as instruments for $k^J(y_{it})$, we consider $\Delta q^L(z_{it})$ as instruments for $\Delta k^J(y_{it})$, which is standard in the panel data literature (Zhang and Zhou, 2021). Alternatively, one can utilize $(q^L(z_{it})', q^L(z_{i,t-1})')'$ as instruments. Whether this will lead to a better performance of the IV estimator is an open question.

In (6), the number of quadratic moments m and linear moments $(d_X + L)$ satisfy $m + L \geq J$ and both of them diverge as $n \to \infty$. Denote $\Delta\boldsymbol{\mathcal{E}}_{n,T-1} \equiv (\Delta\mathcal{E}'_{n2}, \ldots, \Delta\mathcal{E}'_{nT})'$. At the true value $\theta_0 = (\beta_0, h_0)$, $g_{nT}(\theta_0) = (\Delta\boldsymbol{\mathcal{E}}'_{n,T-1}\mathbf{P}_{1n,T-1}\Delta\boldsymbol{\mathcal{E}}_{n,T-1}, \ldots, \Delta\boldsymbol{\mathcal{E}}'_{n,T-1}\mathbf{P}_{mn,T-1}\Delta\boldsymbol{\mathcal{E}}_{n,T-1}, \Delta\boldsymbol{\mathcal{E}}'_{n,T-1}\Delta\mathbf{B}_{n,T-1})'/N$, which has zero expectation. Let $d_g \equiv m + d_X + L$. The sieve GMM estimator for θ_0 is defined as

$$\widehat{\theta}_{nT} = (\widehat{\beta}_{nT}, \widehat{h}_{nT,J}) = \arg\min_{\theta \in \Theta_n} \widehat{\mathcal{Q}}_{nT}(\theta), \tag{7}$$

where

$$\widehat{\mathcal{Q}}_{nT}(\theta) = \frac{1}{d_g} g'_{nT}(\theta) \mathcal{W}_{nT} g_{nT}(\theta),$$

and $\mathcal{W}_{nT}$ is a $d_g \times d_g$ positive-definite weighting matrix. In the paper, we assume that the largest and the smallest eigenvalue of $\mathcal{W}_{nT}$ are bounded and bounded away from zero for all n and T. As in Dong et al. (2023), the involvement of d_g in $\widehat{\mathcal{Q}}_{nT}(\theta)$ takes into account the diverging dimension of $g_{nT}(\theta)$ and ensures that $\widehat{\mathcal{Q}}_{nT}(\theta)$ uniformly converges to its probability limit even if d_g diverges as $n \to \infty$.

Notably, the identification of our model comes from two types of moment conditions. Denote $\Delta u_{it}(\theta) = \Delta y_{it} - \sum_{j=1}^n w_{ij}\Delta h(y_{jt}) - \Delta x'_{it}\beta$ for $\theta = (\beta, h)$. The first type is conditional moment restrictions derived from instrument exogeneity

$$L_{it}(\theta_0) = 0, \quad \text{where } L_{it}(\theta) = \mathrm{E}[\Delta u_{it}(\theta) | \mathbf{w}_{it}, \mathbf{w}_{i,t-1}], \tag{8}$$

and $\mathbf{w}_{it} = \{x_{it}\} \cup \{z_{jt} : j \in \mathcal{N}_{n,i}(\bar{\Delta})\}$. The second set consists of unconditional moment restrictions based on the properties of the model errors

$$M_{nt}(\theta_0) = 0, \quad \text{where } M_{nt}(\theta) \equiv \big(\mathrm{E}(\Delta U'_{nt}(\theta) P_{1n} \Delta U_{nt}(\theta)), \ldots, \mathrm{E}(\Delta U'_{nt}(\theta) P_{mn} \Delta U_{nt}(\theta))\big)', \tag{9}$$

where $\{P_{ln} : 1 \leq l \leq m\}$ is a sequence of $n \times n$ matrix with zero trace. Let the identification set be $\overline{\Theta} = \{\theta \in \Theta : L_{it}(\theta) = 0 \text{ and } M_{nt}(\theta) = 0 \text{ for all } i = 1, \ldots, n,\ t = 2, \ldots, T\}$. Then θ_0 is identified if and only if $\overline{\Theta} = \{\theta_0\}$.

Identification can be achieved through conditional moment restrictions (8) if for any $\theta \in \Theta$, $L_{it}(\theta) = 0$ implies that $\theta = \theta_0$. This corresponds to the so-called completeness condition, which, in our context, is equivalent to the nonexistence of any function $(\tilde{\beta}, \tilde{h}) = (\beta - \beta_0, h - h_0) \neq 0$ such that $\mathrm{E}[\sum_{j=1}^n w_{ij}\Delta\tilde{h}(y_{jt}) + \Delta x'_{it}\tilde{\beta} | \mathbf{w}_{it}, \mathbf{w}_{i,t-1}] = 0$ for all $i = 1, \ldots, n$. As noted by Hoshino (2022), identifiability of h_0 based on such condition depends crucially on the magnitude of the parameter β_0.[6] If it were the case that $\beta_0 = \mathbf{0}$, z_{it} cannot be used as valid instruments and h_0 would not be identified. For illustration, consider the interaction

[6]Here we assume that the instruments z_{it} is a subvector of x_{it}.

model (3) where X_{nt} is independent of $(\mathbf{c}_{n0}', \mathcal{E}_{nt}')'$. If $\beta_0 = \mathbf{0}$, the value of y_{it}'s are determined solely by ϵ_{it}'s and c_{i0}'s, implying that $\mathrm{E}\big[\sum_{j=1}^n w_{ij}\Delta\tilde{h}(y_{jt})|\mathbf{w}_{it}, \mathbf{w}_{i,t-1}\big] = \mathrm{E}[\sum_{j=1}^n w_{ij}\Delta\tilde{h}(y_{jt})]$. Clearly, $\mathrm{E}[\sum_{j=1}^n w_{ij}\Delta\tilde{h}(y_{jt})] = 0 \not\Rightarrow \tilde{h}(y_{jt}) = 0$. In this case, the true parameter θ_0 is observationally equivalent to any $(\beta_0, h) \in \Theta$ such that $\mathrm{E}[\sum_{j=1}^n w_{ij}(\Delta h(y_{jt}) - \Delta h_0(y_{jt}))] = 0$ for all i and t.

In the absence of completeness assumptions, the quadratic moment conditions (9) have nontrivial identification power, and can provide meaningful restrictions on the identification set.[7] By the decomposition $\Delta U_{nt}(\theta) = \Delta\mathcal{E}_{nt} + W_n(\Delta\mathbf{h}_0(Y_{nt}) - \Delta\mathbf{h}(Y_{nt})) + \Delta X_{nt}(\beta_0 - \beta)$ and moment conditions (9), it is easily seen that θ_0 is locally identified on the set $\Xi \equiv \{(\beta_0, h) \in \Theta : \mathrm{E}\big[\Delta U_{nt}'(\theta)P_{ln}\Delta U_{nt}(\theta)\big] \neq 0 \text{ for some } l = 1, \ldots, m\}$, where

$$\begin{aligned}\mathrm{E}[\Delta U_{nt}'(\theta)P_{ln}\Delta U_{nt}(\theta)] = &\mathrm{E}\big[\Delta\mathcal{E}_{nt}'P_{ln}^s W_n(\Delta\mathbf{h}_0(Y_{nt}) - \Delta\mathbf{h}(Y_{nt}))\big] \\ &+ \mathrm{E}\big[(\Delta\mathbf{h}_0(Y_{nt}) - \Delta\mathbf{h}(Y_{nt}))'W_n'P_{ln}W_n(\Delta\mathbf{h}_0(Y_{nt}) - \Delta\mathbf{h}(Y_{nt}))\big].\end{aligned}$$

The set Ξ is a nontrivial set. For example, for any function $h \in \mathcal{H}$ such that $h(y) = h_0(y) + ay$ with $a \neq 0$, it belongs to the set Ξ as long as for some $l = 1, \ldots, m$, $\mathrm{E}(\Delta Y_{nt}'W_n'P_{ln}W_n\Delta Y_{nt}) \neq 0$, and $a \neq -\frac{\mathrm{E}(\Delta\mathcal{E}_{nt}'P_{ln}^s W_n\Delta Y_{nt})}{\mathrm{E}(\Delta Y_{nt}'W_n'P_{ln}W_n\Delta Y_{nt})}$. Clearly, the identification power of quadratic moments depends crucially on the choice of $\{P_{ln}\}_{l=1}^m$ and the joint distribution of $\mathcal{E}_{nt}$ and Y_{nt}. In the linear SAR model where $h_0(y) = \lambda_0 y$, we can derive some preliminary conditions to ensure the identifiability of λ_0 through the quadratic moment conditions, see Lee (2007). In the nonparametric model, the quadratic moment conditions can improve the identification if the completeness condition does not hold, but may not guarantee the point-identification of h_0 in general cases.

3 Asymptotic properties of sieve GMME

We first list assumptions used to derive the convergence rate and limiting distribution of the proposed estimator.

Assumption 3.1. *(i) The function h_0 belongs to a weighted Hölder ball $\mathcal{H} = \Lambda_{c_1}^{\mu_1}(\mathcal{R}_\mathcal{Y}, \omega_1)$ with $\mu_1 > 1/2$ and $\omega_1 \geq 0$. (ii) $\sup_{n,T}\sup_{i,t}\mathrm{E}|y_{it}|^{4\omega} < \infty$ for some $\omega > \max\{1, \omega_1\}$. (iii) X_{nt}'s are nonstochastic with $\sup_{n,T}\sup_{i,t}|x_{it}^{(l)}| < \infty$ for $l = 1, \ldots, d_X$, where $x_{it}^{(l)}$ is the (i,l)th element of X_{nt}.*

Assumption 3.2. *The disturbances $\{\epsilon_{it}\}$, $i = 1, 2, \ldots, n$ and $t = 1, 2, \ldots, T$, are i.i.d. across i and t with zero mean, variance σ_0^2 and $\sup_{n,T}\sup_{i,t}\mathrm{E}|\epsilon_{it}|^p < \infty$ for some $p \geq 8$.*

[7] In the literature on NPIV models, there is a growing interest in the use of shape restrictions to achieve or improve identification, see, e.g., Freyberger and Horowitz (2015) and Chetverikov and Wilhelm (2017). The use of quadratic moment conditions shares the same spirit with these work in the sense that both quadratic moment conditions and shape restrictions provide useful auxiliary information about the parameter.

Assumption 3.3. *n goes to infinity. T is a nondecreasing function of n, which can be finite or tend to infinity as $n \to \infty$.*

Assumption 3.1(i) is a standard weighted smoothness condition imposed on the function h_0; see Chen et al. (2005) for more technical details. The moment conditions imposed on $\{y_{it}\}$ and $\{\epsilon_{it}\}$ are stronger than those of Hoshino (2022) due to the quadratic moments in use. Similar to Remark 3.1 in Hoshino (2022), we can find some primitive conditions for Assumption 3.1(ii). For example, assume that h_0 is differentiable with $\mathcal{K} = \sup_{y \in \mathcal{R}_\mathcal{Y}} |\nabla h_0(y)|$, where $\mathcal{K}$ is given in Assumption 2.3(ii). Then Assumption 3.1(iii) and $\sup_{n,T} \sup_{i,t} \mathrm{E}|\epsilon_{it}|^{4\omega} < \infty$ imply Assumption 3.1(ii). The i.i.d. property in Assumption 3.2 simplifies the presentation and calculations. Under a heteroscedasticity setting, we can choose P_{ln} with zero diagonals so that the quadratic moment conditions still hold, and the asymptotic analysis can be extended in a similar way as in Lin and Lee (2010). Assumption 3.3 allows two cases of interest: (i) n is large and T is fixed, and (ii) both n and T are large.

3.1 Consistency and convergence rates

We establish the consistency of $\widehat{\theta}_{nT}$ under the metric $d(\theta, \theta_0) \equiv \|\beta - \beta_0\| + |h - h_0|_{\infty,\omega}$, where $|\cdot|_{\infty,\omega}$ signifies the weighted sup-norm, i.e.,

$$|h|_{\infty,\omega} \equiv \sup_{y \in \mathcal{R}_\mathcal{Y}} \left[|h(y)| \cdot (1 + |y|^2)^{-\omega/2} \right]$$

for some $\omega \geq 0$; see, e.g., Chen et al. (2005) and Su and Jin (2012). As Chen et al. (2005) remark, the weight function $(1+|y|^2)^{-\omega/2}$ can be regarded as an alternative to the trimming function used in kernel estimation when the support is unbounded.

Assumption 3.4. *(i) The basis functions $\{q_l(\cdot) : l = 1, 2, \ldots\}$ are uniformly bounded on $\mathcal{R}_Z$. (ii) The basis functions $\{k_j(\cdot) : j = 1, 2, \ldots\}$ satisfy $\sup_{i,t} \sup_{1 \leq j \leq J} \mathrm{E}|k_j(y_{it})|^{4\delta} < \infty$ for some $\delta > 1$. For all j, k_j is Lipschitz continuous on $\mathcal{R}_\mathcal{Y}$ with Lipschitz constant $\bar{c}_{kj}$. (iii) The sequence $\{P_{ln} : 1 \leq l \leq m\}$ satisfies $tr(P_{ln}) = 0$ and $\sup_n \sup_{1 \leq l \leq m} (\|P_{ln}\|_1 + \|P_{ln}\|_\infty) < \infty$.*

Assumption 3.5. *There exists a sequence of vectors $\gamma_0 \equiv \gamma_{0J}$ such that $|h_0(\cdot) - k^J(\cdot)'\gamma_0|_{\infty,\omega} = o(1)$.*

Assumption 3.4(i) and (ii) impose some conditions on the choice of basis functions. It can be verified that the B-spline basis and Fourier series satisfy the conditions. Assumption 3.4(iii) ensures the orthogonality between $P_{ln}\Delta\mathcal{E}_{nt}$ and $\Delta\mathcal{E}_{nt}$. The uniform boundedness of P_{ln}'s is used in justifying the uniform convergence of the quadratic moment functions to their probability limit over Θ_n. Assumption 3.5 is from the definition of sieve spaces. For $h_0 \in \Lambda_{c_1}^{\mu_1}(\mathcal{R}_\mathcal{Y}, \omega_1)$, it is trivially satisfied by the aforementioned basis.

Theorem 3.1 in Chen (2007) and Lemma A.2 in Chen and Pouzo (2012) provide sufficient conditions for a general sieve estimator to be consistent. We shall verify those conditions for our sieve GMM estimator. To do so, denote $\mathcal{Q}_{nT}(\theta) \equiv d_g^{-1}\mathrm{E}[g_{nT}(\theta)]'\mathcal{W}_{nT}\mathrm{E}[g_{nT}(\theta)]$. It is obvious that $\mathcal{Q}_{nT}(\theta) \geq 0$ for $\theta \in \Theta$ and $\mathcal{Q}_{nT}(\theta_0) = 0$. For consistency, we need the following identification condition to hold.

Assumption 3.6. *For any $\epsilon > 0$, $\liminf_{n\to\infty} \inf_{\theta\in\Theta:d(\theta,\theta_0)\geq\epsilon} \mathcal{Q}_{nT}(\theta) > 0$.*

Assumption 3.6 guarantees the uniqueness of the true parameter θ_0 in the parameter space satisfying the moment conditions. This type of condition is commonly made in the conventional literature, see, e.g., White and Wooldridge (1991), Chen (2007) and Dong et al. (2023).

A consistency result follows from the above conditions.

Theorem 3.1. *Under Assumptions 2.1-2.3, 3.1, 3.3, 3.4(i)(iii) and 3.5-3.6, the sieve GMM estimator $(\widehat{\beta}_{nT}, \widehat{h}_{nT,J})$ given by (7) is consistent, i.e., $\|\widehat{\beta}_{nT} - \beta_0\| + |\widehat{h}_{nT,J} - h_0|_{\infty,\omega} = o_p(1)$.*

We proceed to study the convergence rate of the sieve GMM estimator. For any square matrix A, let $\mathrm{vec}(A)$ be the column vector formed by stacking the columns of A, and $\mathrm{vec}_D(A)$ be the column vector formed with the diagonal elements of A and $A^s \equiv A + A'$. It follows from Lemma A.8 in the Supplementary Material that $\mathrm{Var}(\sqrt{N}g_{nT}(\theta_0)) = \Omega_{nT}$, where

$$\Omega_{nT} = \frac{1}{N}\begin{pmatrix} 2(2T-3)(\mu_4 - 3\sigma_0^4)\omega_{nm}'\omega_{nm} & \mu_3\omega_{nm}'(\Delta B_{n2} - \Delta B_{nT}) \\ \mu_3(\Delta B_{n2} - \Delta B_{nT})'\omega_{nm} & \mathbf{0}_{(d_X+L)\times(d_X+L)} \end{pmatrix} + \mathcal{V}_{nT}, \tag{10}$$

with $\mu_3 \equiv \mathrm{E}(\epsilon_{it}^3)$, $\mu_4 \equiv \mathrm{E}(\epsilon_{it}^4)$, $\omega_{nm} \equiv [\mathrm{vec}_D(P_{1n}), \ldots, \mathrm{vec}_D(P_{mn})]$, and

$$\mathcal{V}_{nT} \equiv \frac{\sigma_0^4}{N}\begin{pmatrix} 2n(3T-4)\Psi_{n,m} & \mathbf{0}_{m\times(d_X+L)} \\ \mathbf{0}_{(d_X+L)\times m} & \frac{1}{\sigma_0^2}\mathcal{V}_{n,T-1}^B \end{pmatrix} \tag{11}$$

with $\Psi_{n,m} \equiv (1/n)[\mathrm{vec}(P_{1n}'), \ldots, \mathrm{vec}(P_{mn}')]'[\mathrm{vec}(P_{1n}^s), \ldots, \mathrm{vec}(P_{mn}^s)]$, and $\mathcal{V}_{n,T-1}^B \equiv 2\sum_{t=2}^{T}\Delta B_{nt}'\Delta B_{nt} - \sum_{t=2}^{T-1}(\Delta B_{nt}'\Delta B_{n,t+1} + \Delta B_{n,t+1}'\Delta B_{nt})$. The first term in (10) vanishes if ϵ_{it}'s are normally distributed or the matrices P_{ln}'s have zero diagonals.

Denote $\Psi_{nT,L} \equiv \Delta\mathbf{B}_{n,T-1}'\Delta\mathbf{B}_{n,T-1}/N$, $\Psi_{nT,J} \equiv \mathrm{E}[\Delta\mathbf{K}_{n,T-1}'\Delta\mathbf{K}_{n,T-1}/N]$, and

$$D_{nT} \equiv N^{-1}\mathrm{E}\begin{pmatrix} \mathbf{0}_{d_X}' & \Delta\boldsymbol{\mathcal{E}}_{n,T-1}'\mathbf{P}_{1n,T-1}^s\mathbf{W}_{n,T-1}\Delta\mathbf{K}_{n,T-1} \\ \vdots & \vdots \\ \mathbf{0}_{d_X}' & \Delta\boldsymbol{\mathcal{E}}_{n,T-1}'\mathbf{P}_{mn,T-1}^s\mathbf{W}_{n,T-1}\Delta\mathbf{K}_{n,T-1} \\ \Delta\mathbf{B}_{n,T-1}'\Delta\mathbf{X}_{n,T-1} & \Delta\mathbf{B}_{n,T-1}'\mathbf{W}_{n,T-1}\Delta\mathbf{K}_{n,T-1} \end{pmatrix} = (D_{nT,X}, D_{nT,J}), \tag{12}$$

where $\mathbf{W}_{n,T-1} \equiv I_{T-1} \otimes W_n$, $D_{nT,X}$ and $D_{nT,J}$ are $d_g \times d_X$ and $d_g \times J$ sub-matrices of D_{nT}, respectively. Finally, let $\Gamma_{nT,J} \equiv D_{nT,J} - D_{nT,X}(D'_{nT,X}\mathcal{W}_{nT}D_{nT,X})^{-1}D'_{nT,X}\mathcal{W}_{nT}D_{nT,J}$. Clearly, the matrix D_{nT} is related to the covariance between the variables $(p^s_{1,i\cdot}\Delta\mathcal{E}_{nt}, \ldots, p^s_{m,i\cdot}\Delta\mathcal{E}_{nt}, \Delta b'_{it})'$ and the regressors $(\Delta x'_{it}, \Delta\bar{k}'_{it})'$, where $p^s_{l,i\cdot}$ and $\Delta b'_{it}$ denote the ith row of P^s_{ln} and ΔB_{nt}, respectively. Besides, $\mathcal{W}^{1/2}_{nT}\Gamma_{nT,J}$ is the orthogonal complement of $\mathcal{W}^{1/2}_{nT}D_{nT,J}$ on the column space of $\mathcal{W}^{1/2}_{nT}D_{nT,X}$. Note that we suppress the dependence of D_{nT} on (J, L, m) and the dependence of $D_{nT,X}$, $D_{nT,J}$ and $\Gamma_{nT,J}$ on (L, m) for brevity.

Assumption 3.7. *(i) Uniformly in L and m, there exist constants $0 < \underline{c}_\Omega < \bar{c}_\Omega < \infty$ such that $\underline{c}_\Omega < \liminf_{n\to\infty} \lambda_{min}(\Omega_{nT}) \leq \limsup_{n\to\infty} \lambda_{max}(\Omega_{nT}) < \bar{c}_\Omega$.*

(ii) Uniformly in L and m, there exist constants $0 < \underline{c}_V < \bar{c}_V < \infty$ such that $\underline{c}_V < \liminf_{n\to\infty} \lambda_{min}(\Psi_{nT,L}) \leq \limsup_{n\to\infty} \lambda_{max}(\Psi_{nT,L}) < \bar{c}_V$ and $\underline{c}_V < \liminf_{n\to\infty} \lambda_{min}(\Psi_{n,m}) \leq \limsup_{n\to\infty} \lambda_{max}(\Psi_{n,m}) < \bar{c}_V$.

(iii) Uniformly in J, there exist constants $0 < \underline{c}_k < \bar{c}_k < \infty$ such that $\underline{c}_k < \liminf_{n\to\infty} \lambda_{min}\big(\Psi_{nT,J}\big) \leq \limsup_{n\to\infty} \lambda_{max}\big(\Psi_{nT,J}\big) < \bar{c}_k$.

(iv) Let $\bar{\nu}_J \equiv \limsup_{n\to\infty} \max\{\sigma^2_{max}(D_{nT,J}), \sigma^2_{max}(\Gamma_{nT,J})\}$ and $\underline{\nu}_J \equiv \liminf_{n\to\infty} \min\{\sigma^2_{min}(D_{nT,J}), \sigma^2_{min}(\Gamma_{nT,J})\}$. Uniformly in J, L and m, there exists a positive constant $\bar{c}_D < \infty$ such that $\bar{\nu}_J < \bar{c}_D$ and $\underline{\nu}_J > 0$.

Assumption 3.7(i) and (ii) ensure the nonsingularity of the matrices Ω_{nT}, $\Psi_{nT,L}$ and $\Psi_{n,m}$ for sufficiently large n. Assumption 3.7(iii) together with Assumption 2.2(ii) ensures that $\lambda_{\max}(\mathrm{E}(\Delta\mathbf{K}'_{n,T-1}\mathbf{W}'_{n,T-1}\mathbf{W}_{n,T-1}\Delta\mathbf{K}_{n,T-1}/N)) = O(1)$. For Assumption 3.7(iv), $\bar{\nu}_J$ and $\underline{\nu}_J$ depend on the strength of both linear and quadratic moments, where the former is related to the value of β_0 and the latter to the correlation between $\mathcal{E}_{nt}$ and the endogenous outcome Y_{nt}. These factors can indirectly affect the performance of the sieve estimator through $\bar{\nu}_J$ and $\underline{\nu}_J$. In particular, the quantity $\underline{\nu}_J$ can be interpreted as the sieve measure of ill-posedness of our model, see Remark 3.1 for a brief discussion.

Assumption 3.5′. *There exists a sequence of vectors $\gamma_0 \equiv \gamma_{0J}$ such that $k^J(\cdot)'\gamma_0 \in \mathcal{H}_J$ and $|h_0(\cdot) - k^J(\cdot)'\gamma_0|_{\infty,\omega} = O(J^{-\mu})$ for some $\mu > 1/2$.*

Assumption 3.5′ quantifies the approximation error of the function h_0 by the sieve space $\mathcal{H}_J$ in terms of the weighted sup-norm. The number μ depends on the smoothness of the function h_0, the choice of sieve spaces, and the size of ω. As we assume $h_0 \in \Lambda^{\mu_1}_{c_1}(\mathcal{R}_\mathcal{Y}, \omega_1)$, this assumption is satisfied by commonly used basis functions such as splines, wavelets and Fourier series with $\mu = \mu_1$ and $\omega = \omega_1 + \mu_1$; see Ai and Chen (2003) and Chen et al. (2005).

Assumption 3.8. *(i) The number of moments satisfies that $L + m \geq J$, and there exist constants $0 < c_1 < c_2 < \infty$ such that $c_1 < (L + m)/J < c_2$ uniformly in n and T, (ii) $J^3/(\underline{\nu}^5_J N) = o(1)$, and (iii) $\underline{\nu}^{-5}_J J^{1-2\mu} = o(1)$.*

Assumption 3.8(i) excludes the case where the number of moment conditions grows faster than the number of estimation parameters (i.e., $(L+m) \asymp J$). Assumption 3.8(ii) and (iii) jointly determine the growth order of the smoothing parameter J and the magnitude of μ. Compared with Assumption 3.5 in Hoshino (2022), the growth rate of J is required to be slower due to quadratic moments in use.

Remark 3.1. *An NPIV model is by nature ill-posed since conditional expectations "smooth out" nonlinearities. In a typical NPIV model with endogenous variable Y and instrumental variable W, the measure of ill-posedness refers to the speed at which the singular values of an operator $\mathcal{T}h = \mathrm{E}[h(Y)|W]$ decays (Blundell et al., 2007). However, in presence of spatial autocorrelation, such measure of ill-posedness could be inconvenient to use. By examining the relationship between the reduced form coefficient and the structural form coefficient, we can obtain an alternative (sieve) measure of ill-posedness of our model with only linear moment conditions. Specifically, for a given $h_J = k^J(\cdot)'\gamma \in \mathcal{H}_J$, consider the projection of $W_n \Delta \mathbf{h}_J(Y_{nt})$ onto the column space of the instrumental variable matrix ΔB_{nt}. The projection coefficient can be obtained by minimizing the average mean-squared error, i.e., $\min_{\delta \in \mathbb{R}^{d_X+L}} \sum_{t=2}^{T} \mathrm{E}\|W_n \Delta K_{nt}\gamma - \Delta B_{nt}\delta\|^2$, which has a closed-form solution $\Psi_{nT,L}^{-1} C_{nT,J}\gamma$, where $C_{nT,J} = \mathrm{E}[N^{-1}\sum_{t=2}^{T} \Delta B_{nt}' W_n \Delta K_{nt}]$. The projection of $\sum_{j=1}^{n} w_{ij}\Delta h_J(y_{jt})$ onto the instrument space is obtained as $\Delta b_{it}' \Psi_{nT,L}^{-1} C_{nT,J}\gamma$. Thus, the quantity $\sigma_{\min}^2(C_{nT,J})$ delivers information on the diminishing rate of the projection coefficients. The reciprocal of it can be interpreted as a sieve measure of ill-posedness for our model with only linear moments.*

Augmenting linear with quadratic moment conditions provides additional information about the nonlinearity of h_0, and helps alleviate the ill-posedness in the NPIV model. In view of the asymptotic expansion of the estimated sieve coefficients (see the proof of Lemma A.7 in the Supplementary Material), the sieve GMM estimator (7) is asymptotically equivalent to a sieve 2SLS estimator with "instruments" $\Delta V_{nt} \equiv (P_{1n}^s \Delta \mathcal{E}_{nt}, \ldots, P_{mn}^s \Delta \mathcal{E}_{nt}, \Delta B_{nt})$. This implies that the role of quadratic moments can be understood as providing a new source of "instruments" that are powerful in detecting nonlinearity. In this connection, we define the measure of ill-posedness in the GMM setting as $\tau_J = 1/\sqrt{\underline{\nu}_J}$, where $\underline{\nu}_J$ is given by Assumption 3.7(iv). This quantity measures the decaying rate of the minimum singular value of the matrix $D_{nT,J} = \mathrm{E}[N^{-1}\sum_{t=2}^{T} \Delta V_{nt}' W_n \Delta K_{nt}]$, which can be viewed as an augmented version of the matrix $C_{nT,J}$ above. We check the impact of quadratic instruments on the degree of ill-posedness numerically by conducting Monte Carlo experiments, the details of which are presented in Section C in the Supplementary Material. The experiments reveal that quadratic instruments significantly ameliorate the ill-posed problem especially when β_0 is close to zero.

We proceed to establish the convergence rate of the sieve estimator. Following Chen and Christensen (2015) and Su and Zhang (2016), we consider the uniform convergence rate on an expanding sequence of compact sets $\mathscr{D}_{nT} \subset \mathcal{R}_{\mathcal{Y}}$. For instance, we could take

$\mathscr{D}_{nT} = [-r_{nT}, r_{nT}]$ with $r_{nT} > 0$ being an increasing sequence of n and T. The following assumption specifies the expanding rate of $\mathscr{D}_{nT}$ and magnitude of the basis functions and their first derivations.

Assumption 3.9. *(i) There are a sequence of constants $\zeta_{nT,J}$ and a sequence of expanding compact sets $\mathscr{D}_{nT} \subset \mathcal{R}_\mathcal{Y}$ such that (a) $\sup_{y\in\mathscr{D}_{nT}} |y| \lesssim \zeta_{nT,J}^{1/\omega}$ for ω given in Assumption 3.5′, (b) $\sup_{y\in\mathscr{D}_{nT}} \|k^J(y)\| \leq \zeta_{nT,J}$, (c) $J^{1/2} \lesssim \zeta_{nT,J} \lesssim N^{\xi_1} J^{\xi_2}$ for some $\xi_1 \geq 0$ and $\xi_2 > 0$, and (d) $\zeta_{nT,J}^3 J^{1/2} \ln N/N = o(1)$. (ii) There exist constants $c_p > 0$ and $\varpi_1, \varpi_2 \geq 0$ such that $\|k^J(y) - k^J(y')\| \leq c_p N^{\varpi_1} J^{\varpi_2} |y - y'|$ for any finite y, $y' \in \mathscr{D}_{nT}$.*

Theorem 3.2. *Suppose that Assumptions 2.1-2.3 and 3.1-3.8 hold (where 3.5 is replaced by 3.5′).*

(i) Let $\mathcal{S}_1 = (I_{d_X}, \mathbf{0}_{d_X\times J})$. If $\lambda_{\max}(\mathcal{S}_1(D_{nT}'\mathcal{W}_{nT}D_{nT})^{-1}\mathcal{S}_1') = O(1)$, then $\|\widehat{\beta}_{nT} - \beta_0\| = O_p(J^{-\mu} + N^{-1/2})$.

(ii) If Assumption 3.9 also holds and the diagonals of the matrices P_{ln}'s are zero, then $\sup_{y\in\mathscr{D}_{nT}} |\widehat{h}_{nT,J}(y) - h_0(y)| = O_p\big(\tau_J\zeta_{nT,J}\big(J^{1/(2(p-1))}(\ln N/N)^{(p-2)/(2(p-1))} + J^{-\mu}\big)\big)$, where p is a constant given in Assumption 3.2.

Theorem 3.2(i) implies that if the number of basis terms J is such that $J^{-\mu} \asymp N^{-1/2}$, the GMM estimator $\widehat{\beta}_{nT}$ becomes $\sqrt{N}$-consistent. This coincides with the convergence rate of the sieve 2SLS estimator, see Theorem 3.4(i) of Hoshino (2022). Theorem 3.2(ii) gives the uniform convergence rate of $\widehat{h}_{nT,J}$ on $\mathscr{D}_{nT}$. In deriving the rate, we split $|\widehat{h}_{nT,J}(y) - h_0(y)|$ into "bias" and "standard deviation" terms and show that they are of order $O_p(\tau_J\zeta_{nT,J}J^{-\mu})$ and $O_p(\tau_J\zeta_{nT,J}J^{1/(2(p-1))}(\ln N/N)^{(p-2)/(2(p-1))})$, respectively. Chen and Christensen (2018) derived optimal sup-norm convergence rates for nonparametric IV regression models with i.i.d. data; that is, in our context, $\sup_{y\in\mathscr{D}_{nT}} |\widehat{h}_{nT,J}(y) - h_0(y)| = O_p\big(\tau_J\zeta_{nT,J}\sqrt{\ln N/N}\big) + O_p(J^{-\mu})$. Due to the use of concentration inequality for quadratic forms,[8] the sieve GMM estimator fails to attain such an optimal rate in terms of both bias and variance part.

3.2 Asymptotic normality

Now we study the asymptotic distribution of $(\widehat{\beta}_{nT}, \widehat{h}_{nT,J}(y))$ for a given $y \in \mathcal{R}_\mathcal{Y}$. Let

$$\Sigma_{nT}(y) \equiv \Gamma_{n,J}(y)(D_{nT}'\mathcal{W}_{nT}D_{nT})^{-1}D_{nT}'\mathcal{W}_{nT}\Omega_{nT}\mathcal{W}_{nT}D_{nT}(D_{nT}'\mathcal{W}_{nT}D_{nT})^{-1}\Gamma_{n,J}'(y), \tag{13}$$

where Ω_{nT} and D_{nT} are defined in (10) and (12), respectively, and

$$\Gamma_{n,J}(y) = \begin{pmatrix} I_{d_X} & \\ & k^J(y)' \end{pmatrix}.$$

[8] See the proof of Theorem 3.2(ii) in Section B of the Supplementary Material.

If we take $\mathcal{W}_{nT} = \Omega_{nT}^{-1}$ in (7), the corresponding estimator becomes optimal GMM (OGMM) estimator and the variance matrix (13) reduces to $\Sigma_{o,nT}(y) \equiv \Gamma_{n,J}(y)(D_{nT}'\Omega_{nT}^{-1}D_{nT})^{-1}\Gamma_{n,J}'(y)$.

Theorem 3.3. *Suppose that Assumptions 2.1-2.3 and 3.1-3.8 hold (where 3.5 is replaced by 3.5′). Then for a given finite $y \in \mathcal{R}_{\mathcal{Y}}$ such that $\|k^J(y)\| > 0$, the sieve GMM estimator $\widehat{\theta}_{nT} = (\widehat{\beta}_{nT}, \widehat{h}_{nT,J})$ derived from $\min_{\theta\in\Theta_n} g_{nT}'(\theta)\mathcal{W}_{nT}g_{nT}(\theta)$ satisfies*

$$\sqrt{N}\Sigma_{nT}^{-1/2}(y)\begin{pmatrix}\widehat{\beta}_{nT} - \beta_0 \\ \widehat{h}_{nT,J}(y) - h_0(y)\end{pmatrix} \xrightarrow{d} N(\mathbf{0}, I_{d_X+1}), \tag{14}$$

provided that $\sqrt{N}J^{-\mu} = o(\min\{\|k^J(y)\|, \underline{\nu}_J^{1/2}\})$. Also, the sieve OGMM estimator derived from $\min_{\theta\in\Theta_n} g_{nT}'(\theta)\Omega_{nT}^{-1}g_{nT}(\theta)$ satisfies (14) with $\Sigma_{nT}^{-1/2}(y)$ replaced by $\Sigma_{o,nT}^{-1/2}(y)$.

The proof of Theorem 3.3 is given in Section B in the Supplementary Material and employs the central limit theorem (CLT) for linear-quadratic forms in Kelejian and Prucha (2001). The requirements that $\sqrt{N}J^{-\mu} = o(\min\{\|k^J(y)\|, \underline{\nu}_J^{1/2}\})$ is an "undersmoothing" condition ensuring that the sieve approximation error does not affect the limiting distribution. We also note that apart from the diverging dimensions of D_{nT} and Ω_{nT}, the form of the variance matrix $\Sigma_{nT}(y)$ is the same as in the standard semiparametric literature such as Pakes and Pollard (1989) and Chen et al. (2005), etc.

Remark 3.2. *(Feasible sieve OGMM estimator) The optimal weighting matrix Ω_{nT}^{-1} involves the true parameters σ_0^2, μ_3 and μ_4, which can be consistently estimated using an initial GMM estimator $\widehat{\theta}_{nT}$. For example, we may obtain $\widehat{\theta}_{nT}$ from $\min_{\theta\in\Theta_n} d_g^{-1}g_{nT}'(\theta)g_{nT}(\theta)$. Denote $\Delta\widehat{\epsilon}_{it} \equiv \Delta y_{it} - \sum_{j=1}^n w_{ij}\Delta\widehat{h}_{nT,J}(y_{jt}) - \Delta x_{it}'\widehat{\beta}_{nT}$. Then σ_0^2, μ_3 and μ_4 can be estimated by $\widehat{\sigma}^2 = (1/(2N))\sum_{i=1}^n\sum_{t=2}^T \Delta\widehat{\epsilon}_{it}^2$, $\widehat{\mu}_3 = -(1/(6N))\sum_{i=1}^n\sum_{t=3}^T(\Delta\widehat{\epsilon}_{it} - \Delta\widehat{\epsilon}_{i,t-1})^3$ and $\widehat{\mu}_4 = (1/(2N))\sum_{i=1}^n\sum_{t=2}^T \Delta\widehat{\epsilon}_{it}^4 - 3\widehat{\sigma}^4$, respectively. Plugging these estimates to the formula of Ω_{nT} given in (10), we obtain a consistent estimate for Ω_{nT}, say $\widehat{\Omega}_{nT}$. We show in Section B in the Supplementary Material that under Assumptions 3.1-3.8, $\|\widehat{\Omega}_{nT} - \Omega_{nT}\| = o_p(1)$ and the feasible OGMM estimator derived from $\min_{\theta\in\Theta_n} g_{nT}'(\theta)\widehat{\Omega}_{nT}^{-1}g_{nT}(\theta)$ has the same asymptotic distribution as the infeasible OGMM estimator given in Theorem 3.3.*

Moreover, for the OGMM estimator, the best selection of P_{ln}'s in terms of efficiency may not be available due to the unknown form of the correlation between $\Delta\mathcal{E}_{nt}$ and $W_n\Delta K_{nt}$. In the linear case where $h_0(y) = \lambda_0 y$, Lee (2007) suggests the use of $G_n - tr(G_n)I_n/n$ as P_n in terms of efficiency under the normally distributed disturbances, where $G_n = W_n(I_n - \lambda_0 W_n)^{-1}$. Following this guidance, in practice, we can choose P_{ln} as $W_n - tr(W_n)I_n/n$, $W_n^2 - tr(W_n^2)I_n/n$, etc.

The asymptotic distribution of $\widehat{\beta}_{nT}$ and $\widehat{h}_{nT,J}(y)$ can be obtained separately from Theorem 3.3, as given in the following corollaries. Let $\mathscr{P}(A) = A(A'A)^{-1}A'$ for any matrix A with

full column rank. Define

$$\Sigma_{nT,X} \equiv D'_{nT,X}\mathcal{W}_{nT}^{1/2}\left(I_{d_g} - S_{nT}^{J}\right)\mathcal{W}_{nT}^{1/2}D_{nT,X}, \quad \Omega_{nT,X} \equiv \xi_{nT,X}\Omega_{nT}\xi'_{nT,X}, \tag{15}$$

where $S_{nT}^{J} \equiv \mathscr{P}(\mathcal{W}_{nT}^{1/2}D_{nT,J})$ and $\xi_{nT,X} \equiv D'_{nT,X}\mathcal{W}_{nT}^{1/2}(I_{d_g} - S_{nT}^{J})\mathcal{W}_{nT}^{1/2}$ with $D_{nT,X}$ and $D_{nT,J}$ given by (12).

Corollary 3.1. *Suppose that Assumptions 2.1-2.3 and 3.1-3.8 (where 3.5 is replaced by 3.5′) hold. If $N^{1/2}J^{-\mu} = o(1)$, then $\sqrt{N}(\widehat{\beta}_{nT}-\beta_0) \xrightarrow{d} N(0, \Sigma_X^{-1}\Omega_X\Sigma_X^{-1})$, where $\Sigma_X \equiv \lim_{n\to\infty}\Sigma_{nT,X}$ and $\Omega_X \equiv \lim_{n\to\infty}\Omega_{nT,X}$ are assumed to be positive definite.*

To derive the limiting distribution of $\widehat{\beta}_{nT}$, we require that J increases to infinity at a rate faster than $N^{1/(2\mu)}$. Along with Assumption 3.8, when $J = O(N^{\kappa})$ for some constant $0 < \kappa < \infty$, we need to choose κ such that $1/(2\mu) < \kappa$ and $N^{3\kappa-1}/\underline{\nu}_J^5 = o(1)$ simultaneously. Moreover, if the ill-posedness is not severe so that $1/\sqrt{\underline{\nu}_J} = O(J^r)$ for some $r > 0$, above conditions can be reduced to $1/(2\mu) < \kappa < 1/(10r+3)$.

Remark 3.3. *(Efficiency gains from quadratic moments) If we choose the weighting matrix $\mathcal{W}_{nT} = \widehat{\Omega}_{nT}^{-1}$ and assume that $\mu_3 = 0$, the asymptotic variance matrix of $\widehat{\beta}_{nT}$ (OGMME) can be reduced to $\Sigma_X^{-1} = \lim_{n\to\infty}\Sigma_{nT,X}^{-1}$ where*

$$\begin{aligned}\Sigma_{nT,X} = {} & \Psi'_{nT,LX}\Omega_{nT,22}^{-1}\Psi_{nT,LX} - \Psi'_{nT,LX}\Omega_{nT,22}^{-1}\Psi_{nT,LJ} \\ & \times\left(\Psi'_{nT,mJ}\Omega_{nT,11}^{-1}\Psi_{nT,mJ} + \Psi'_{nT,LJ}\Omega_{nT,22}^{-1}\Psi_{nT,LJ}\right)^{-1}\Psi'_{nT,LJ}\Omega_{nT,22}^{-1}\Psi_{nT,LX}\end{aligned}$$

with $\Psi_{nT,LX} = \Delta\mathbf{B}'_{n,T-1}\Delta\mathbf{X}_{n,T-1}/N$, $\Psi_{nT,LJ} = \mathrm{E}[\Delta\mathbf{B}'_{n,T-1}\mathbf{W}_{n,T-1}\Delta\mathbf{K}_{n,T-1}/N]$, $\Psi_{nT,mJ} = \mathrm{E}[(\mathbf{P}^s_{1n,T-1}\Delta\boldsymbol{\mathcal{E}}_{n,T-1}, \ldots, \mathbf{P}^s_{mn,T-1}\Delta\boldsymbol{\mathcal{E}}_{n,T-1})'\mathbf{W}_{n,T-1}\Delta\mathbf{K}_{n,T-1}/N]$, $\Omega_{nT,11}$ is the $m\times m$ principal sub-matrices of Ω_{nT} and $\Omega_{nT,22} = \sigma_0^2\mathcal{V}^B_{n,T-1}/N$ with $\mathcal{V}^B_{n,T-1}$ defined in (11).

In contrast, if we only use linear moments and choose $\mathcal{W}_{nT} = \Omega_{nT,22}^{-1}$, the asymptotic variance matrix of $\widehat{\beta}_{nT}$ (2SLSE) is $\widetilde{\Sigma}_X^{-1} = \lim_{n\to\infty}\widetilde{\Sigma}_{nT,X}^{-1}$ with

$$\begin{aligned}\widetilde{\Sigma}_{nT,X} = {} & \Psi'_{nT,LX}\Omega_{nT,22}^{-1}\Psi_{nT,LX} \\ & - \Psi'_{nT,LX}\Omega_{nT,22}^{-1}\Psi_{nT,LJ}\left(\Psi'_{nT,LJ}\Omega_{nT,22}^{-1}\Psi_{nT,LJ}\right)^{-1}\Psi'_{nT,LJ}\Omega_{nT,22}^{-1}\Psi_{nT,LX}.\end{aligned}$$

Since $\Psi'_{nT,mJ}\Omega_{nT,11}^{-1}\Psi_{nT,mJ}$ is positive semidefinite, we have $\Sigma_{nT,X} \succeq \widetilde{\Sigma}_{nT,X}$, which implies that the OGMME is more efficient than the 2SLSE due to the additional quadratic moments.

Finally, we examine the asymptotic distribution of the $\widehat{h}_{nT,J}(y)$. Let

$$\Sigma_{nT,h} \equiv D'_{nT,J}\mathcal{W}_{nT}^{1/2}(I_{d_g} - S_{nT}^{X})\mathcal{W}_{nT}^{1/2}D_{nT,J}, \quad \Omega_{nT,h} \equiv \xi_{nT,J}\Omega_{nT}\xi'_{nT,J}, \tag{16}$$

and $v_{nT}^2(y) \equiv k^J(y)'\Sigma_{nT,h}^{-1}\Omega_{nT,h}\Sigma_{nT,h}^{-1}k^J(y)$ for a given $y \in \mathcal{R}_\mathcal{Y}$, where $S_{nT}^X \equiv \mathscr{P}(\mathcal{W}_{nT}^{1/2}D_{nT,X})$ and $\xi_{nT,J} \equiv D_{nT,J}'\mathcal{W}_{nT}^{1/2}(I_{d_g} - S_{nT}^X)\mathcal{W}_{nT}^{1/2}$.

Corollary 3.2. *Suppose that Assumptions 2.1-3.8 hold (where 3.5 is replaced by 3.5′). If* $\sqrt{N}J^{-\mu}/\min\left\{\underline{\nu}_J^{1/2}, \|k^J(y)\|\right\} = o(1)$, *then for a given* $y \in \mathcal{R}_\mathcal{Y}$ *such that* $\|k^J(y)\| > 0$, *we have* $\sqrt{N}v_{nT}^{-1}(y)(\widehat{h}_{nT,J}(y) - h_0(y)) \xrightarrow{d} N(0,1)$.

Similar to Remark 3.3, we can show that the quadratic moments help reduce the asymptotic variance of $\widehat{h}_{nT,J}(y)$ for a given $y \in \mathcal{R}_\mathcal{Y}$. The details are omitted here.

Remark 3.4. *The asymptotic variance of the estimator* $\widehat{\beta}_{nT}$ *and* $\widehat{h}_{nT,J}(y)$ *can be consistently estimated by their sample analogs. Let* $\Delta\widehat{\mathcal{E}}_{nt} \equiv \Delta Y_{nt} - W_n\Delta\widehat{\mathbf{h}}_{nT,J}(Y_{nt}) - \Delta X_{nt}\widehat{\beta}_{nT}$, *where* $\widehat{\mathbf{h}}_{nT,J}(Y_{nt}) \equiv (\widehat{h}_{nT,J}(y_{1t}), \ldots, \widehat{h}_{nT,J}(y_{nt}))'$, *and let* $\widehat{D}_{nT,J} \equiv N^{-1}\sum_{t=2}^T(P_{1n}^s\Delta\widehat{\mathcal{E}}_{nt}, \ldots, P_{mn}^s\Delta\widehat{\mathcal{E}}_{nt}, \Delta B_{nt})'W_n\Delta K_{nt}$. *The variance of* $\widehat{\beta}_{nT}$ *can be estimated by* $\widehat{V}_{nT,X} \equiv \widehat{\Sigma}_{nT,X}^{-1}\widehat{\Omega}_{nT,X}\widehat{\Sigma}_{nT,X}^{-1}$, *where* $\widehat{\Sigma}_{nT,X}$ *and* $\widehat{\Omega}_{nT,X}$ *are given by (15) with* $D_{nT,J}$ *replaced by* $\widehat{D}_{nT,J}$ *and* Ω_{nT} *replaced by* $\widehat{\Omega}_{nT}$ *given in Remark 3.2. An estimate for* $v_{nT}^2(y)$, *say* $\widehat{v}_{nT}^2(y)$, *can be obtained similarly. We show in Section B of the Supplementary Material that* $\|\widehat{V}_{nT,X} - \Sigma_{nT,X}^{-1}\Omega_{nT,X}\Sigma_{nT,X}^{-1}\| = o_p(1)$ *and* $\widehat{v}_{nT}^2(y) = v_{nT}^2(y)(1 + o_p(1))$ *under the assumptions of Corollaries 3.1 and 3.2.*

4 Testing the linearity of the endogenous effect

In this section, we consider testing the functional form of the endogenous interaction effect h_0. In particular, it is important to test whether the widely used linear specification is appropriate. Since the time-invariant additive components in h_0 can be absorbed in the fixed effects, we consider the null hypothesis of the form

$$\mathbb{H}_0 : h_0(y) = \rho_0 y \text{ for some } \rho_0 \in \Xi \text{ and for all } y \in \mathcal{R}_\mathcal{Y}, \tag{17}$$

where Ξ is a compact subset of $\mathbb{R}$. The alternative hypothesis is the negation of $\mathbb{H}_0$. We first extend the test in Hoshino (2022) to the panel data setting, then propose alternative test statistics, which are more robust to the degree of ill-posedness.

Under the cross-sectional setting, Hoshino (2022) proposes a test based on comparing the restricted estimate under $\mathbb{H}_0$ and the unrestricted estimate under $\mathbb{H}_1$. We extend his test statistic to our panel data model as

$$\mathbf{T}_{nT} = \sum_{i=1}^n\sum_{t=1}^T\left(\widehat{h}_{nT,J}(y_{it}) - \overline{\widehat{h}}_{nT,J} - \widehat{\rho}_{nT}(y_{it} - \bar{y})\right)^2, \tag{18}$$

where $\overline{\widehat{h}}_{nT,J} \equiv (nT)^{-1}\sum_{i=1}^{n}\sum_{t=1}^{T}\widehat{h}_{nT,J}(y_{it})$ and $\bar{y} \equiv (nT)^{-1}\sum_{i=1}^{n}\sum_{t=1}^{T} y_{it}$ are for re-cencering purpose,[9] and $\widehat{\rho}_{nT}$ is a consistent estimator of ρ_0. For example, it could be the 2SLS estimator in Kelejian and Prucha (1998) when $\beta_0 \neq 0$ or the quasi-maximum likelihood estimator in Lee and Yu (2010). Clearly, $\mathbf{T}_{nT}$ will be close to zero under $\mathbb{H}_0$ and will be "large" otherwise. Let

$$\begin{aligned}\mathbb{B}_{nT} &= N^{-1}\mathrm{tr}(\Sigma_{nT,h}^{-1}\mathrm{E}(\widetilde{\mathbf{K}}_{nT}'\widetilde{\mathbf{K}}_{nT})),\\ s_{nT}^2 &= 2N^{-2}\mathrm{tr}(\Sigma_{nT,h}^{-1}\mathrm{E}(\widetilde{\mathbf{K}}_{nT}'\widetilde{\mathbf{K}}_{nT})\Sigma_{nT,h}^{-1}\mathrm{E}(\widetilde{\mathbf{K}}_{nT}'\widetilde{\mathbf{K}}_{nT})),\end{aligned}$$

where $\widetilde{\mathbf{K}}_{nT} \equiv (I_{nT} - (nT)^{-1}\mathbf{1}_{nT}\mathbf{1}_{nT}')\mathbf{K}_{nT}$ and $\mathbf{1}_{nT}$ is a $(nT)\times 1$ vector of ones. As is shown in Theorem 4.1, when $\widehat{h}_{nT,J}$ in (18) is the sieve OGMM estimator, $\mathbb{B}_{nT}$ and s_{nT}^2 will serve as the asymptotic bias and variance for $\mathbf{T}_{nT}$, respectively.

The test statistic $\mathbf{T}_{nT}$ requires estimation of h_0 under both the null and alternative. However, the sieve estimate of h_0 under the alternative may be variable if the degree of ill-posedness is severe. To avoid this problem, we propose two new tests under the GMM framework. We first nest the null models in the alternative by taking $k_1(y) = y$, as in Korolev (2019).[10] Then the endogenous variable matrix can be split as $K_{nt} = (Y_{nt}, \tilde{K}_{nt})$ and the sieve coefficients as $\gamma_0 = (\gamma_{0,1}, \gamma_{0,2}')'$, where $\tilde{K}_{nt} = (k_2(Y_{nt}), \ldots, k_J(Y_{nt}))$ is an $n\times(J-1)$ sub-matrix of additional series terms used to approximate the nonparametric SAR model and $\gamma_{0,2}$ is the corresponding sieve coefficients of length $J-1$. The alternative model (5) could be written as

$$\Delta Y_{nt} = \gamma_{0,1}W_n\Delta Y_{nt} + W_n\Delta\widetilde{K}_{nt}\gamma_{0,2} + \Delta X_{nt}\beta_0 + \Delta U_{nt}, \quad t = 2,\ldots,T. \tag{19}$$

If $\mathbb{H}_0$ is correct, then the additional series terms $\Delta\widetilde{K}_{nt}$ should not enter the model. Thus we can test $\mathbb{H}_0$ by way of the 'approximate' null $\mathbb{H}_0^{\mathrm{app}} : \gamma_{0,2} = 0$.

Testing the 'approximate' null can be viewed as testing the joint significance of high-order sieve terms, the number of which diverges to infinity. For this purpose, we propose Lagrangian multiplier (LM) and distance metric (DM) statistics as in Gupta (2018). For ease of presentation, we use the sieve coefficient as the argument of moment functions in this section. For a $(d_X + J)$-dimensional vector $\delta = (\beta', \gamma')'$, let $\Delta U_{nt}(\delta) = \Delta Y_{nt} - W_n\Delta K_{nt}\gamma - \Delta X_{nt}\beta$, $\Delta\mathbf{U}_{n,T-1}(\delta) = (\Delta U_{n2}'(\delta), \ldots, \Delta U_{nT}'(\delta))'$, and $g_{nT}(\delta) = N^{-1}(\mathbf{P}_{1n,T-1}\Delta\mathbf{U}_{n,T-1}(\delta), \ldots, \mathbf{P}_{mn,T-1}\Delta\mathbf{U}_{n,T-1}(\delta), \Delta\mathbf{B}_{n,T-1})'\Delta\mathbf{U}_{n,T-1}(\delta)$. The

[9] Due to the first difference, the intercept term in h_0 is eliminated and thus cannot be estimated. It is meaningful to compare $\Delta\widehat{h}_{nT,J}(y_{it})$ with $\widehat{\rho}_{nT}\Delta y_{it}$, or $\widehat{h}_{nT,J}(y_{it}) - \overline{\widehat{h}}_{nT,J}$ with $\widehat{\rho}_{nT}(y_{it} - \bar{y})$. We choose the second one to form the test statistic, but the simulation results (available upon request) show that the two tests have similar finite sample performances.

[10] In practice, one can add y to the basis functions and eliminate the linearly dependent columns.

restricted OGMM estimator $\bar{\delta}_{nT} = (\bar{\beta}'_{nT}, \bar{\gamma}_{nT,1}, \mathbf{0}'_{J-1})'$ can be obtained from the minimization of $g'_{nT}(\beta, \gamma_1, \mathbf{0}_{J-1})\widehat{\Omega}^{-1}_{nT} g_{nT}(\beta, \gamma_1, \mathbf{0}_{J-1})$, where $\widehat{\Omega}_{nT}$ is the consistent estimator of the variance matrix Ω_{nT} given in Remark 3.2. Let $G_{nT}(\delta) \equiv \frac{\partial g_{nT}(\delta)}{\partial \delta'} = -N^{-1}(\mathbf{P}^s_{1n,T-1}\Delta\mathbf{U}_{n,T-1}(\delta), \ldots, \mathbf{P}^s_{mn,T-1}\Delta\mathbf{U}_{n,T-1}(\delta), \Delta\mathbf{B}_{n,T-1})'\Delta\boldsymbol{\mathcal{X}}_{n,T-1}$ with $\Delta\boldsymbol{\mathcal{X}}_{n,T-1} \equiv (\Delta\mathbf{X}_{n,T-1}, \mathbf{W}_{n,T-1}\Delta\mathbf{K}_{n,T-1})$. The LM and DM test statistics are defined as

$$\mathbf{LM}_{nT} \equiv N g'_{nT}(\bar{\delta}_{nT})\widehat{\Omega}^{-1}_{nT} G_{nT}(\bar{\delta}_{nT})[G'_{nT}(\bar{\delta}_{nT})\widehat{\Omega}^{-1}_{nT} G_{nT}(\bar{\delta}_{nT})]^{-} G'_{nT}(\bar{\delta}_{nT})\widehat{\Omega}^{-1}_{nT} g_{nT}(\bar{\delta}_{nT}), \tag{20}$$

and

$$\mathbf{DM}_{nT} \equiv N[g'_{nT}(\bar{\delta}_{nT})\widehat{\Omega}^{-1}_{nT} g_{nT}(\bar{\delta}_{nT}) - g'_{nT}(\widehat{\delta}_{nT})\widehat{\Omega}^{-1}_{nT} g_{nT}(\widehat{\delta}_{nT})], \tag{21}$$

where $\widehat{\delta}_{nT}$ is the unrestricted optimal estimator of the coefficients δ_0. LM test measures how close the gradient of GMM criterion $g'_{nT}(\delta)\widehat{\Omega}^{-1}_{nT} g_{nT}(\delta)$ evaluated as the restricted estimator is to zero. It only requires the estimate for the null model. DM test is based on the difference between the GMM criterion evaluated at the restricted and unrestricted estimator.

The number of restrictions in the approximate null is $J-1$. For a fixed J, $\mathbf{LM}_{nT}$ and $\mathbf{DM}_{nT}$ have an asymptotic χ^2_{J-1} distribution under $\mathbb{H}_0$. Such a distribution has mean $J-1$ and variance $2(J-1)$, and it is a well-known fact that $(\chi^2_J - J)/\sqrt{2J} \xrightarrow{d} N(0,1)$, as $J \to \infty$. This motivates us to consider the following standardized quantity

$$\overline{\mathbf{LM}}_{nT} \equiv \frac{\mathbf{LM}_{nT} - (J-1)}{\sqrt{2(J-1)}}, \quad \overline{\mathbf{DM}}_{nT} \equiv \frac{\mathbf{DM}_{nT} - (J-1)}{\sqrt{2(J-1)}}. \tag{22}$$

Theorem 4.1 shows that such a normalization will yield a non-degenerate limiting distribution of LM and DM statistics.

Denote $D_{nT,1} \equiv \mathrm{E}[\Delta\mathbf{V}'_{n,T-1}(\Delta\mathbf{X}_{n,T-1}, \mathbf{W}_{n,T-1}\Delta\mathbf{Y}_{n,T-1})/N]$, where $\Delta\mathbf{V}_{n,T-1} = (\Delta V_{n2}, \ldots, \Delta V_{nT})'$ with $\Delta V_{nt} \equiv (P^s_{1n}\Delta\mathcal{E}_{nt}, \ldots, P^s_{mn}\Delta\mathcal{E}_{nt}, \Delta B_{nt})$. To establish the limiting distribution of the test statistics, we add the following assumptions.

Assumption 3.2′. *Assumption 3.2 holds. In addition,* $\sup_{n,T}\sup_{i,t}\mathrm{E}(\epsilon^{12}_{it}) < \infty$ *if* $vec_D(P_{ln}) \neq \mathbf{0}$ *for some* $1 \leq l \leq m$.

Assumption 3.7′. *Assumption 3.7 holds. In addition, there exist finite positive constants* $\underline{c}_D, \underline{c}_k$, *and* $\bar{c}_k$ *such that (i)* $\liminf_{n\to\infty}\sigma^2_{min}(D_{nT,1}) > \underline{c}_D$, *and (ii)* $\underline{c}_k < \liminf_{n\to\infty}\lambda_{min}\big(\mathrm{E}\big(\widetilde{\mathbf{K}}'_{nT}\widetilde{\mathbf{K}}_{nT}/N\big)\big) \leq \limsup_{n\to\infty}\lambda_{max}\big(\mathrm{E}\big(\widetilde{\mathbf{K}}'_{nT}\widetilde{\mathbf{K}}_{nT}/N\big)\big) < \bar{c}_k$.

Assumption 3.8′. *Assumption 3.8 holds. In addition,* $J^{1-\mu}/\sqrt{\underline{\nu}^5_J} = o(1)$.

Assumption 4.1. *Under* $\mathbb{H}_0$, $\widehat{\rho}_{nT}$ *is a* $\sqrt{N}$*-consistent estimator for* ρ_0.

Since the test statistics have a quadratic form, we assume the existence of higher moments of the errors as in Assumption 3.2′. Assumptions 3.7′(ii) and 4.1 are parallel to Assumptions

3.4′ and 3.8 in Hoshino (2022) and are used to derive the limiting distribution of $\mathbf{T}_{nT}$ in (18). Assumption 3.8′ requires that the ill-posedness is not severe.

In deriving the asymptotic null distribution of $\overline{\mathbf{LM}}_{nT}$ and $\overline{\mathbf{DM}}_{nT}$, we first show that they are asymptotically equivalent to $(2(J-1))^{-1/2}\big(g'_{nT}(\theta_0)\Omega_{nT}^{-1/2}\mathcal{V}_{nT}\Omega_{nT}^{-1/2}g_{nT}(\theta_0)-(J-1)\big)$, where $\mathcal{V}_{nT}$ is a idempotent matrix of rank $J-1$. This is a quadratic form of $g_{nT}(\theta_0)$, where $g_{nT}(\theta_0)$ itself is a linear-quadratic vector of diverging dimension. To our best knowledge, there is no existing theory for the asymptotic normality of such kind of quantity. We establish a new type of CLT for this, see Lemma A.9 in Section A of the Supplementary Material.

Theorem 4.1. *Suppose that Assumptions 2.1-2.3, 3.1-3.8, and 4.1 hold (where 3.2, 3.7 and 3.8 are replaced by 3.2′, 3.7′ and 3.8′, respectively). Under* $\mathbb{H}_0$,

(i) $\overline{\mathbf{LM}}_{nT} \xrightarrow{d} N(0,1)$ *and* $\overline{\mathbf{DM}}_{nT} \xrightarrow{d} N(0,1)$.

(ii) If the following conditions also hold: (a) $J^3/(\underline{\nu}_J^6 N) = o(1)$, *(b)* $N^{1/2}J^{1/2-\mu}/\sqrt{\underline{\nu}_J} = o(1)$, *and (c)* $\underline{\nu}_J^2 J \to \infty$ *as* $J \to \infty$, *then* $\overline{\mathbf{T}}_{nT} = s_{nT}^{-1}(\mathbf{T}_{nT} - \mathbb{B}_{nT}) \xrightarrow{d} N(0,1)$.

If $\mathbb{H}_0$ is not true, the standardized test statistics tend to deviate to a positive value. Theorem 4.1 implies that we can implement a one-sided test by comparing the value of the standardized test statistics $\overline{\mathbf{LM}}_{nT}$, $\overline{\mathbf{DM}}_{nT}$ and $\overline{\mathbf{T}}_{nT}$ with the upper α-percentile of $N(0,1)$.

Next we study the power properties of our tests. Consider the following sequence of local alternatives

$$\mathbb{H}_1(\alpha_{nT}) : h_0(y) = \rho_0 y + \alpha_{nT} r(y) \quad \text{for some } \rho_0 \in \Xi \text{ and for all } y \in \mathcal{R}_{\mathcal{Y}}, \tag{23}$$

where $r(\cdot)$ is an unknown nonlinear function defined on $\mathcal{R}_{\mathcal{Y}}$, and $\alpha_{nT} \to 0$ is a scalar that specifies the speed at which the local alternatives converge to the null. Under $\mathbb{H}_1(\alpha_{nT})$, α_{nT} affects the convergence rate of the restricted estimator $(\bar{\beta}'_{nT}, \bar{\gamma}_{nT,1})'$ to $(\beta'_0, \rho_0)'$.

Assumption 4.2. *The local shift* $r \in \Lambda_{c_1}^{\mu_1}(\mathcal{R}_{\mathcal{Y}}, \omega_1)$ *and* $\sup_{n,T}\sup_{i,t}\mathrm{E}|r(y_{it})|^{4\omega} < \infty$, *where* μ_1, ω_1 *and* ω *are the same as Assumption 3.1.*

Let $\mathscr{V}_{nT} \equiv \mathscr{P}(\Omega_{nT}^{-1/2}D_{nT}) - \mathscr{P}(\Omega_{nT}^{-1/2}D_{nT,1})$ and $\xi_{nT} \equiv \mathrm{E}\big[\sum_{t=2}^{T}\Delta V'_{nt}W_n\Delta\mathbf{r}(Y_{nt})/N\big]$, where $\mathbf{r}(Y_{nt}) = (r(y_{1t}), \ldots, r(y_{nt}))'$ and $\Delta\mathbf{r}(Y_{nt}) = \mathbf{r}(Y_{nt}) - \mathbf{r}(Y_{n,t-1})$.

Theorem 4.2. *Suppose that Assumptions 2.1-2.3, 3.1-3.8 and 4.2 hold (where 3.2, 3.7 and 3.8 are replaced by 3.2′, 3.7′ and 3.8′, respectively). Under* $\mathbb{H}_{1n}(\alpha_{nT})$ *with* $\alpha_{nT} = J^{1/4}N^{-1/2}$, $\overline{\mathbf{LM}}_{nT}$ *and* $\overline{\mathbf{DM}}_{nT}$ *converge in distribution to* $N(\Xi, 1)$ *with* $\Xi \equiv (1/\sqrt{2})\lim_{n\to\infty}\xi'_{nT}\Omega_{nT}^{-1/2}\mathscr{V}_{nT}\Omega_{nT}^{-1/2}\xi_{nT}$.

Theorem 4.2 states that our tests can distinguish local alternatives $\mathbb{H}_1(\alpha_{nT})$ at the rate $\alpha_{nT} = J^{1/4}N^{-1/2}$. This rate has also been found in the literature, see, e.g., de Jong and Bierens (1994), Gupta (2018) and Gupta and Qu (2022). Theorem 4.2 also implies that the

strength of both linear and quadratic moments could impact the local power of our tests. To see this, note that ξ_{nT} measures the correlation between $(P^s_{1n}\Delta\mathcal{E}_{nt}, \ldots, P^s_{mn}\Delta\mathcal{E}_{nt}, \Delta B_{nt})$ and the endogenous local shift $W_n\Delta\mathbf{r}(Y_{nt})$. When the IVs are weak, the magnitude of $\|\xi_{nT}\|$ will be small, leading to a small local shift Ξ. This makes distinguishing the null and local alternatives more difficult. Under the extreme case that $\beta_0 = 0$, the tests based on only linear moments will have no local power since $\mathrm{E}(\Delta B'_{nt}W_n\Delta\mathbf{r}(Y_{nt})) = 0$. In contrast, the tests will still have nontrivial local power when quadratic moments are added since $\Delta\mathcal{E}_{nt}$ is always correlated with the endogenous local shift in some way.

5 Simulation

5.1 Sieve estimation

In this subsection, we examine the finite sample performance of our sieve GMM estimator. We generate samples from the following three data-generating processes (DGPs):

$$y_{it} = \sum_{j=1}^{n} w_{ij}h_0(y_{jt}) + x^{(1)}_{it}\beta_{01} + x^{(2)}_{it}\beta_{02} + c_{i0} + \epsilon_{it}, \quad i = 1, \ldots, n,\ t = 1, \ldots, T,$$

where

$$\begin{aligned}
&\text{DGP 1: } h_0(y) = \cos(0.8y),\\
&\text{DGP 2: } h_0(y) = 0.9\ln(|y-1|+1)\mathrm{sgn}(y-1),\\
&\text{DGP 3: } h_0(y) = 0.5y,
\end{aligned}$$

with $x^{(1)}_{it} \overset{i.i.d.}{\sim} N(0,4)$, $x^{(2)}_{it} \overset{i.i.d.}{\sim} \text{Uniform}[-1,1]$ and $\epsilon_{it} \overset{i.i.d.}{\sim} N(0, 0.8^2)$ for all DGPs. These DGPs are similar to Hoshino (2022) except that the intercept term in h_0 is dropped. We consider two sets of β coefficients: (i) $\beta_0 = (1,1)$ and (ii) $\beta_0 = (0,0)$. We generate the individual effect c_{i0}'s as $c_{i0} = 0.2(x^{(1)}_{i\cdot} + x^{(2)}_{i\cdot}) + u_i$, where $x^{(j)}_{i\cdot} = T^{-1}\sum_{t=1}^{T} x^{(j)}_{it}$ for $j = 1, 2$, and $\{u_i, i \geq 1\}$ is a sequence of i.i.d. $N(0, 0.5^2)$ random variables. The weights matrix W_n is generated by the Rook contiguity with row normalization. We randomly allocate n units on the lattice of $20 \times l$, such that $20l = n$, in which we consider two sample sizes: $n \in \{100, 400\}$ and $T \in \{5, 10\}$. We generate the dependent variables $\{Y_{nt}, 1 \leq t \leq T\}$ by contraction mapping iterations as in Hoshino (2022). For the choice of basis functions k and q, we use piece-wise cubic splines for both. The linear IV is defined by $q(z_{it}) = (\tilde{q}(x^{(1)}_{it})', \tilde{q}(x^{(2)}_{it})')$ where $\tilde{q}(\cdot)$ is a vector of the univariate B-spline functions. For quadratic moments, we use $P_{ln} = W^l_n - (\mathrm{tr}(W^l_n)/n)I_n$ for $l = 1, \ldots, m$. Let J and $\tilde{L}$ be the order of

sieve expansion for the univariate B-spline bases k and $\tilde{q}$, respectively.[11] To evaluate how our sieve estimator is sensitive to the choice of J, $\tilde{L}$ and the number of quadratic moments m, we consider three pairs of values for each n and β_0: (i) when $\beta_0 = (1,1)$, $(J,\tilde{L},m) \in \{(5,5,5),(5,6,6),(6,6,6)\}$ for $n = 100$, and $(J,\tilde{L},m) \in \{(6,6,6),(6,7,7),(7,7,7)\}$ for $n = 400$; (ii) when $\beta_0 = (0,0)$, $(J,\tilde{L},m) \in \{(4,4,4),(4,5,5),(5,5,5)\}$ for $n = 100$ and $(J,\tilde{L},m) \in \{(5,5,5),(5,6,6),(6,6,6)\}$ for $n = 400$.

The number of repetitions is $1,000$ for each set-up in the Monte Carlo experiment. In each case, we report the estimated bias, the root mean square errors (RMSE), and the coverage rate of the 95% confidence interval (CR95) for 2SLS, GMM, and OGMM estimator of (β_{01},β_{02}). The sieve GMM estimator is obtained by choosing $\mathcal{Q}_{nT}(\theta) = d_g^{-1} g_{nT}'(\theta) g_{nT}(\theta)$ in (7) and the OGMM estimator is obtained by choosing $\mathcal{Q}_{nT}(\theta) = d_g^{-1} g_{nT}'(\theta) \widehat{\Omega}_{nT}^{-1} g_{nT}(\theta)$ with $\widehat{\Omega}_{nT}$ given in Remark 3.2. Following Hoshino (2022), we evaluate the performance of the sieve estimator $\widehat{h}_{nT,J}(y)$ using the integrated squared bias (ISB) and integrated mean squared error (IMSE):

$$\begin{aligned} \text{ISB:} &\int_{y_{2.5}}^{y_{97.5}} \left[\frac{1}{1000} \sum_{r=1}^{1000} \left(\widehat{h}_{nT,J}^{(r)}(y) - h_0(y) - c_{nT}^{(r)} \right) \right]^2 dy, \\ \text{IMSE:} &\int_{y_{2.5}}^{y_{97.5}} \left[\frac{1}{1000} \sum_{r=1}^{1000} \left(\widehat{h}_{nT,J}^{(r)}(y) - h_0(y) - c_{nT}^{(r)} \right)^2 \right] dy, \end{aligned} \tag{24}$$

where $\widehat{h}_{nT,J}^{(r)}(y)$ is the estimate obtained from the rth replicated data, $c_{nT}^{(r)} = (y_{97.5}^{(r)} - y_{2.5}^{(r)})^{-1} \int_{y_{2.5}^{(r)}}^{y_{97.5}^{(r)}} \left(\widehat{h}_{nT,J}^{(r)}(y) - h_0(y) \right) dy$ with $y_{2.5}^{(r)}$ and $y_{97.5}^{(r)}$ being the 2.5% and 97.5% empirical quantiles of y_{it}'s in the rth repetition, and $y_{2.5}$ and $y_{97.5}$ being the empirical quantiles averaged over all repetitions. Here $c_{nT}^{(r)}$ is the amount of vertical translation used to make $\widehat{h}_{nT,J}(y)$ and $h_0(y)$ comparable.

The result for estimation of h_0 when $\beta_0 = (1,1)$ is presented in Table 1. We summarize our findings as follows. First, in DGP 1 where the function h_0 is highly nonlinear, there is a bias-variance trade-off with respect to J for all sieve estimators. The ISB values significantly decrease when J grows, but a larger J also leads to a larger variance. Additionally, the sieve 2SLS estimator is more sensitive to the choice of J than the sieve GMM and OGMM estimators. For example, in DGP 1, when the sample size is $(n,T) = (400,10)$, changing J from 6 to 7 results in a nearly twofold increase in the IMSE of 2SLS estimator, while this has a relatively minor impact on the variance of GMM estimators. Second, in DGPs 2 and 3 where the function h_0 can be well approximated by a linear function, a moderate number of J seems to be a better choice than a larger J in terms of IMSE. Although the function h_0 in DGP 2 is not as smooth as the other DGPs (the derivative of $h_0(y)$ has a kink at $y = 1$),

[11] The number of linear moments equals $d_X(1+\tilde{L})$.

Table 1: Sieve 2SLS and GMM estimation of h_0

				DGP 1		DGP 2		DGP 3	
J	$\tilde{L}$	m	Method	ISB	IMSE	ISB	IMSE	ISB	IMSE
				$n = 100,\ T = 5$					
5	5	5	2SLS	0.1220	1.3636	0.0421	1.4236	0.0007	1.5459
			GMM	0.1439	1.2367	0.0489	1.1154	0.0047	1.2221
			OGMM	0.1309	0.9606	0.0459	0.8733	0.0031	1.0069
5	6	6	2SLS	0.1189	1.1974	0.0426	1.0784	0.0012	1.2741
			GMM	0.1425	1.1843	0.0478	1.0439	0.0039	1.1806
			OGMM	0.1284	0.9310	0.0469	0.8171	0.0027	0.9323
6	6	6	2SLS	0.0463	2.0521	0.0184	2.2901	0.0023	2.6750
			GMM	0.0393	1.3566	0.0201	1.3376	0.0059	1.5081
			OGMM	0.0422	1.1785	0.0159	1.2138	0.0042	1.4015
				$n = 100,\ T = 10$					
5	5	5	2SLS	0.1307	0.5987	0.0416	0.5721	0.0009	0.7635
			GMM	0.1498	0.6087	0.0461	0.5302	0.0012	0.5949
			OGMM	0.1348	0.5021	0.0418	0.4457	0.0008	0.5018
5	6	6	2SLS	0.1273	0.6083	0.0419	0.5092	0.0005	0.6358
			GMM	0.1432	0.5872	0.0469	0.5014	0.0024	0.5941
			OGMM	0.1323	0.5095	0.0425	0.4237	0.0004	0.4821
6	6	6	2SLS	0.0455	1.0164	0.0140	1.0554	0.0004	1.5840
			GMM	0.0426	0.6397	0.0181	0.6399	0.0020	0.7694
			OGMM	0.0505	0.6038	0.0141	0.6331	0.0007	0.7848
				$n = 400,\ T = 5$					
6	6	6	2SLS	0.0815	0.5820	0.0134	0.5730	0.0002	0.7977
			GMM	0.0611	0.4446	0.0146	0.4182	0.0020	0.4961
			OGMM	0.0747	0.4161	0.0137	0.3855	0.0009	0.4698
6	7	7	2SLS	0.0903	0.5423	0.0134	0.5065	0.0000	0.7052
			GMM	0.0631	0.4311	0.0149	0.4162	0.0023	0.4835
			OGMM	0.0804	0.4040	0.0137	0.3695	0.0012	0.4508
7	7	7	2SLS	0.0534	1.0991	0.0027	1.2651	0.0001	1.6589
			GMM	0.0453	0.4985	0.0037	0.5028	0.0026	0.5652
			OGMM	0.0526	0.5252	0.0037	0.5681	0.0011	0.6269
				$n = 400,\ T = 10$					
6	6	6	2SLS	0.0807	0.2826	0.0130	0.2239	0.0009	0.3452
			GMM	0.0549	0.2518	0.0136	0.2257	0.0006	0.2311
			OGMM	0.0775	0.2389	0.0128	0.1893	0.0006	0.2396
6	7	7	2SLS	0.0875	0.2768	0.0129	0.2115	0.0009	0.3177
			GMM	0.0538	0.2507	0.0140	0.2273	0.0007	0.2341
			OGMM	0.0822	0.2386	0.0128	0.1817	0.0006	0.2291
7	7	7	2SLS	0.0530	0.5362	0.0032	0.5597	0.0018	0.8305
			GMM	0.0405	0.2806	0.0030	0.2654	0.0009	0.2879
			OGMM	0.0520	0.3106	0.0028	0.2900	0.0014	0.3548

Note: True parameters $\beta_{01} = \beta_{02} = 1$.

Table 2: 2SLSE, GMME and OGMME for β_{01}

				2SLSE			GMME			OGMME		
DGP	J	$\tilde{L}$	m	Bias	RMSE	CR95	Bias	RMSE	CR95	Bias	RMSE	CR95
							$n = 100, T = 5$					
1	5	5	5	-0.0014	0.0254	0.947	-0.0008	0.0255	0.957	-0.0020	0.0255	0.932
	5	6	6	-0.0014	0.0252	0.941	-0.0007	0.0254	0.956	-0.0020	0.0256	0.933
	6	6	6	-0.0014	0.0259	0.949	-0.0007	0.0257	0.962	-0.0020	0.0257	0.936
2	5	5	5	0.0001	0.0256	0.956	0.0001	0.0257	0.958	-0.0006	0.0253	0.947
	5	6	6	0.0001	0.0254	0.952	-0.0001	0.0257	0.959	-0.0006	0.0253	0.943
	6	6	6	0.0001	0.0262	0.957	-0.0001	0.0260	0.962	-0.0008	0.0257	0.944
3	5	5	5	-0.0013	0.0255	0.944	-0.0021	0.0265	0.963	-0.0024	0.0257	0.938
	5	6	6	-0.0014	0.0252	0.942	-0.0021	0.0265	0.961	-0.0025	0.0256	0.940
	6	6	6	-0.0014	0.0261	0.955	-0.0024	0.0268	0.962	-0.0025	0.0259	0.942
							$n = 100, T = 10$					
1	5	5	5	-0.0003	0.0167	0.949	-0.0004	0.0168	0.959	-0.0007	0.0169	0.943
	5	6	6	-0.0003	0.0167	0.951	-0.0003	0.0168	0.957	-0.0008	0.0168	0.941
	6	6	6	-0.0003	0.0168	0.951	-0.0001	0.0167	0.970	-0.0007	0.0168	0.941
2	5	5	5	0.0004	0.0163	0.958	0.0004	0.0168	0.967	0.0001	0.0164	0.953
	5	6	6	0.0005	0.0164	0.957	0.0004	0.0168	0.968	0.0001	0.0164	0.953
	6	6	6	0.0007	0.0167	0.960	0.0004	0.0168	0.963	0.0002	0.0165	0.950
3	5	5	5	-0.0016	0.0168	0.960	-0.0016	0.0170	0.966	-0.0021	0.0166	0.953
	5	6	6	-0.0017	0.0166	0.958	-0.0017	0.0170	0.963	-0.0021	0.0166	0.950
	6	6	6	-0.0017	0.0171	0.963	-0.0019	0.0171	0.965	-0.0023	0.0166	0.951
							$n = 400, T = 5$					
1	6	6	6	0.0004	0.0119	0.958	0.0004	0.0119	0.966	0.0002	0.0119	0.955
	6	7	7	0.0004	0.0119	0.957	0.0005	0.0120	0.967	0.0002	0.0120	0.953
	7	7	7	0.0003	0.0121	0.965	0.0004	0.0120	0.968	0.0001	0.0120	0.955
2	6	6	6	0.0004	0.0117	0.950	0.0002	0.0119	0.965	0.0001	0.0116	0.949
	6	7	7	0.0003	0.0116	0.949	0.0002	0.0119	0.962	0.0001	0.0116	0.947
	7	7	7	0.0004	0.0117	0.954	0.0002	0.0119	0.973	0.0001	0.0116	0.952
3	6	6	6	-0.0002	0.0126	0.946	-0.0006	0.0129	0.958	-0.0006	0.0126	0.944
	6	7	7	-0.0003	0.0126	0.945	-0.0006	0.0129	0.956	-0.0007	0.0126	0.943
	7	7	7	-0.0003	0.0130	0.951	-0.0006	0.0128	0.966	-0.0007	0.0127	0.948
							$n = 400, T = 10$					
1	6	6	6	-0.0001	0.0083	0.939	-0.0001	0.0084	0.953	-0.0002	0.0084	0.933
	6	7	7	-0.0001	0.0083	0.937	0.0001	0.0084	0.950	-0.0002	0.0084	0.932
	7	7	7	-0.0001	0.0084	0.940	-0.0001	0.0084	0.953	-0.0002	0.0084	0.933
2	6	6	6	-0.0003	0.0085	0.935	-0.0003	0.0086	0.952	-0.0004	0.0084	0.929
	6	7	7	-0.0003	0.0085	0.934	-0.0003	0.0086	0.948	-0.0004	0.0085	0.929
	7	7	7	-0.0003	0.0086	0.938	-0.0003	0.0086	0.948	-0.0004	0.0085	0.931
3	6	6	6	0.0001	0.0082	0.957	-0.0002	0.0082	0.971	-0.0001	0.0081	0.954
	6	7	7	0.0001	0.0082	0.957	-0.0003	0.0082	0.973	-0.0001	0.0081	0.954
	7	7	7	0.0001	0.0083	0.962	-0.0003	0.0082	0.975	-0.0002	0.0081	0.957

Note: True parameters $\beta_{01} = \beta_{02} = 1$. CR95 stands for the coverage rate of the 95% confidence interval computed based on Corollary 3.1.

Table 3: Sieve 2SLS and GMM estimation of h_0 when X_{nt} is irrelevant

				DGP 1		DGP 2		DGP 3	
J	$\tilde{L}$	m	Method	ISB	IMSE	ISB	IMSE	ISB	IMSE
				$n = 100,\ T = 5$					
4	4	4	2SLS	0.2118	4.2967	0.3912	4.7165	0.5251	4.4727
			GMM	0.0164	0.4525	0.0044	0.4882	0.0093	0.6831
			OGMM	0.0192	0.4715	0.0076	0.4937	0.0099	0.5359
4	5	5	2SLS	0.2357	2.6997	0.4179	3.2825	0.5436	2.8340
			GMM	0.0184	0.4686	0.0035	0.4957	0.0106	0.6721
			OGMM	0.0207	0.4423	0.0083	0.4809	0.0116	0.5114
5	5	5	2SLS	0.2159	4.1655	0.4193	4.7776	0.5727	4.3418
			GMM	0.0168	0.6225	0.0073	0.6811	0.0079	0.7089
			OGMM	0.0226	0.6235	0.0143	0.6902	0.0122	0.6742
				$n = 100,\ T = 10$					
4	4	4	2SLS	0.2444	4.2356	0.4745	4.4806	0.5876	3.9815
			GMM	0.0110	0.2302	0.0016	0.3040	0.0094	0.4500
			OGMM	0.0109	0.2676	0.0013	0.2928	0.0038	0.3282
4	5	5	2SLS	0.2327	2.7671	0.4601	2.8673	0.6070	2.7745
			GMM	0.0135	0.2484	0.0018	0.2674	0.0069	0.3810
			OGMM	0.0135	0.2678	0.0017	0.2828	0.0040	0.3239
5	5	5	2SLS	0.2223	4.0951	0.4887	4.3850	0.5814	4.1294
			GMM	0.0121	0.3301	0.0012	0.3206	0.0083	0.5000
			OGMM	0.0136	0.3468	0.0020	0.3784	0.0033	0.4190
				$n = 400,\ T = 5$					
5	5	5	2SLS	0.3245	4.2994	0.4306	4.3911	0.7196	4.5000
			GMM	0.0155	0.2865	0.0024	0.2767	0.0035	0.3071
			OGMM	0.0160	0.2691	0.0020	0.2464	0.0023	0.2842
5	6	6	2SLS	0.2878	3.0818	0.4013	3.2139	0.7321	3.3254
			GMM	0.0132	0.2392	0.0019	0.2063	0.0034	0.2976
			OGMM	0.0164	0.2659	0.0021	0.2505	0.0024	0.2855
6	6	6	2SLS	0.3304	4.2344	0.4245	4.2390	0.7493	4.5596
			GMM	0.0144	0.2403	0.0023	0.2521	0.0049	0.4168
			OGMM	0.0169	0.3085	0.0023	0.2909	0.0030	0.3350
				$n = 400,\ T = 10$					
5	5	5	2SLS	0.3255	4.4217	0.3759	4.2506	0.7229	4.4582
			GMM	0.0175	0.1916	0.0037	0.1604	0.0049	0.2896
			OGMM	0.0176	0.1683	0.0045	0.1585	0.0010	0.1720
5	6	6	2SLS	0.2716	3.0287	0.4292	2.9766	0.7487	3.2943
			GMM	0.0167	0.1934	0.0033	0.1413	0.0074	0.3016
			OGMM	0.0173	0.1681	0.0035	0.1571	0.0012	0.1765
6	6	6	2SLS	0.2682	4.0249	0.4158	4.3374	0.7720	4.6846
			GMM	0.0170	0.1637	0.0038	0.1855	0.0041	0.2792
			OGMM	0.0171	0.1927	0.0044	0.1809	0.0012	0.1966

Note: True parameters: $\beta_{01} = \beta_{02} = 0$.

Table 4: 2SLSE, GMME and OGMME for β_{01} when X_{nt} is irrelevant

				2SLSE			GMME			OGMME		
DGP	J	$\tilde{L}$	m	Bias	RMSE	CR95	Bias	RMSE	CR95	Bias	RMSE	CR95
							$n = 100, T = 5$					
1	4	4	4	-0.0001	0.0284	0.981	-0.0003	0.0246	0.968	-0.0001	0.0247	0.951
	4	5	5	0.0001	0.0263	0.971	-0.0002	0.0247	0.963	0.0000	0.0245	0.947
	5	5	5	-0.0004	0.0281	0.981	0.0001	0.0249	0.970	0.0001	0.0246	0.957
2	4	4	4	0.0007	0.0276	0.984	0.0009	0.0240	0.977	0.0010	0.0240	0.958
	4	5	5	0.0010	0.0259	0.977	0.0011	0.0240	0.974	0.0011	0.0242	0.954
	5	5	5	0.0009	0.0285	0.986	0.0010	0.0241	0.976	0.0010	0.0244	0.964
3	4	4	4	-0.0004	0.0273	0.985	-0.0009	0.0242	0.963	-0.0009	0.0240	0.952
	4	5	5	-0.0004	0.0260	0.977	-0.0010	0.0243	0.962	-0.0006	0.0241	0.949
	5	5	5	-0.0004	0.0277	0.985	-0.0009	0.0243	0.969	-0.0007	0.0246	0.956
							$n = 100, T = 10$					
1	4	4	4	-0.0010	0.0190	0.987	-0.0010	0.0161	0.969	-0.0010	0.0160	0.968
	4	5	5	-0.0009	0.0178	0.983	-0.0009	0.0160	0.970	-0.0009	0.0161	0.966
	5	5	5	-0.0008	0.0194	0.990	-0.0009	0.0161	0.974	-0.0009	0.0162	0.969
2	4	4	4	0.0006	0.0190	0.989	0.0004	0.0164	0.969	0.0005	0.0166	0.959
	4	5	5	0.0002	0.0180	0.972	0.0004	0.0164	0.965	0.0005	0.0166	0.957
	5	5	5	0.0001	0.0190	0.980	0.0003	0.0164	0.974	0.0005	0.0166	0.960
3	4	4	4	0.0003	0.0200	0.986	0.0002	0.0164	0.975	0.0001	0.0166	0.964
	4	5	5	0.0003	0.0184	0.980	0.0002	0.0164	0.970	0.0002	0.0166	0.958
	5	5	5	0.0007	0.0192	0.987	0.0003	0.0167	0.981	0.0003	0.0166	0.963
							$n = 400, T = 5$					
1	5	5	5	-0.0003	0.0144	0.981	-0.0001	0.0120	0.975	-0.0002	0.0120	0.965
	5	6	6	-0.0005	0.0134	0.981	-0.0001	0.0120	0.976	-0.0001	0.0120	0.958
	6	6	6	-0.0004	0.0140	0.987	-0.0001	0.0121	0.981	-0.0001	0.0120	0.963
2	5	5	5	0.0007	0.0140	0.986	0.0006	0.0120	0.975	0.0006	0.0121	0.958
	5	6	6	0.0006	0.0132	0.981	0.0006	0.0120	0.970	0.0006	0.0121	0.957
	6	6	6	0.0005	0.0142	0.992	0.0006	0.0120	0.979	0.0006	0.0121	0.964
3	5	5	5	0.0004	0.0134	0.987	0.0005	0.0113	0.973	0.0005	0.0115	0.969
	5	6	6	0.0004	0.0125	0.985	0.0004	0.0114	0.973	0.0005	0.0115	0.966
	6	6	6	0.0002	0.0135	0.986	0.0004	0.0114	0.982	0.0005	0.0116	0.973
							$n = 400, T = 10$					
1	5	5	5	0.0001	0.0101	0.986	0.0001	0.0082	0.973	0.0001	0.0083	0.958
	5	6	6	0.0001	0.0094	0.978	0.0001	0.0082	0.969	0.0001	0.0083	0.955
	6	6	6	0.0001	0.0100	0.987	0.0001	0.0082	0.977	0.0001	0.0083	0.960
2	5	5	5	-0.0005	0.0095	0.981	-0.0005	0.0080	0.969	-0.0005	0.0080	0.952
	5	6	6	-0.0004	0.0090	0.979	-0.0005	0.0080	0.968	-0.0004	0.0080	0.955
	6	6	6	-0.0004	0.0096	0.988	-0.0005	0.0080	0.973	-0.0005	0.0080	0.959
3	5	5	5	-0.0002	0.0098	0.983	-0.0002	0.0082	0.976	-0.0003	0.0083	0.958
	5	6	6	-0.0002	0.0091	0.976	-0.0002	0.0083	0.968	-0.0003	0.0082	0.950
	6	6	6	-0.0001	0.0095	0.983	-0.0002	0.0082	0.974	-0.0002	0.0083	0.952

Note: True parameters $\beta_{01} = \beta_{02} = 0$. CR95 stands for the coverage rate of the 95% confidence interval computed based on Corollary 3.1.

all sieve estimators work satisfactorily. Finally, The choice of $\tilde{L}$ and m has relatively minor impacts on the performance of the estimator in all DGPs. This finding is consistent with that in Hoshino (2022). The result of estimating β_{01} when $\beta_0 = (1, 1)$ is presented in Table 2. We see that all sieve estimators perform quite well. The bias and RMSE are satisfactorily small, and the coverage rate of the 95% confidence interval is close to the nominal level. Similar pattern is found in estimation results of β_{02}, see Table D.1 in the Supplementary Material for details.

Table 3 reports the results for the estimation of h_0 when $\beta_0 = (0, 0)$. The exogenous variables $X_{nt}^{(1)}$ and $X_{nt}^{(2)}$ are irrelevant in this case. We see that the ISB and IMSE of the sieve 2SLS estimator are much larger than that of the sieve GMM and OGMM estimators for all DGPs, as expected in theory. The performance of the sieve 2SLS estimator does not show any improvement when the sample size grows. In contrast, the sieve GMM method gives satisfactory estimates even with a moderate sample size. This indicates that the performance of the sieve GMM estimator is more robust to the strength of IVs generated by exogenous variables. The results of estimating β_{01} and β_{02} are presented in Table 4 and Table D.2 in the Supplementary Material, respectively. We see that when β_0 reduces to zero, the performance of the sieve 2SLS estimator for β is still acceptable. The bias of 2SLS and GMM estimators are similar, while the sieve OGMM estimator outperforms the 2SLS estimator in terms of RMSE and CR95 in all DGPs. Comparing this with Tables 1, 2 and Table D.1 in the Supplementary Material, we conclude that the difference between 2SLS and GMM estimators occurs mainly for the estimation of h_0 but not for the estimation of β_0.

5.2 Testing for the linearity of h_0

Next, we examine the finite sample performance of the proposed LM and DM tests. We also compare it with the test statistic $\mathbf{T}_{nT}$. The settings for basis functions and moment functions are the same as in the above subsection.

Table 5 presents the rejection frequency of three test statistics based on 1,000 Monte Carlo repetitions when $\beta_0 = (1, 1)$. We see that when the function h_0 is highly nonlinear, as in DGP 1, all three tests exhibit good power properties even when the sample size is relatively small. However, when h_0 can be well approximated by a linear function, as in DGP 2, the power of the tests significantly decreases, but it can be improved by increasing the sample size. In addition, for DGPs 1 and 2, the choice of J has an obvious impact on the power performance for the test statistic $\mathbf{T}_{nT}$. For example, when $n = 400$ and $T = 10$, the empirical power for $\mathbf{T}_{nT}$ test shrinks by half when J goes from 6 to 7 for DGP 2. As Hoshino (2022) remark, the reason behind is that when too many basis terms are used to approximate h_0, the increase in the estimation variance will reduce the power of the test considerably. In contrast, increasing J has a relatively minor impact on the power of our

LM and DM tests. In terms of empirical power, LM and DM tests dominate the $\mathbf{T}_{nT}$ test in all the alternatives we consider. Finally, the result for DGP 3 indicates that the empirical sizes for three tests are all close to the nominal levels when (n, T) is large.

6 Empirical application

6.1 China EKC

As the first empirical illustration, we revisit the EKC in China using the nonlinear SAR model with a nonparametric endogenous effect. We use PM2.5 concentrations as the pollution index. The GDP per capita and some other economic variables are also collected for 284 cities in China from 2004 to 2015. We analyze the relationship between PM2.5 concentrations and GDP per capita using the following SAR panel data model:

$$p_{it} = \sum_{j=1}^{n} w_{ij} h_0(p_{jt}) + x_{it}'\beta_0 + Y_{it}'\gamma_0 + c_i + \mu_t + \epsilon_{it}, \tag{25}$$

where p_{it} is the level of PM2.5 concentrations of ith city at year t, w_{ij} is the row-normalized spatial weight[12] between city i and city j. The $x_{it}'s$ are control variables, including population, secondary industry as a percentage of GDP (Sec_acc_gdp), number of buses, number of taxis, and road ratio of the city. Apart from the shape of reaction function h_0, we are also interested in the variables $Y_{it} = (y_{it}, y_{it}^2)$, where y_{it} is the GDP per capita (GDPPC). The estimation result of the corresponding coefficient γ_0 indicates the relationship between the environmental indicator (PM2.5) and income. Both individual fixed effects and time fixed effects are specified in model (25). The time fixed effects μ_t are eliminated by demeaning from the cross-sectional average and the individual fixed effects c_i are eliminated by the first difference.

We estimate the model (25) using two approaches: the parametric GMM procedure by Lee (2007) assuming that $h_0(y) = \rho_0 y$ and the sieve GMM method. The instrument z_{it} is taken as all elements of $(x_{it}', y_{it})'$. We use the cubic B-spline basis function with one internal knot for both k and q. This yields $J = \tilde{L} = 4$. For the quadratic moments, we choose $P_{ln} = W_n^l - \frac{1}{n-1}\text{tr}(W_n^l J_n)J_n$ for $l = 1, \ldots, 4$, where $J_n = I_n - (1/n)\mathbf{1}_n'\mathbf{1}_n$.[13]

The value of the standardized test statistics $\overline{\mathbf{LM}}_{nT}$ and $\overline{\mathbf{T}}_{nT}$ are 0.998 and -0.707, re-

[12]The element w_{ij} is decaying functions of the distance between city i and j. Specifically, $w_{ij} = e^{-\alpha \times d_{ij}} 1(d_{ij} < \bar{\Delta})$, where $\alpha > 0$, $100 \times d_{ij}$ is the distance between city i and j measured in kilometers, and $\bar{\Delta}$ is the threshold distance, i.e., if $d_{ij} > \bar{\Delta}$, then $w_{ij} = 0$. A larger α_d denotes a more rapid decline in the size of the weights as the distance increases. We set $\bar{\Delta} = 2.34$ and $\alpha = -1.5$ as in Chang et al. (2021).

[13]In presence of time fixed effects, we need $\text{tr}(P_{ln}J_n) = 0$ for $l = 1, \ldots, m$ so that the quadratic orthogonal conditions hold, see Lee and Yu (2014).

Table 5: Rejection frequency

				LM			**DM**			**T**		
DGP	J	$\tilde{L}$	m	10%	5%	1%	10%	5%	1%	10%	5%	1%
							$n = 100$, $T = 5$					
1	5	5	5	0.993	0.985	0.977	0.998	0.991	0.985	0.955	0.932	0.866
	5	6	6	0.994	0.992	0.979	0.997	0.995	0.988	0.978	0.960	0.905
	6	6	6	0.991	0.988	0.971	0.998	0.993	0.989	0.921	0.892	0.804
2	5	5	5	0.437	0.368	0.250	0.482	0.413	0.293	0.310	0.248	0.148
	5	6	6	0.453	0.398	0.259	0.494	0.430	0.312	0.330	0.268	0.176
	6	6	6	0.430	0.347	0.226	0.474	0.403	0.271	0.249	0.189	0.122
3	5	5	5	0.101	0.075	0.035	0.108	0.072	0.038	0.125	0.098	0.051
	5	6	6	0.118	0.085	0.040	0.122	0.081	0.040	0.139	0.096	0.056
	6	6	6	0.110	0.066	0.036	0.104	0.062	0.033	0.100	0.066	0.034
							$n = 100$, $T = 10$					
1	5	5	5	1	1	1	1	1	1	1	1	0.997
	5	6	6	1	1	1	1	1	1	1	1	0.998
	6	6	6	1	1	1	1	1	1	0.995	0.985	0.962
2	5	5	5	0.655	0.600	0.472	0.681	0.623	0.511	0.407	0.343	0.248
	5	6	6	0.684	0.607	0.482	0.707	0.640	0.518	0.463	0.383	0.257
	6	6	6	0.631	0.561	0.426	0.681	0.609	0.466	0.299	0.232	0.146
3	5	5	5	0.104	0.069	0.037	0.109	0.063	0.027	0.124	0.087	0.048
	5	6	6	0.108	0.077	0.035	0.120	0.071	0.028	0.120	0.082	0.049
	6	6	6	0.107	0.068	0.025	0.090	0.050	0.022	0.090	0.061	0.027
							$n = 400$, $T = 5$					
1	6	6	6	1	1	1	1	1	1	1	1	0.998
	6	7	7	1	1	1	1	1	1	1	1	1
	7	7	7	1	1	1	1	1	1	0.997	0.987	0.965
2	6	6	6	0.882	0.827	0.731	0.897	0.846	0.765	0.485	0.387	0.279
	6	7	7	0.890	0.841	0.731	0.900	0.860	0.772	0.527	0.432	0.310
	7	7	7	0.866	0.816	0.681	0.872	0.832	0.721	0.310	0.237	0.148
3	6	6	6	0.119	0.084	0.038	0.113	0.073	0.029	0.112	0.088	0.050
	6	7	7	0.127	0.088	0.034	0.123	0.085	0.031	0.114	0.091	0.049
	7	7	7	0.131	0.080	0.031	0.109	0.065	0.023	0.075	0.053	0.023
							$n = 400$, $T = 10$					
1	6	6	6	1	1	1	1	1	1	1	1	1
	6	7	7	1	1	1	1	1	1	1	1	1
	7	7	7	1	1	1	1	1	1	1	1	0.998
2	6	6	6	0.994	0.991	0.977	0.996	0.994	0.987	0.834	0.739	0.592
	6	7	7	0.995	0.994	0.980	0.997	0.994	0.989	0.860	0.783	0.623
	7	7	7	0.994	0.992	0.969	0.990	0.989	0.976	0.457	0.357	0.229
3	6	6	6	0.107	0.068	0.032	0.104	0.065	0.030	0.113	0.094	0.058
	6	7	7	0.109	0.066	0.032	0.101	0.065	0.030	0.111	0.078	0.052
	7	7	7	0.102	0.066	0.030	0.085	0.051	0.023	0.082	0.061	0.027

Note: True parameters: $\beta_{01} = \beta_{02} = 1$.

Table 6: Parameter estimates of SAR models of China air pollution data

	Parametric GMM		Sieve GMM	
	Coef.	se	Coef.	se
ρ	0.8875	0.0070		
Population	0.0019	0.0019	0.0023	0.0020
GDPPC	0.2561	0.1079	0.2620	0.1102
GDPPC^2	-0.0111	0.0054	-0.0091	0.0055
Sec_acc_gdp	-0.0002	0.0091	-0.0030	0.0096
Number of buses	-0.0415	0.0592	-0.1004	0.0608
Number of taxies	-0.0506	0.0574	-0.0193	0.0602
Road ratio	-0.3555	0.3278	-0.4417	0.3367

Note: Number of observations = 3408.

spectively, implying that the null hypothesis of linearity is not rejected at 5% level. Table 6 presents the results for estimating $(\beta_0', \gamma_0')'$. From Table 6, we see that the magnitude and significance of the coefficient estimates are similar between the linear SAR model and the nonlinear model. Besides, for both specifications, the coefficients of interest, i.e., GDP per capita and its quadratic term, are significantly positive and negative, respectively, implying an inverted U-shaped relationship between income and pollution levels. The estimation result of the nonlinear SAR model shows that the turning point of EKC is 144,472 RMB, which is similar to the one calculated by parametric SAR models (115,255 RMB).

6.2 Knowledge spillover

The seminal work of Coe and Helpman (1995) showed that a country's total factor productivity depends not only on its domestic stock of research and development (R&D) but also on the R&D stock of its neighbors, and international trade is an essential channel of technology spillover. Parametric SAR panel data models have been widely used to model such spatial diffusion and analyze the contribution of domestic and foreign R&D to productivity growth, see, e.g., Autant-Bernard and LeSage (2011), Lin and Kwan (2016) and Ho et al. (2018), among others. In this example, we examine international knowledge spillover through bilateral trade using the nonlinear SAR panel data model.

The dataset investigated in this empirical analysis covers 61 countries over the period 1998-2017. The time periods correspond to 4-year intervals. Denote $a_{it} = \ln(A_{it})$, $r_{it} = \ln(R_{it})$, and $l_{it} = \ln(L_{it})$, where A_{it} is the annual patent applications (divided by the total population in million), R_{it} is the R&D expenditure as a percentage of GDP, and L_{it} is the

total population in million. The model under consideration is

$$a_{it} = \sum_{j=1}^{n} w_{ij,t-1} h_0(a_{jt}) + \beta_{01} r_{i,t-1} + \beta_{02} l_{i,t-1} + c_i + \mu_t + \epsilon_{it}. \tag{26}$$

Similar to Ho et al. (2018), we exploit the panel structure of bilateral trade flows to construct the time-varying spatial weights, $W_{n,t-1} = [w_{ij,t-1}]_{ij=1,\ldots,n}$. The (i,j)th entry of the weights matrix $W_{n,t-1}$ is the bilateral import or export flow (the nominal US dollar value) between country i and j at time $t-1$. Equation (26) includes a full set of country dummies c_i and time dummies μ_t. The former captures any time-invariant country characteristics that affect its rate of innovation, while the latter captures common shocks to innovation across countries.

We estimate model (26) using the parametric GMM method by Lee (2007) assuming $h_0(y) = \rho_0 y$ and sieve GMM method. The times and individual fixed effects are eliminated using the same method as the first example. We use the quadratic B-spline basis functions with one internal knot for both k and q, which yields $J = \tilde{L} = 3$. Linear instruments z_{it} are generated from both r_{it-1} and l_{it-1}. For quadratic moments, we choose $\mathbf{P}_{ln,T-1}$ as an $N \times N$ block diagonal matrix with tth block being $W_{t-1}^l - (n-1)^{-1}\text{tr}(W_{t-1}^l J_n) J_n$ for $t = 1, \ldots, T$ and $l = 1, \ldots, 4$.

Table 7 presents the results for estimating β_0, and specification tests. From Columns 1-3, we see that when the spatial weights are constructed from either bilateral imports, exports or the sum of them, the coefficient ρ under the linear specification is significantly negative at the 1% level. This implies that innovation of neighboring countries could lead to less domestic innovation through bilateral trade, which is counter-intuitive as commodity trade can facilitate the technological spillover. We conduct LM and DM test, and the null hypothesis of linear specification is rejected at 5% level for all three cases. Figure 1 presents the estimated result for h_0 when the spatial weights are constructed from bilateral imports. We see that the spatial knowledge spillover effect is nonlinear and non-monotone. Specifically, if the knowledge production levels of neighboring regions are relatively low, a one percent increase in them results in less than a one percent increase in one's own productivity. However, as the innovation level of neighboring regions increases, the knowledge spillover effect could taper off and even be negative. Lastly, if the knowledge production levels of neighboring regions are high, a one percent increase in them results in more than a one percent increase in one's own productivity.

Table 7: Estimates and tests of SAR models of knowledge spillover data

	(1)	(2)	(3)
Parametric GMM			
ρ	-0.5274***	-0.4040***	-0.5694***
	(0.1359)	(0.0839)	(0.1036)
$r_{i,t-1}$	0.6456***	0.6837***	0.6715***
	(0.0817)	(0.0813)	(0.0809)
$l_{i,t-1}$	0.7876*	0.4750	0.7305*
	(0.4255)	(0.3797)	(0.3951)
Sieve GMM			
$r_{i,t-1}$	0.6446***	0.6843***	0.6947***
	(0.0831)	(0.0816)	(0.0783)
$l_{i,t-1}$	0.1050	-0.0272	0.1093
	(0.4998)	(0.4375)	(0.4703)
$\overline{\mathbf{LM}}_{nT}$	13.7135	5.0370	3.2429
$\overline{\mathbf{DM}}_{nT}$	5.7063	2.2983	3.3491
W.Matrix	IM	EX	IM+EX

Note: Number of observations = 305. Standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. For W. Matrices, IM, EX and IM+EX indicate that the weights matrices are constructed with bilateral imports, bilateral exports, and the sum of bilateral imports and exports, respectively. Ws are row normalized.

7 Conclusion

This paper provides sieve GMM estimation for SAR panel data models with a nonparametric endogenous effect. The new estimator utilizes both linear and quadratic moment conditions and is capable of accommodating a variety of DGPs, including a pure SAR panel data model with no relevant regressors. We establish the consistency and asymptotic normality of the estimator and show the efficiency gains from additional quadratic moments. We also propose LM and DM test statistics for testing the linearity of endogenous effect and derive their asymptotic distributions under the null and a sequence of local alternatives. Monte Carlo simulations show that our estimators and tests perform well for finite samples.

There are several possible extensions. These include choosing an optimal order of the basis expansion J and the number of moment conditions, investigating whether the proposed estimator can attain the optimal convergence rate, and selecting the best linear and quadratic moment conditions. In addition, considering a bootstrap version of our test could be helpful to improve the finite sample properties of the test. Moreover, it is meaningful to consider dynamic SAR panel data models with nonlinear endogenous effects. We leave these topics for future research.

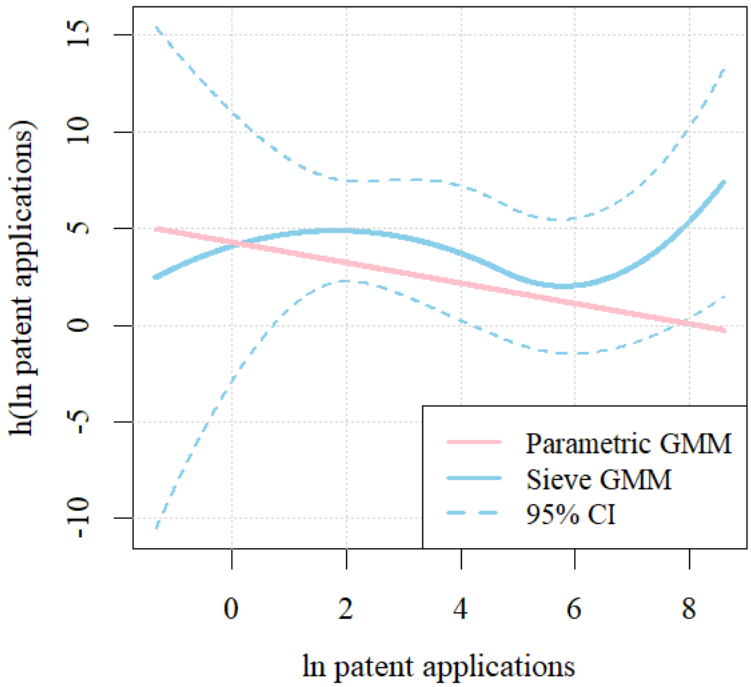

Figure 1: Estimation of h_0 for knowledge spillover data.

Appendix

List of Notations. The following list summarizes some frequently used notations in the main text and the supplementary material. Let $N \equiv n(T-1)$. For any vector/matrix Z_{nt}, let $\Delta Z_{nt} \equiv Z_{nt} - Z_{n,t-1}$, and $\Delta \mathbf{Z}_{n,T-1} \equiv (\Delta Z_{n2}, \ldots, \Delta Z_{nT})'$. Let $\mathscr{P}(A) = A(A'A)^{-1}A'$ for any matrix A with full column rank.

$$\Delta \mathcal{X}_{nt} \equiv (\Delta X_{nt}, W_n \Delta K_{nt}), \quad \Delta V_{nt} \equiv (P_{1n}^s \Delta \mathcal{E}_{nt}, \ldots, P_{mn}^s \Delta \mathcal{E}_{nt}, \Delta B_{nt}),$$

$$\begin{aligned}
&\widehat{D}_{nT,X} \equiv \Delta \mathbf{V}'_{n,T-1} \Delta \mathbf{X}_{n,T-1}/N, && D_{nT,X} \equiv \mathrm{E}\widehat{D}_{nT,X},\\
&\widehat{D}_{nT,J} \equiv \Delta \mathbf{V}'_{n,T-1} \mathbf{W}_{n,T-1} \Delta \mathbf{K}_{n,T-1}/N, && D_{nT,J} \equiv \mathrm{E}\widehat{D}_{nT,J},\\
&\widehat{D}_{nT} \equiv \Delta \mathbf{V}'_{n,T-1} \Delta \boldsymbol{\mathcal{X}}_{n,T-1}/N = (\widehat{D}_{nT,X}, \widehat{D}_{nT,J}), && D_{nT} \equiv \mathrm{E}\widehat{D}_{nT},\\
&\widehat{\Psi}_{nT,J} \equiv \Delta \mathbf{K}'_{n,T-1} \Delta \mathbf{K}_{n,T-1}/N, && \Psi_{nT,J} \equiv \mathrm{E}\widehat{\Psi}_{nT,J},\\
&\widehat{\Phi}^{(l)}_{nT} \equiv \Delta \boldsymbol{\mathcal{X}}'_{n,T-1} \mathbf{P}_{ln,T-1} \Delta \boldsymbol{\mathcal{X}}_{n,T-1}/N, && \Phi^{(l)}_{nT} \equiv \mathrm{E}\widehat{\Phi}^{(l)}_{nT},
\end{aligned}$$

where $P^s_{ln} = P_{ln} + P'_{ln}$. Also, denote $\Psi_{nT,L} \equiv \Delta\mathbf{B}'_{n,T-1}\Delta\mathbf{B}_{n,T-1}/N$ and

$$\Psi_{n,m} \equiv \frac{1}{n}\begin{pmatrix} \mathrm{tr}(P_{1n}P^s_{1n}) & \dots & \mathrm{tr}(P_{1n}P^s_{mn}) \\ \vdots & & \vdots \\ \mathrm{tr}(P_{mn}P^s_{1n}) & \dots & \mathrm{tr}(P_{mn}P^s_{mn}) \end{pmatrix}.$$

Additionally,

$$\Sigma_{nT}(y) \equiv \Gamma_{n,J}(y)(D'_{nT}\mathcal{W}_{nT}D_{nT})^{-1}D'_{nT}\mathcal{W}_{nT}\Omega_{nT}\mathcal{W}_{nT}D_{nT}(D'_{nT}\mathcal{W}_{nT}D_{nT})^{-1}\Gamma'_{n,J}(y),$$

$$\begin{aligned}
&\Sigma_{nT,X} \equiv D'_{nT,X}\mathcal{W}^{1/2}_{nT}\big(I_{d_g} - S^J_{nT}\big)\mathcal{W}^{1/2}_{nT}D_{nT,X}, && \Omega_{nT,X} \equiv \xi_{nT,X}\Omega_{nT}\xi'_{nT,X},\\
&\Sigma_{nT,h} \equiv D'_{nT,J}\mathcal{W}^{1/2}_{nT}\big(I_{d_g} - S^X_{nT}\big)\mathcal{W}^{1/2}_{nT}D_{nT,J}, && \Omega_{nT,h} \equiv \xi_{nT,J}\Omega_{nT}\xi'_{nT,J},\\
&v^2_{nT}(y) \equiv k^J(y)'\Sigma^{-1}_{nT,h}\Omega_{nT,h}\Sigma^{-1}_{nT,J}k^J(y), && \Sigma_{o,nT}(y) \equiv \Gamma_{n,J}(y)(D'_{nT}\Omega^{-1}_{nT}D_{nT})^{-1}\Gamma_{n,J}(y),
\end{aligned}$$

where $\Omega_{nT} \equiv \mathrm{Var}(\sqrt{N}g_{nT}(\theta_0))$, $S^J_{nT} \equiv \mathscr{P}(\mathcal{W}^{1/2}_{nT}D_{nT,J})$, $S^X_{nT} \equiv \mathscr{P}(\mathcal{W}^{1/2}_{nT}D_{nT,X})$, $\xi_{nT,X} \equiv D'_{nT,X}\mathcal{W}^{1/2}_{nT}\big(I_{d_g} - S^J_{nT}\big)\mathcal{W}^{1/2}_{nT}$, and $\xi_{nT,J} \equiv D'_{nT,J}\mathcal{W}^{1/2}_{nT}\big(I_{d_g} - S^X_{nT}\big)\mathcal{W}^{1/2}_{nT}$. Lastly, $\mathscr{V}_{nT} \equiv \mathscr{P}(\Omega^{-1/2}_{nT}D_{nT}) - \mathscr{P}(\Omega^{-1/2}_{nT}D_{nT,1})$ with $D_{nT,1} \equiv \mathrm{E}[N^{-1}\Delta\mathbf{V}'_{n,T-1}(\Delta\mathbf{X}_{n,T-1}, \mathbf{W}_{n,T-1}\Delta\mathbf{Y}_{n,T-1})]$.

References

Ai, C. and Chen, X. (2003), 'Efficient estimation of models with conditional moment restrictions containing unknown functions', *Econometrica* **71**(6), 1795–1843.

Autant-Bernard, C. and LeSage, J. P. (2011), 'Quantifying knowledge spillovers using spatial econometric models', *Journal of Regional Science* **51**(3), 471–496.

Baltagi, B. H., Egger, P. and Pfaffermayr, M. (2013), 'A generalized spatial panel data model with random effects', *Econometric Reviews* **32**(5-6), 650–685.

Belhaj, M., Bramoullé, Y. and Deroïan, F. (2014), 'Network games under strategic complementarities', *Games and Economic Behavior* **88**, 310–319.

Blundell, R., Chen, X. and Kristensen, D. (2007), 'Semi-nonparametric IV estimation of shape-Invariant Engel curves', *Econometrica* **75**(6), 1613–1669.

Boucher, V. and Fortin, B. (2016), Some Challenges in the Empirics of the Effects of Networks, *in* 'The Oxford Handbook of the Economics of Networks', Oxford University Press.

Boucher, V., Rendall, M., Ushchev, P. and Zenou, Y. (2024), 'Toward a general theory of peer effects', *Econometrica* **92**(2), 543–565.

Chang, H.-Y., Song, X. and Yu, J. (2024), 'Trending time-varying coefficient spatial panel data models', *Journal of Business & Economic Statistics* **0**(0), 1–13.

Chang, H.-Y., Wang, W. and Yu, J. (2021), 'Revisiting the environmental Kuznets curve in China: A spatial dynamic panel data approach', *Energy Economics* **104**, 105600.

Chen, X. (2007), Chapter 76 large sample sieve estimation of semi-nonparametric models, Vol. 6 of *Handbook of Econometrics*, Elsevier, pp. 5549–5632.

Chen, X. and Christensen, T. M. (2015), 'Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions', *Journal of Econometrics* **188**(2), 447–465. Heterogeneity in Panel Data and in Nonparametric Analysis in honor of Professor Cheng Hsiao.

Chen, X. and Christensen, T. M. (2018), 'Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric IV regression', *Quantitative Economics* **9**(1), 39–84.

Chen, X., Hong, H. and Tamer, E. (2005), 'Measurement error models with auxiliary data', *The Review of Economic Studies* **72**(2), 343–366.

Chen, X. and Pouzo, D. (2012), 'Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals', *Econometrica* **80**(1), 277–321.

Chetverikov, D. and Wilhelm, D. (2017), 'Nonparametric instrumental variable estimation under monotonicity', *Econometrica* **85**(4), 1303–1320.

Cliff, A. D. and Ord, K. (1973), *Spatial Autocorrelation*, London: Pion Ltd.

Coe, D. T. and Helpman, E. (1995), 'International R&D spillovers', *European Economic Review* **39**(5), 859–887.

de Jong, R. M. and Bierens, H. J. (1994), 'On the limit behavior of a chi-square type test if the number of conditional moments tested approaches infinity', *Econometric Theory* **10**(1), 70–90.

Dong, C., Gao, J. and Linton, O. (2023), 'High dimensional semiparametric moment restriction models', *Journal of Econometrics* **232**(2), 320–345.

Egger, P., Pfaffermayr, M. and Winner, H. (2005), 'An unbalanced spatial panel data approach to US state tax competition', *Economics letters* **88**(3), 329–335.

Elhorst, J. P. (2003), 'Specification and estimation of spatial panel data models', *International Regional Science Review* **26**(3), 244–268.

Ertur, C. and Koch, W. (2007), 'Growth, technological interdependence and spatial externalities: theory and evidence', *Journal of Applied Econometrics* **22**(6), 1033–1062.

Freyberger, J. and Horowitz, J. L. (2015), 'Identification and shape restrictions in nonparametric instrumental variables estimation', *Journal of Econometrics* **189**(1), 41–53.

Fruehwirth, J. C. (2013), 'Identifying peer achievement spillovers: Implications for desegregation and the achievement gap', *Quantitative Economics* **4**(1), 85–124.

Gupta, A. (2018), 'Nonparametric specification testing via the trinity of tests', *Journal of Econometrics* **203**(1), 169–185.

Gupta, A. and Qu, X. (2022), 'Consistent specification testing under spatial dependence', *Econometric Theory* p. 1–42.

Ho, C.-Y., Wang, W. and Yu, J. (2018), 'International knowledge spillover through trade: A time-varying spatial panel data approach', *Economics Letters* **162**, 30–33.

Hoshino, T. (2022), 'Sieve IV estimation of cross-sectional interaction models with nonparametric endogenous effect', *Journal of Econometrics* **229**(2), 263–275.

Hoxby, C. and Weingarth, G. (2005), 'Taking race out of the equation: School reassignment and the structure of peer effects', **18**. Unpublished working paper.

Jenish, N. and Prucha, I. R. (2009), 'Central limit theorems and uniform laws of large numbers for arrays of random fields', *Journal of Econometrics* **150**(1), 86–98.

Jenish, N. and Prucha, I. R. (2012), 'On spatial processes and asymptotic inference under near-epoch dependence', *Journal of Econometrics* **170**(1), 178–190.

Kelejian, H. H. and Prucha, I. R. (1998), 'A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances', *The Journal of Real Estate Finance and Economics* **17**(1), 99–121.

Kelejian, H. and Prucha, I. R. (2001), 'On the asymptotic distribution of the Moran I test statistic with applications', *Journal of Econometrics* **104**(2), 219–257.

Korolev, I. (2019), A consistent heteroskedasticity robust LM type specification test for semiparametric models, Papers 1810.07620, arXiv.org.

Kuersteiner, G. M. and Prucha, I. R. (2020), 'Dynamic spatial panel models: Networks, common shocks, and sequential exogeneity', *Econometrica* **88**(5), 2109–2146.

Lee, L.-F. (2007), 'GMM and 2SLS estimation of mixed regressive, spatial autoregressive models', *Journal of Econometrics* **137**(2), 489–514.

Lee, L.-F. and Yu, J. (2010), 'Estimation of spatial autoregressive panel data models with fixed effects', *Journal of Econometrics* **154**(2), 165–185.

Lee, L.-F. and Yu, J. (2012), 'QML estimation of spatial dynamic panel data models with time varying spatial weights matrices', *Spatial Economic Analysis* **7**(1), 31–74.

Lee, L.-F. and Yu, J. (2014), 'Efficient GMM estimation of spatial dynamic panel data models with fixed effects', *Journal of Econometrics* **180**(2), 174–197.

LeSage, J. P. and Fischer, M. M. (2020), 'Cross-sectional dependence model specifications in a static trade panel data setting', *Journal of geographical systems* **22**(1), 5–46.

Lin, M. and Kwan, Y. K. (2016), 'FDI technology spillovers, geography, and spatial diffusion', *International Review of Economics & Finance* **43**, 257–274.

Lin, X. and Lee, L.-F. (2010), 'GMM estimation of spatial autoregressive models with unknown heteroskedasticity', *Journal of econometrics* **157**(1), 34–52.

Malikov, E. and Sun, Y. (2017), 'Semiparametric estimation and testing of smooth coefficient spatial autoregressive models', *Journal of Econometrics* **199**(1), 12–34.

Newey, W. K. (2013), 'Nonparametric instrumental variables estimation', *The American Economic Review* **103**(3), 550–556.

Newey, W. K. and Powell, J. L. (2003), 'Instrumental variable estimation of nonparametric models', *Econometrica* **71**(5), 1565–1578.

Pakes, A. and Pollard, D. (1989), 'Simulation and the asymptotics of optimization estimators', *Econometrica* **57**(5), 1027–1057.

Su, L. and Jin, S. (2012), 'Sieve estimation of panel data models with cross section dependence', *Journal of Econometrics* **169**(1), 34–47. Recent Advances in Panel Data, Nonlinear and Nonparametric Models: A Festschrift in Honor of Peter C.B. Phillips.

Su, L. and Zhang, Y. (2016), 'Semiparametric estimation of partially linear dynamic panel data models with fixed effects', *Advances in Econometrics* **36**, 137 – 204.

Sun, Y. and Malikov, E. (2018), 'Estimation and inference in functional-coefficient spatial autoregressive panel data models with fixed effects', *Journal of Econometrics* **203**(2), 359–378.

White, H. and Wooldridge, J. M. (1991), Some results on sieve estimation with dependent observations, Nonparametric and Semiparametric Methods in Econometrics and Statistics. Cambridge University Press, Cambridge, pp. 459–493.

Xu, X. and Lee, L.-F. (2015), 'Maximum likelihood estimation of a spatial autoregressive tobit model', *Journal of Econometrics* **188**(1), 264–280.

Zhang, Y. and Zhou, Q. (2021), 'Partially linear functional-coefficient dynamic panel data models: sieve estimation and specification testing', *Econometric Reviews* **40**(10), 983–1006.