Networked instrumental variable estimation: The case of Hausman-style instruments

Shi, Xiangyu

Yale University

June 2024

Networked instrumental variable estimation: The case of Hausman-style instruments[*]

Xiangyu Shi[†]

May 28, 2024

Summary

In this paper, I argue that in situations of complex network dependence, the traditional and widely used Hausman-style instrumental variable estimation may not be valid for causal identification. This is the case for inter-regional migration networks when evaluating place-based labor market policies, and for correlated unobserved consumer tastes in the product and geographic space in demand estimation. I build an economic model for these two cases, respectively, to derive the estimating equation and to shed light on the fallacy—omitted variable bias and the resulting violation of exclusion restriction—of the traditional econometric framework. I then build an alternative econometric framework and propose a new approach to estimation that exploits higher-order network neighbors and, then, I establish its desirable properties. I conduct Monte Carlo simulations and two empirical analyses that each correspond to the two economic models to validate this new approach of estimation.

JEL Classification: C01, C13, C21, C33

Keywords: treatment effect; network; instrumental variable; Hausman IV; spatial linkages; migration network; demand estimation

[*] We thank the insightful discussions and comments by Wayne Yuan Gao, Yuichi Kitamura, Lihua Lei, Michael Leung, Zhentao Shi, Michael Sullivan, Elie Tamer, Yiqing Xu, and participants in various seminars and conferences.

[†] Shi: Department of Economics, Yale University. Email: xiangyu.shi@yale.edu.

1 Introduction

Instrumental variable estimation is widely used to evaluate causal effects. Hausman-style instruments, first brought up in Hausman, Leonard, and Zona (1994), have been a popular choice for identification in the fields of industrial organization and spatial economics (Nevo, 2001; Lanoie et al., 2011; Crawford and Yurukoglu, 2012; Houde, 2012; and Azar, Berry, and Marinescu, 2022). The idea of Hausman-style instruments is to exploit cross-sectional dependence of some plausibly exogenous characteristics, and construct a (weighted) average to serve as the instrument for the endogenous variable. For example, in Azar, Berry, and Marinescu (2022), the authors use the labor market conditions of other cities (geographical neighbors) as instruments of the wage of the focal city. In Nevo (2001), the author uses the prices of other cities (also geographical neighbors) as instruments of the price of the focal city.[1]

However, this approach is questionable when we take into account more complicated network linkages. For example, when estimating the effects of various factors on local labor market outcomes as in Azar, Berry, and Marinescu (2022), migration flows among different localities, or inter-regional migration networks, are a mechanism at play that leads to complex cross-sectional or network dependence. Exploiting a Hausman-style instrument that does not correctly take it into consideration may lead to biases in the estimation. Another example is demand estimation, as in the seminal work of Berry (1994) and Nevo (2001), in which the correlation of unobserved factors of consumer tastes in different localities of the geographical or product space[2] may cause unaccounted cross-sectional or network dependence that may also lead to biases in the estimation.

In this paper, I reevaluate and redesign instrumental variable estimation, using Hausman-style instruments as a case in point, from the perspective of networks or cross-sectional dependence. I first lay out two economic models, one on regional labor markets and migration networks, and the other on consumer demand in the product space, to motivate and illustrate the idea of unaccounted network dependence and derive the estimating framework. Next, I construct the econometric model based on the previous economic models, which illustrates the idea of how network dependence introduces an omitted variable bias and, thus, leads to an inconsistent and biased estimator. I then propose an appropriate approach to estimation to account for this issue. The idea of this new estimator is to employ the characteristics of higher-order network neighbors as the instruments, similar to that in Bramoullé, Djebbari, and Fortin (2009).[3] The intuition is that under complicated and often unknown network dependence, more distant higher-order neighbors are more likely to satisfy exclusion restrictions. The asymptotic properties of the new estimator can then be established in a standard way. Next, I conduct Monte Carlo simulations to validate the new estimation approach. Finally, I conduct two sets of empirical analyses to compare the traditional and the newly proposed Hausman IV estimation. I first estimate the effects of infrastructure on the regional labor market, which corresponds to the case of migration networks. I then estimate the demand equation for the US airline market, which corresponds to the case of correlated consumer tastes in the product space.[4] In particular, I propose that when the detailed data-generating process is unknown, researchers should take a data-driven approach to use different orders of neighbors as instruments for robustness checks and evaluate whether results are stable. In fact, the results may become stable when the order of neighbors becomes large, which also implies that the newly proposed Hausman-style IV estimation can

[1] I provide a formal construction of the instrument in Section 2.

[2] Product space is defined as the relationship of complementarity and substitution between different products. For example, an air ticket from New York to Los Angeles is a substitute for an air ticket for transferring the route from New York to Los Angeles via Orlando. These two tickets can be seen as neighbors in the product space.

[3] Bramoullé, Djebbari, and Fortin (2009) uses higher-order neighbors as instruments to estimate endogenous peer effects. My paper is similar to it since we both exploit the plausibly exogenous variations of higher-order neighbors.

[4] Note that in the case of the airline market, the product space is also associated with the geographical space, since a ticket is always associated with geographical locations.

avoid biases due to cross-sectional network dependence.[5]

The remainder of this paper proceeds as follows. Section 2 lays out two economic models to shed light on the estimating equation and the correct econometric framework. Section 3 constructs the econometric model. Section 4 introduces the new approach to Hausman IV estimation. Section 5 conducts a series of Monte Carlo simulations. Section 6 conducts two sets of empirical analyses. Finally, Section 7 concludes.

2 Economic Models

2.1 A Model of Inter-regional Migration Networks

The model herein is an adapted version of Caliendo, Dvorkin, and Parro (2019). There are N localities, indexed by $i = 1, 2, ..., N$. There is a mass of L_i of households in location i who provide a unit of labor inelastically. Assume that in each location, the households can only migrate to a subset of other locations, denoted by $N(i)$. Such locations can be regarded as the network neighbors of i. The pattern of inter-regional migration corresponds to the structure of the migration network. The preference takes a log form, in which $U(C_i) = A_i \log(C_i)$, where C_i is the consumption, and A_i is the preference shifter or the amenities of location i, including the level of infrastructure that is endogenously determined by government policies. The migration choice problem is formulated as follows:

$$v_i = A_i \log C_i + \max_{j \in N(i)} (v_j - \tau_{i,j} + \nu_j \epsilon_j), \tag{1}$$

where v_i is the value function, or the utility of residing in location i, $\tau_{i,j} = \tau_{j,i} \geq 0$ is the (symmetric) migration cost, ν_j is an exogenous preference shifter of j, and ϵ_j is the idiosyncratic shock. Assume that ϵ is i.i.d. and distributed type-I extreme value with zero mean, and let $V_i = E(v_i)$ denote the expected utility, let $\mu_{j,i}$ be the fraction of households that relocate from j to i, we have that

$$V_i = A_i \log C_i + \mu_i \log(\sum_{j \in N(i)} \exp(V_j - \tau_{i,j})^{1/\nu_j}). \tag{2}$$

and

$$\mu_{j,i} = \frac{\exp(V_i - \tau_{i,j})^{1/\nu_i}}{\sum_{h \in N(i)} \exp(V_h - \tau_{h,j})^{1/\nu_h}}. \tag{3}$$

The proofs and derivations are relegated to Appendix A. By simple algebra, we have that for $j_1, j_2 \in N(i)$,

$$\log(\frac{\mu_{j_1,i}}{\mu_{j_2,i}}) = \frac{1}{\nu_{j_1}}(V_{j_1} - \tau_{i,j_1}) - \frac{1}{\nu_{j_2}}(V_{j_2} - \tau_{i,j_2}). \tag{4}$$

Let the labor supply in equilibrium in i be denoted as μ_i, given equation (4), we can express μ_i as a (non-parametric) function of variables A_i, ν_i and all network neighbors[6]:

$$\mu_i = \sum_{j \in N(i)} \mu_{j,i} = f(A_i, \nu_i, e_i), \tag{5}$$

where e_i is a (non-parametric) function of network neighbors

[5] In the two examples of Section 6, the results are stable when the order exceeds 2.

[6] Due to the nonlinearity in equation (4), the network neighbors' term cannot be perfectly canceled out under general conditions when calculating $\log \sum_{j \in N(i)} \mu_{j,i}$, which is the labor supply in i, $\log \mu_i$

$$e_i = g(A_{j_1}, \nu_{j_1}, \tau_{i,j_1}, ..., A_{j_K}, \nu_{j_K}, \tau_{i,j_K}), j_1, ..., j_K \in N(i). \tag{6}$$

We may be interested in estimating the effects of place-based labor market policies, or $\frac{\partial f}{\partial A_i}$, in which A_i may be endogenously determined. To estimate (a linear approximation of) equation (5), one may use a Hausman-style instrument, $\sum_{j\in N(i)} w_{ij}\nu_j$, where w_{ij} is the weight, but this is invalid and produces a biased estimate because $cov(e_i, \sum_{j\in N(i)} w_{ij}A_j)$ or $cov(e_i, \sum_{j\in N(i)} w_{ij}\nu_j)$ is not zero. This corresponds to the discussion in the next section.

2.2 A Model of Correlated Consumer Tastes in Product Space

This model is an adapted version of Berry (1994). Assume that there are N products that a generic consumer i considers to purchase, indexed as $j = 1, 2, ..., N$. We further restrict that a neighborhood $N(j)$ of j is a closer substitute of j that may exhibit the network structure of the product space. The full model is relegated to Appendix A, and we start from the mean utility of product j, which is

$$\delta_j = x_j\beta_1 + \sum_{k\in N(j)} w_{jk}z_k\beta_2 - \alpha p_j + \xi_j, \tag{7}$$

where x_j is the observed characteristics of j, and $\sum_{k\in N(j)} w_{jk}z_k$ is the weighted average of the other observed characteristics of j's network neighbors in the product space.[7] For example, in the case of airline networks, the demand for the route "New York—Los Angeles" is affected by the characteristics of "New York—Orlando" because passengers from New York can transfer in Orlando to finally arrive in Los Angeles. Another example is Berry, Gandhi, and Haile (2013), in which some "connected substitutes" may be subject to correlated demand shocks, which in turn justifies the inclusion of $\sum_{k\in N(j)} w_{jk}z_k$.

Therefore, $\sum_{k\in N(j)} w_{jk}z_k$ enters equation (7) for at least two reasons. First, consumer tastes across different products in the product space are correlated. The change in characteristics of other products may also change the consumer's evaluation of the focal product. Second, the quality or other product characteristics are endogenously determined and change over time, which in turn affects consumer tastes. For example, when there is a large traffic flow in the airport of New York due to a surge of passengers from Orlando to New York, ticket holders of the flight New York-Los Angeles are subject to traffic congestion, and, their utility may be compromised.

In addition, the neighbors can also be defined in a geographical space, as in Nevo (2001), and the variations of geographically neighboring observations driven by migration and traffic flows may matter for the estimation of demand. p_j is the endogenously determined price, and ξ_j is an i.i.d. shock. Given the standard logistic setting as in the demand estimation literature, we have the following estimating equation

$$\log(s_j) - \log(s_0) = x_j\beta_1 + \sum_{k\in N(j)} w_{jk}z_k\beta_2 - \alpha p_j + \xi_j, \tag{8}$$

where s_j is the market share of j, and s_0 is the market share of the outside option. To deal with the endogeneity of the price p_j, one may consider using a Hausman-style instrument $\sum_{k\in N(j)} w_{jk}\omega_k$, where ω_k constitutes exogenous variations. However, given equation (8), it is likely that $cov(\sum_{k\in N(j)} w_{jk}\omega_k, \sum_{k\in N(j)} w_{jk}z_k) \neq 0$ and $\beta_2 \neq 0$, thus resulting in a biased estimator. This corresponds to the discussion in the next section.

[7] Here I make a specific restriction on how network neighbors may affect δ_j. There are other parametric and even non-parametric modeling choices. In general, the equation can be written as $\delta_j = x_j\beta_1 + f_j(\{z_k\}_{k\in N(j)}, \beta_2) - \alpha p_j + \xi_j$, where $f_j(\cdot)$ is an arbitrary function.

3 Econometric Model

In the data set, there are N agents, indexed by $i = 1, 2, 3, ..., N$. The network structure among the agents is denoted as an $N \times N$ matrix $W(N)$,[8] and is fixed, pre-determined, and can be observed by all agents and econometricians. Each entry in $W(N)$, w_{ij}, is a weakly positive real number that describes the strength of the connection between agents i and j. The sum of each row is normalized to 1, or $\sum_j w_{ij} = 1$. By convention, $w_{ii} = 0$ for all $i = 1, 2, ..., N$. It is also associated with the concept of spatial weights in the spatial econometric models. $N(i)$ denotes the set of neighbors of agent i. For all $j \in N(i)$, $w_{ij} > 0$. In the Hausman-style instrumental variable estimation, generic second-stage and first-stage equations are, respectively, the following:

$$Y_i = X_i\beta_1 + Z_i\beta_2 + u_i, X_i = \alpha \sum_{j\in N(i), j\neq i} w_{ji}Z_j + v_i, \tag{9}$$

where Y_i is the outcome variable of interest, X_i is a vector of (predicted) endogenous regressors, Z_i is a vector of exogenous regressors, and u_i and v_i are error terms. [9] $\sum_{j\in N(i),j\neq i} w_{ji}Z_j$ is the Hausman-style instrumental variable. When taking a simple arithmetic average, we have $w_{ji} = \frac{1}{|N(i)|}$ for all j, where $|N(i)|$ is the cardinality of $N(i)$. β_1 is the parameter of interest. We assume that $E(u_i) = E(v_i) = E(Z_iu_i) = E(\sum_{j\in N(i),j\neq i} w_{ji}Z_jv_i) = 0$. Given the standard econometrics textbook materials, we have the following lemma:

Lemma 1. *In equation (9), the Hausman-style instrumental variable estimation produces a consistent estimator if and only if* $cov(\sum_{j\in N(i),j\neq i} w_{ji}Z_j, X_i) \neq 0$ *and* $cov(\sum_{j\in N(i),j\neq i} w_{ji}Z_j, v_i) = cov(\sum_{j\in N(i),j\neq i} w_{ji}Z_j, u_i) = 0$.

When there is (first-order) network dependence that is unaccounted for, we have

$$u_i = \gamma^X \sum_{j\in N(i),j\neq i} w_{ji}(X_j - E(X_j)) + \gamma^Z \sum_{j\in N(i),j\neq i} w_{ji}(Z_j - E(Z_j)) + e_i, \tag{10}$$

where $\sum_{j\in N(i),j\neq i} w_{ji}(X_j - E(X_j))$ and $\sum_{j\in N(i),j\neq i} w_{ji}(Z_j - E(Z_j))$ are a first-order weighted sum of neighbors' variables and both capture network or cross-section dependence that is omitted in the second-stage regression. These two terms correspond to e_i (equation (5)) in the case of inter-regional migration networks and to $\sum_{k\in N(j)} w_{jk}z_k$ in the case of correlated consumer tastes (equation (8)). I add the constant terms $\sum_{j\in N(i),j\neq i} w_{ji}(-E(X_j))$ and $\sum_{j\in N(i),j\neq i} w_{ji}(-E(Z_j))$ to make sure that $E(u_i) = E(e_i) = 0$. To formulate the above equations in the matrix form, we have

$$\mathbf{Y} = \mathbf{X}\beta_1 + \mathbf{Z}\beta_2 + \mathbf{u}, \mathbf{X} = \mathbf{WZ} + \mathbf{v}, \tag{11}$$

and

$$\mathbf{u} = \mathbf{a} + \gamma^X\mathbf{WX} + \gamma^Z\mathbf{WZ} + \mathbf{e}. \tag{12}$$

The existence of the constant term, $\mathbf{a}$, is to make the mean of $\mathbf{u}$ and $\mathbf{e}$ zero. $\mathbf{u}$ corresponds to e_i in equation (5) and to $\sum_{k\in N(j)} w_{jk}z_k\beta_2 + \xi_j$ in equation (7). Here, the Hausman-style instrument fails because the standard exogeneity condition may not be satisfied, i.e., $cov(\mathbf{u}, \mathbf{WZ}) = E(\mathbf{Z'Wu}) \neq \mathbf{0}$. Instead, we have

$$E(\mathbf{Z'Wu}) = E(\mathbf{Z'W}(\gamma^X\mathbf{WX} + \gamma^Z\mathbf{WZ} + \mathbf{e})) = \gamma^X E(\mathbf{Z'W}^2\mathbf{X}) + \gamma^Z E(\mathbf{Z'W}^2\mathbf{Z}) \neq \mathbf{0}. \tag{13}$$

[8] For the ease of exposition, we use the matrix form $\mathbf{W}$ instead of $W(N)$ below, when there is no confusion.

[9] Note that X_i and Z_i can both be i.i.d. vectors that do not exhibit cross-sectional or network dependence. Under this assumption, we still get an invalid Hausman-IV estimation.

Under network or cross-sectional dependence, we generically have that $E(\mathbf{Z}'\mathbf{W}^2\mathbf{X}), E(\mathbf{Z}'\mathbf{W}^2\mathbf{Z}) \neq \mathbf{0}$. If $\gamma^X, \gamma^Z \neq 0$ and $||\gamma^X E(\mathbf{Z}'\mathbf{W}^2\mathbf{X})|| \neq \gamma^Z ||E(\mathbf{Z}'\mathbf{W}^2\mathbf{Z})||$, we have that $cov(\mathbf{u}, \mathbf{WZ}) \neq \mathbf{0}$. This is a source of omitted variable bias. Note that equation (12) is only a special case in which only the first-order neighbors enter the equation. By Lemma 1, the traditional Hausman-style instrumental variable estimation yields an inconsistent estimator. For a general case, we have

$$\mathbf{u} = \mathbf{a} + \sum_{p=1}^{K^X} \gamma_p^X \mathbf{W}^p \mathbf{X} + \sum_{q=1}^{K^Z} \gamma_q^Z \mathbf{W}^q \mathbf{Z} + \mathbf{e}, \tag{14}$$

where up to K^Xth-order of X and K^Zth-order of Z are accounted for.[10] In the exposition below, we consider this general case. This is a standard omitted variable bias model.

4 New Estimation Approach: High-order Neighbors as IV

4.1 New Approach

I first establish the following lemma regarding the covariance of high-order network neighbors. It is used for establishing the validity of the newly proposed Hausman IV.

Lemma 2. *For two **i.i.d.** random vectors, $\boldsymbol{X} = (X_1, ..., X_N)$ and $\boldsymbol{Z} = (Z_1, ..., Z_N)$, and a well-defined $N \times N$ network weight matrix $\boldsymbol{W}$ whose all entries are weakly positive, if $cov(\boldsymbol{X}, \boldsymbol{Z}) = \boldsymbol{0}$, then for any positive integers p and q, $cov(\boldsymbol{W}^p\boldsymbol{X}, \boldsymbol{W}^q\boldsymbol{Z}) = \boldsymbol{0}$; if $cov(\boldsymbol{X}, \boldsymbol{Z}) \neq \boldsymbol{0}$, then for any positive integers p and q, $cov(\boldsymbol{W}^p\boldsymbol{X}, \boldsymbol{W}^q\boldsymbol{Z}) \neq \boldsymbol{0}$.*

The proof is obvious using the fact that the two vectors are **i.i.d.** and all entries of the weight matrix are weakly positive. Recall that the econometric model consists of equations (11) and (14). Rearranging yields

$$\mathbf{Y} = \mathbf{a} + \mathbf{X}\beta_1 + \mathbf{Z}\beta_2 + \sum_{p=1}^{K^X} \gamma_p^X \mathbf{W}^p \mathbf{X} + \sum_{q=1}^{K^Z} \gamma_q^Z \mathbf{W}^q \mathbf{Z} + \mathbf{e}. \tag{15}$$

Note that in this equation, there are K^X+1 endogenous regressors, since generically, $cov(\mathbf{X}, \mathbf{e}), cov(\mathbf{W}^p\mathbf{X}, \mathbf{e}) \neq \mathbf{0}$, for $p = 1, 2, ..., K^X$. However, at the same time, $cov(\mathbf{Z}, \mathbf{e}), cov(\mathbf{W}^q\mathbf{Z}, \mathbf{e}) = \mathbf{0}$, for $q = 1, 2, ..., 2\max\{K^X + K^Z\}$. Moreover, we have that $cov(\mathbf{v}, \mathbf{W}^q\mathbf{Z}) = \mathbf{0}$ and $cov(\mathbf{W}^p\mathbf{X}, \mathbf{W}^q\mathbf{Z}) \neq \mathbf{0}$. Therefore, we propose a new Hausman-style instrumental variable estimator, in which $\mathbf{V}^Z = (\mathbf{W}^{\max\{K^Z, K^X\}+1}\mathbf{Z}, ..., \mathbf{W}^{\max\{K^Z, K^X\}+K^X}\mathbf{Z})$ serves as the instrument for $\mathbf{V}^X = (\mathbf{X}, \mathbf{WX}, ..., \mathbf{W}^{K^X}\mathbf{X})$,[11] and, in the second-stage regression equation (15), $(\mathbf{WZ}, ..., \mathbf{W}^{K^Z}\mathbf{Z})$ is is added as a vector of additional control variables. Note that the instrument starts from the $\max\{K^Z, K^X\} + 1$th-order neighbors to satisfy the exclusion restriction. Even higher-order neighbors of $\mathbf{Z}$ can be included as the instruments since they satisfy exclusion restrictions. I provide the formal definition of the estimator in Appendix B.

Given the above equations, the new approach to the estimation is as follows. In the first step, include high-order neighbors of Z, up to K^Zth order, as the control variables. In the second step, use the high-order neighbors of Z, $\mathbf{V}^Z$, as the instruments for the vector of endogenous regressors, $\mathbf{V}^X$, and conduct the ordinary instrumental variable estimation. This procedure is also similar to that in Bramoullé, Djebbari, and Fortin (2009), where higher-order neighbors' exogenous characteristics are also used to estimate the parameter, which

[10] Here K^X and K^Z are two scalars that can be the same or different.

[11] If we use more instrumental variables, the estimation will be more efficient, but there will be a larger small-sample bias. We leave the choice of the optimal number of IVs to future research.

is the peer effect in that case. The intuition of the new approach is that under complicated and often unknown network dependence, more distant higher-order neighbors are more likely to satisfy exclusion restrictions. Using standard arguments (including Lemma 1) and under some standard regularity conditions, we have that such a new estimator produces an asymptotic consistent and normal estimator of the true parameter, β_1. Due to space limitations, I relegate the discussion to Appendix B. Here we only provide a lemma on the consistency. Using this lemma and Lemma 2, we can establish that the new Hausman IV estimator is consistent.

Lemma 3. *In equation (15), the Hausman-style instrumental variable estimation produces a consistent estimator if and only if* $cov(\boldsymbol{Z}, \boldsymbol{e}) = cov(\boldsymbol{WZ}, \boldsymbol{e}) = ... = cov(\boldsymbol{W}^{K^X}\boldsymbol{Z}, \boldsymbol{e}) = \boldsymbol{0}$*, and* $cov(\boldsymbol{V}^Z, \boldsymbol{V}^X) \neq \boldsymbol{0}$*.*

In the above analysis, we assume that the parameters K^X and K^Z are known. However, it is the case if we have a model that explicitly dictates so, but in many cases, it is not. For feasibility, we could set these parameters flexibly and use a data-driven approach to check how the estimates vary correspondingly. The reason is that, in principle, choosing a high K^X and K^Z leads to less efficient estimation, but an estimator that is more likely to be unbiased and consistent. Therefore, this is a bias-efficiency tradeoff. We can use some tests to address this issue proposed below.

4.2 Some Tests

In this section, we discuss some relevant statistical tests, including (1) Weak IV tests, (2) Hausman specification tests, and (3) Overidentification tests. I employ these tests in the Monte Carlo simulations and empirical analyses. These tests are needed to evaluate the validity of the new approach of Hausman-style IV estimation. A valid Hausman-IV estimation should satisfy the following three conditions: (1) reject the null of the weak IV test; (2) does not reject the null of the Hausman specification test, where the alternative model is the one that exploits higher-order neighbors; and (3) does not reject the null of the overidentification test. I propose the detailed tests as follows.

4.2.1 Weak IV Tests

We first discuss the test of weak instruments, following a multiple endogenous regressors framework in Sanderson and Windmeijer (2016). We follow the same weak-IV asymptotics and employ the conditional F-test similar to that of Cragg and Donald (1993). We need a sufficiently large F-statistic to reject the null that instruments are weak. In the Monte-Carlo simulations and the empirical analysis below, we exploit the same weak-IV test as in Sanderson and Windmeijer (2016). The technical details can be found in Section 4.3 of Sanderson and Windmeijer (2016), I omit the discussion here to avoid repetition.[12]

4.2.2 Hausman Specification Tests

Finally, we can conduct a Hausman-style specification test (Hausman, 1978) to compare these two econometric models—the traditional model and the new proposed model—and the associated approach to estimation. We can also use this Hausman test to choose between two models with different K^X and K^Z. The null hypothesis is that the model with a smaller K^X and K^Z yields more efficient estimates, and both models yield consistent estimates. The alternative hypothesis is that only the model with a larger K^X and K^Z yields consistent estimates.

[12] The authors also provided STATA codes for the statistical tests on the journal website.

4.2.3 Overidentification Tests

Since it is possible that the number of instruments is larger than the number of the endogenous variables, we can employ the classical Sargan-Hansen overidentification test (Sargan, 1958; Hansen, 1982). It is also called the J-test in practice. The J-statistic J follows χ_m, where m is the number of additional instruments used for estimation. We need a small J to avoid rejecting the null hypothesis that the model is not overidentified.

5 Monte Carlo Simulations

In this section, I conduct three sets of Monte Carlo simulations. I first set that X and Z are scalar, and $\beta_1 = 0.5$, $\beta_2 = 0.3$. For Model 1, I consider first-order neighbors in equation (14). For Model 2, I consider second-order neighbors for X and first-order neighbors for Z. For Model 3, I consider first-order neighbors for X and second-order neighbors for Z. Regardless of the order of neighbors, I set $\gamma_k^X = \gamma_k^Z = 0.2$ for all k. I also assume that e and v distributed i.i.d. normally with zero mean and variance 0.1. Simulations are repeated for 10000 times.

I vary the sample size by $N = 50, 100, 200, 500, 1000$, and generate the data according to equations (11) and (14). For each sample size, I randomly generate the same matrix of network neighbors, $\mathbf{W}$, and then fix it for each simulation. I conduct the traditional Hausman-style estimation and the new Hausman-style estimation and report the results in Table 1. For all sample sizes and all models, the traditional approach produces a sizable bias at a magnitude of 0.1 (20% of the true value), while the new approach produces a much smaller finite-sample bias (3% of the true value). Hausman specification tests always reject the null that both estimators are consistent.[13] The new approach outperforms the traditional one when we take into account network dependence. Moreover, using the new approach, I calculate the probability associated with different critical values and report the results in Table 2. The results suggest that the estimator is indeed asymptotically normal.

Next, I consider the choice of orders K^X and K^Z. I first discuss the consequences of misspecifying these orders. I fix the sample size to $N = 200$, and try different K^X and K^Z in the true model and in the actually-specified estimation. I report the results in Table 3. If we over-specify K^X and K^Z, there will not be a significant bias. But if we under-specify K^X and K^Z, there will be a significant bias. Next, I discuss the choice of the number of instruments. It is a well-established result in the literature that incorporating more instruments may lead to more efficient estimation but a larger finite-sample bias. I also report the results in Table 3. Including more IVs than necessary slightly increases the bias. Though not reported, the standard deviations of including more IVs are smaller. The Monte Carlo results are consistent with the theoretical arguments in the literature. Finally, I conduct a series of weak-IV tests using the method of Sanderson and Windmeijer (2016). The p-values associated with the F-statistics are all 0.000 when the model is correctly specified as using the new Hausman-style estimation.

6 Empirical Analysis

In this section, I conduct two sets of empirical analyses to further validate the new approach of Hausman IV estimation. The first analysis is to estimate the effects of infrastructure on city labor markets in China, which corresponds to the case of migration networks. The second analysis is the demand estimation in the

[13] The p-values, though not reported in the associated tables, are all 0.000.

US airline market, which corresponds to the case of correlated consumer tastes in the product market. The summary statistics of the data used in these analyses are presented in Table 4.

6.1 Effects of Infrastructure on Labor Market

I first estimate the effects of infrastructure on city labor markets in China. This will help us understand the effects of the place-based policy of infrastructure improvement on the labor market. The data source is the China City Statistics Yearbooks for 1999-2018. The dependent variable is the share of the labor force in a city for the primary, secondary, and tertiary industries. The main endogenous variable of interest is a measure of the level of infrastructure: log per capita road area in a certain city. I use two sets of instrumental variables separately: the log distance to the nearest railroad interacted with linear year trends, and the log distance to the province capital. Neighbors are defined as cities within a 250-kilometer radius. The estimating equations for the new approach of Hausman IV estimation are as follows:

$$LaborShare_{it} = \alpha_1 \log(Road_{it}) + \alpha_2 W \log(Road_{it}) + \beta \log(Distance_i) \times Year_t + \lambda_i + \lambda_t + u_{it}, \tag{16}$$

where $LaborShare_{it}$ is the labor share in city i and year t, $\log(Road_{it})$ is the per capita road area, $\log(Distance_i)$ is the log distance to the nearest railroad or the province capital, λ_i is the city fixed effects, λ_t is the year fixed effects, and u_{it} is the error term. The first-stage equations are as follows:

$$\begin{aligned} \log(Road_{it}) &= \gamma_1 W \log(Distance_i) \times Year_t + \gamma_1 W^2 \log(Distance_i) \times Year_t + \lambda_i + \lambda_t + v_{1,it}, \\ W\log(Road_{it}) &= \delta_1 W \log(Distance_i) \times Year_t + \delta_1 W^2 \log(Distance_i) \times Year_t + \lambda_i + \lambda_t + v_{2,it}, \end{aligned} \tag{17}$$

where $W\log(Distance_i)$ is the first-order network neighbors' simple arithmetic average distance, and $W^2\log(Distance_i)$ is the second-order network neighbors' simple arithmetic average distance, and $W\log(Road_{it})$ is the first-order network neighbors' simple arithmetic average road area.

I report the results in Table 5. For Panel A, I employ an OLS estimation with no instruments. For Panel B, the instruments in columns (1) through (3) are first-order neighbors of log distance to the railroad; The instruments in columns (4) through (6) are first-order neighbors of log distance to the province capital. For Panel C, the instruments in columns (1) through (3) are first- and second-order neighbors of log distance to the railroad; The instruments in columns (4) through (6) are first- and second-order neighbors of log distance to the province capital. The coefficients on the main parameter of interest are much different across different settings in Panels A through C. For example, the coefficient in column (2) in Panel C is nearly twice as much in magnitude as that in Panel B. The Hausman specification test for each pair of settings yields a p-value of 0.000. Thus, the traditional way of Hausman IV may produce a significant bias. However, if I increase K^X and K^Z as in Panels D and E, the results remain quite quantitatively stable, and the p-value of the Hausman specification test is larger than 0.1. Therefore, for the case of evaluating the effects of transportation infrastructure on the labor market, it is sufficient to use the new estimation approach with $K^X = K^Z = 1$. The p-values of the weak-IV test associated with Panels B through D are all smaller than 0.01, rejecting the null that the IVs are weak. Thus, Panels D and E illustrate a data-driven approach to choosing an appropriate order of neighbors.

6.2 Demand Estimation in Airline Market

Next, I conduct a demand estimation using data from the US airline market. The data source is the Airline Origin and Destination Survey (DB1B).[14] A product corresponds to a market between a specific airport pair (of different airlines). The product space of the airline market exhibits the property of network dependence since passengers can transfer and take more than one route to travel between places. Each observation corresponds to an airline market, which is an airport pair at a certain time. The dependent variable is the log market share, constructed as in equation (8). The main endogenous dependent variable is the log price, which is the average of the product of the year-quarter. The instrument is the log distance of the air route between an airport pair for a certain airline company interacting with linear year-quarter time trends. Neighbors are defined as airline markets that share an origin or a destination airport.[15] The estimating equations for the new approach of Hausman IV estimation are as follows:

$$\begin{aligned}\log(MarketShare_{ijat}) = {} & \alpha \log(p_{ijat}) + \beta_1 W \times OnTimeRatio_{ijat} + \beta_2 \log(Distance_{ija}) \times Quarter_t \\ & + \lambda_{ij} + \lambda_a + \lambda_t + \lambda_i \times Quarter_t + \lambda_j \times Quarter_t + \lambda_a \times Quarter_t + u_{it},\end{aligned} \tag{18}$$

where $\log(MarketShare_{ijat})$ is the market share in market ij, airline a, and year-quarter t, $\log(p_{ijat})$ is the price, $W * OnTimeRatio_{ijat}$ is the first-order neighbors of the on-time ratio, $\log(Distance_i)$ is the log distance of the market, λ_{ij} is the market fixed effects, λ_a is the airline company fixed effects, λ_t is the year-quarter fixed effects, and u_{it} is the error term. The first-stage equations are as follows:

$$\begin{aligned}\log(p_{ijat}) &= \gamma_1 W \log(Distance_{ija}) \times Quarter_t + \gamma_2 W^2 \log(Distance_{ija}) \times Quarter_t + FE_{ijat} + v_{1,it}, \\ W \times OnTimeRatio_{ijat} &= \gamma_1 W \log(Distance_{ija}) \times Quarter_t + \gamma_2 W^2 \log(Distance_{ija}) \times Quarter_t + FE_{ijat} + v_{1,it},\end{aligned} \tag{19}$$

where $W \log(Distance_i)$ is the first-order network neighbors' simple arithmetic average distance, and $W^2 \log(Distance_i)$ is the second-order network neighbors' simple arithmetic average distance. FE_{ijat} is the abbreviation of various fixed effects and time trends, the same as in the first-stage regression equation.

I report the results in Table 6. In column (1), I conduct the OLS estimation. In column (2), I conduct the old Hausman IV estimation, in which the instrument is the first-order neighbor of the log distance of the airport pair. In column (3), I conduct the new Hausman IV estimation, in which the instruments are the first-order and second-order neighbors of the airport pair. Again, the coefficients on the main parameter of interest are much different across different settings. For example, the coefficient in column (2) is nearly twice as much in magnitude as that in column (3). The Hausman specification test for each pair of settings yields a p-value of 0.000. Thus, again, the traditional way of Hausman IV may produce a significant bias. In column (4), I use a model with $K^X = K^Z = 2$, and the results are not much different from those in column (3). The Hausman specification test also yields a p-value larger than 0.1. The weak-IV tests for columns (2) through (4) all yield a p-value less than 0.01, and the overidentification test for column (4) yields a p-value larger than 0.1. Thus, there is no weak IV and overidentification issue.

Some lessons can be learned from these two empirical analyses. First, it is indeed possible that cross-sectional dependence in the geographical space or the product space is unaccounted for, leading to biases in the traditional Hausman-style IV estimation. Second, in practice, we do not know a priori the order of unaccounted

[14] `https://www.transtats.bts.gov/DatabaseInfo.asp?QO_VQ=EFI&Yv0x=D`

[15] In this case, neighbors are associated with both the product space and the geographical space.

network dependence. Thus, we can try a data-driven approach to increase the order of neighbors used as IV to check whether the estimates vary with the order. From the two analyses above, the results become stable when the order of neighbors as used in IV is higher.

7 Conclusion

In this paper, I argue that when complex network dependence is an important mechanism at play, the traditional and widely used Hausman-style instrumental variable estimation in the industrial organization and trade literature may not be valid for causal identification. This is the case for migration networks in evaluating place-based labor market policies, and for correlated unobserved consumer tastes in the product space in demand estimation. However, the existing literature is almost silent about this source of bias and consistently uses it in econometric exercises.

To address this issue, I first build an economic model for the two cases of inter-regional migration networks and correlated consumer tastes in the product space, respectively, to shed light on the econometric framework. With the help of the economic models, I argue that network dependence, in theory, and the resulting omitted variable bias, in practice, can be a mechanism at play. Next, I establish an econometric model, and I show that not correctly accounting for such network dependence leads to biases and propose a new approach of estimation that exploits higher-order network neighbors. Finally, I conduct Monte Carlo simulations and empirical analyses to validate this new approach of estimation.

References

[1] Azar, José A, Steven T Berry, and Ioana Marinescu. 2022. "Estimating labor market power." Tech. rep., National Bureau of Economic Research.

[2] Berry, Steven, Amit Gandhi, and Philip Haile. 2013. "Connected substitutes and invertibility of demand." *Econometrica* 81 (5):2087–2111.

[3] Berry, Steven T. 1994. "Estimating discrete-choice models of product differentiation." *The RAND Journal of Economics* :242–262.

[4] Bramoullé, Yann, Habiba Djebbari, and Bernard Fortin. 2009. "Identification of peer effects through social networks." *Journal of econometrics* 150 (1):41–55.

[5] Caliendo, Lorenzo, Maximiliano Dvorkin, and Fernando Parro. 2019. "Trade and labor market dynamics: General equilibrium analysis of the china trade shock." *Econometrica* 87 (3):741–835.

[6] Cragg, John G and Stephen G Donald. 1993. "Testing identifiability and specification in instrumental variable models." *Econometric Theory* 9 (2):222–240.

[7] Crawford, Gregory S and Ali Yurukoglu. 2012. "The welfare effects of bundling in multichannel television markets." *American Economic Review* 102 (2):643–685.

[8] Hansen, Lars Peter. 1982. "Large sample properties of generalized method of moments estimators." *Econometrica: Journal of the econometric society* :1029–1054.

[9] Hausman, Jerry, Gregory Leonard, and J Douglas Zona. 1994. "Competitive analysis with differenciated products." *Annales d'Economie et de Statistique* :159–180.

[10] Hausman, Jerry A. 1978. "Specification tests in econometrics." *Econometrica: Journal of the econometric society* :1251–1271.

[11] Houde, Jean-François. 2012. “Spatial differentiation and vertical mergers in retail markets for gasoline.” *American Economic Review* 102 (5):2147–2182.

[12] Lanoie, Paul, Jérémy Laurent-Lucchetti, Nick Johnstone, and Stefan Ambec. 2011. “Environmental policy, innovation and performance: new insights on the Porter hypothesis.” *Journal of Economics & Management Strategy* 20 (3):803–842.

[13] Nevo, Aviv. 2001. “Measuring market power in the ready-to-eat cereal industry.” *Econometrica* 69 (2):307–342.

[14] Sanderson, Eleanor and Frank Windmeijer. 2016. “A weak instrument F-test in linear IV models with multiple endogenous variables.” *Journal of econometrics* 190 (2):212–221.

[15] Sargan, John D. 1958. “The estimation of economic relationships using instrumental variables.” *Econometrica: Journal of the econometric society* :393–415.

Table 1: Monte Carlo simulations

Model 1				
	Traditional Hausman-style estimation		New Hausman-style estimation	
N	Estimate of β_1	Bias	Estimate of β_1	Bias
50	0.592	0.092	0.517	0.017
100	0.593	0.093	0.516	0.016
200	0.598	0.098	0.522	0.022
500	0.612	0.112	0.523	0.023
1000	0.584	0.084	0.516	0.016
Model 2				
	Traditional Hausman-style estimation		New Hausman-style estimation	
N	Estimate of β_1	Bias	Estimate of β_1	Bias
50	0.611	0.111	0.523	0.023
100	0.584	0.084	0.523	0.023
200	0.595	0.095	0.518	0.018
500	0.614	0.114	0.516	0.016
1000	0.589	0.089	0.518	0.018
Model 3				
	Traditional Hausman-style estimation		New Hausman-style estimation	
N	Estimate of β_1	Bias	Estimate of β_1	Bias
50	0.582	0.082	0.524	0.024
100	0.590	0.090	0.520	0.020
200	0.582	0.082	0.523	0.023
500	0.595	0.095	0.517	0.017
1000	0.583	0.083	0.521	0.021

Table 2: Monte Carlo simulations: Asymptotic distribution

N	Prob($\|z_{\beta_1}\| < 1.645$)	Prob($\|z_{\beta_1}\| < 1.96$)	Prob($\|z_{\beta_1}\| < 2.33$)	Prob($\|z_{\beta_1}\| < 2.575$)
		Model 1		
50	0.8653	0.9768	0.9702	0.9860
100	0.9117	0.9304	0.9832	0.9866
200	0.8732	0.9489	0.9785	0.9907
500	0.8642	0.9433	0.9784	0.9906
1000	0.9225	0.9449	0.9871	0.9926
		Model 2		
50	0.9128	0.9419	0.9806	0.9878
100	0.9063	0.9418	0.9762	0.9921
200	0.9063	0.9358	0.9715	0.9899
500	0.8839	0.9680	0.9883	0.9885
1000	0.9200	0.9691	0.9851	0.9884
		Model 3		
50	0.9263	0.9590	0.9872	0.9929
100	0.8804	0.9511	0.9747	0.9918
200	0.8994	0.9563	0.9873	0.9882
500	0.9141	0.9592	0.9720	0.9858
1000	0.8869	0.9491	0.9850	0.9880

Table 3: More Monte Carlo Analysis

True K^X	True K^Z	Actually specified K^X	Actually specified K^Z	Number of IVs	Estimate of β_1	Bias
2	2	2	2	4	0.503	0.003
2	2	1	2	4	0.532	0.032
2	2	2	1	4	0.525	0.025
2	2	2	0	4	0.528	0.028
2	2	3	2	4	0.502	0.002
2	2	2	3	4	0.504	0.004
2	2	2	2	5	0.504	0.004
2	2	2	2	6	0.504	0.004
2	2	2	2	7	0.505	0.005
2	2	2	2	8	0.505	0.005
3	3	3	3	5	0.502	0.002
3	3	2	3	5	0.517	0.017
3	3	1	3	5	0.52	0.020
3	3	3	2	5	0.522	0.022
3	3	3	1	5	0.524	0.024
3	3	2	2	5	0.527	0.027
3	3	1	2	5	0.531	0.031
3	3	2	1	5	0.534	0.034
3	3	1	1	5	0.533	0.033
3	3	3	3	6	0.502	0.002
3	3	3	3	7	0.503	0.003
3	3	3	3	8	0.503	0.003

Table 4: Summary statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
Panel A: City Labor Market in China					
Primary share	5,985	3.781	7.730	0.01	95.580
Secondary share	6,022	40.654	14.986	1.403	97.375
Tertiary share	6,018	52.560	14.790	3.281	94.269
log per capita road area	4,563	2.043	0.668	-3.912	6.093
log distance to railroad	4,911	2.148	1.398	0	6.644
log distance to province capital	4,911	4.947	0.586	2.664	6.164
Panel B: US airline market					
log market share	444,779	-2.305	0.231	-3.455	-1.365
log price	444,779	5.449	0.283	-1.470	9.426
On-time rate	444,439	0.949	0.039	0.459	1
log market distance	444,788	7.164	0.661	2.398	8.912

Table 5: Estimating the effects of infrastructure on the labor market

Panel A: OLS

	(1)	(2)	(3)	(4)	(5)	(6)
	log(Primary Share)	log(Second Share)	log(Third Share)	log(Primary Share)	log(Second Share)	log(Third Share)
log(Per capita road area)	0.0668	-0.0379**	0.0158	0.0860**	-0.0365**	0.0201*
	(0.0430)	(0.0164)	(0.0115)	(0.0427)	(0.0169)	(0.0116)
City FE	Y	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y	Y
Observations	4,830	4,871	4,872	4,830	4,871	4,872
R-squared	0.841	0.806	0.804	0.846	0.801	0.805

Panel B: Old Hausman

	(1)	(2)	(3)	(4)	(5)	(6)
	log(Primary Share)	log(Second Share)	log(Third Share)	log(Primary Share)	log(Second Share)	log(Third Share)
log(Per capita road area)	8.392***	-1.764***	2.174***	5.187***	-1.710***	0.676***
	(0.831)	(0.263)	(0.188)	(0.942)	(0.399)	(0.260)
City FE	Y	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y	Y
Observations	4,830	4,871	4,872	4,830	4,871	4,872
R-squared	0.843	0.794	0.788	0.841	0.783	0.779

Panel C: New Hausman, $K^X = 1, K^Z = 1$

	(1)	(2)	(3)	(4)	(5)	(6)
	log(Primary Share)	log(Second Share)	log(Third Share)	log(Primary Share)	log(Second Share)	log(Third Share)
log(Per capita road area)	15.49***	-3.260***	4.018***	18.06***	-6.628***	6.308***
	(1.535)	(0.487)	(0.346)	(3.183)	(0.887)	(0.646)
City FE	Y	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y	Y
Observations	4,830	4,871	4,872	4,830	4,871	4,872
R-squared	0.844	0.794	0.788	0.848	0.792	0.792

Panel D: New Hausman, $K^X = 1$, $K^Z = 2$

	(1)	(2)	(3)	(4)	(5)	(6)
	log(Primary Share)	log(Second Share)	log(Third Share)	log(Primary Share)	log(Second Share)	log(Third Share)
log(Per capita road area)	14.85***	-3.052***	3.728***	13.35***	-4.617***	4.041***
	(1.462)	(0.470)	(0.330)	(1.340)	(0.362)	(0.267)
Observations	4,830	4,871	4,872	4,830	4,871	4,872
R-squared	0.843	0.794	0.788	0.848	0.791	0.791

Panel E: New Hausman, $K^X = 2, K^Z = 2$

	(1)	(2)	(3)	(4)	(5)	(6)
	log(Primary Share)	log(Second Share)	log(Third Share)	log(Primary Share)	log(Second Share)	log(Third Share)
log(Per capita road area)	15.41***	-3.229***	3.972***	13.62***	-4.795***	4.206***
	(1.520)	(0.482)	(0.343)	(1.371)	(0.362)	(0.273)
City FE	Y	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y	Y
Observations	4,830	4,871	4,872	4,830	4,871	4,872
R-squared	0.844	0.794	0.788	0.848	0.792	0.792

Notes: The sample covers about 4830 city-year cells during 1999-2018. The dependent variables are the log share of the labor force of the primary, secondary, and tertiary industries. In all columns city and year fixed effects are included. For Panel B, the instruments in columns (1) through (3) are first-order neighbors of log distance to the railroad; The instruments in columns (4) through (6) are first-order neighbors of log distance to the province capital. For Panels C through E, I use a new Hausman IV approach with different K^X and K^Z. * Significant at 10%, ** 5%, *** 1%.

Table 6: Demand estimation in airline market

	(1)	(2)	(3)	(4)
			log(Market Share)	
	OLS	Old Hausman IV	New Hausman IV ($K^X = 1, K^Z = 1$)	New Hausman IV ($K^X = 2, K^Z = 2$)
log(Price)	-0.0765***	-1.437***	-0.692***	-0.599***
	(0.00164)	(0.0383)	(0.0378)	(0.0445)
On-time ratio	-2.672***	-2.633***	-9.710***	-12.256***
	(0.00884)	(0.00878)	(0.558)	(0.717)
log(Distance)	-0.0349***	-0.0475***	-0.0649***	-0.0649***
	(0.000686)	(0.000569)	(0.000614)	(0.000614)
Airport Pair FE	Y	Y	Y	Y
Year-quarter FE	Y	Y	Y	Y
Origin Airport FE*Year-quarter	Y	Y	Y	Y
Destination Airport FE*Year-quarter	Y	Y	Y	Y
Airline FE*Year-quarter	Y	Y	Y	Y
Observations	444,438	444,438	444,779	444,779
R-squared	0.584	0.596	0.465	0.465

Notes: The sample covers 444,438 airport-pair-quarter-airline cells during 2008Q1-2018Q4. The dependent variable is the market share of a certain airline in the market of an airport pair. In all columns airport pair and year-quarter fixed effects, airline fixed effects interacted with linear year-quarter trends, origin airport fixed effects interacted with linear year-quarter trends, and destination airport fixed effects interacted with linear year-quarter trends are controlled for. The instrument in column (2) is the first-order neighbor of the log distance of the airport pair. The instruments in column (3) are the first-order and second-order neighbors of the airport pair. Neighbors are defined as airline markets that share an origin or a destination airport. * Significant at 10%, ** 5%, *** 1%.

Online Appendix

Appendix A Details of the Economic Models

A.1 Proofs in the Model of Migration Networks

The migration choice problem is formulated as follows:

$$v_i = A_i \log C_i + \max_{j \in N(i)} (v_j - \tau_{i,j} + \nu_j \epsilon_j), \tag{A1}$$

We assume that the idiosyncratic preference shock ϵ is i.i.d. and is a realization of a Type-I Extreme Value distribution with zero mean. The c.d.f of ϵ is $F(\epsilon) = \exp(-\exp(-\epsilon - \bar{e}))$, where $\bar{e}$ is the Euler constant. The p.d.f. is $f(\epsilon) = \frac{\partial F}{\partial \epsilon}$. We are trying to solve for

$$\Phi_i = E[\max_{j \in N(i)} (E(v_j) - \tau_{i,j} + \nu_j \epsilon_j)]. \tag{A2}$$

Let $\bar{\epsilon}_{i,j} = ((v_i - v_j) - (\tau_{i,h} - \tau_{j,h}))/\mu_i$, we have

$$\Phi_i = \sum_{j \in N(i)} \int_{-\infty}^{\infty} (V_j - \tau_{i,j} + \nu_j \epsilon_j) f(\epsilon_j) \prod_{h \neq j, h \in N(i)} F(\bar{\epsilon}_{j,h} + \epsilon_j) d\epsilon_j. \tag{A3}$$

Defining $\lambda_i = \log(\sum_{j \in N(i)} \exp(-\bar{\epsilon}_{i,j}))$ and considering the change of variables $\xi_i = \bar{e} + \epsilon_i$, we get

$$\Phi_i = \sum_{j \in N(i)} \int_{-\infty}^{\infty} (V_j - \tau_{i,j} + \nu_j(\epsilon_j - \bar{e})) \exp(-\xi_j - \exp(-(-\xi_j - \lambda_j))) d\xi_j. \tag{A4}$$

We conduct an additional change of variables: let $y_i = \xi_i - \lambda_i$, we have

$$\Phi_i = \sum_{j \in N(i)} \exp(-\lambda_i)(V_j - \tau_{i,j} + \nu_j(\lambda_j - \bar{e}) + \int_{-\infty}^{\infty} \mu_j y_j \exp(-y_j - \exp(-y_j)) dy_j. \tag{A5}$$

Using the definition of Euler constant, $\bar{e}$, we have

$$\Phi_i = \sum_{j \in N(i)} \exp(-\lambda_j)(V_j - \tau_{i,j} + \nu_j \lambda_j). \tag{A6}$$

Plug into λ, we have

$$\Phi_i = \nu_i (\log \sum_{j \in N(i)} \exp(V_j - \tau_{i,j})^{1/\nu_j}). \tag{A7}$$

Therefore, we have equation (2). Define $\mu_{j,i}$ as the fraction of households that relocate from j to i, that is, the probability that the expected utility of moving to i is higher than the expected utility in any other location, we have that

$$\mu_{j,i} = Prob(V_i - \tau_{i,j} + \nu_j \epsilon_j \geq \max_{k \in N(j)} V_k - \tau_{k,j} + \nu_k \epsilon_k). \tag{A8}$$

Given our assumptions on the idiosyncratic preference shock, we obtain

$$\mu_{j,i} = \int_{-\infty}^{\infty} f(\epsilon_i) \prod_{h \neq i, h \in N(j)} F((V_i - V_h) - (\tau_{i,j} - \tau_{h,j}) + \epsilon_i) d\epsilon_i. \tag{A9}$$

From the above derivations, and using the definitions above, we have that

$$\mu_{j,i} = \exp(-\lambda_i) \int_{-\infty}^{\infty} \exp(-y_i - \exp(-y_i)) dy_i. \tag{A10}$$

Solving for this integral, we have equation (3).

A.2 Full Model of Correlated Consumer Tastes in Product Space

We start by defining the utility function of consumer i: $U(x_j, \xi_j, p_j, v_i, z_k)$, where x_j is the observed product characteristics, ξ_j is the unobserved product characteristics, p_j is the price, z_k is other exogenous observed characteristics of other goods than j used to construct the Hausman-style instrument, v_i is the consumer's characteristic that is not observed by econometricians. Same as Berry (1994), I focus on a random coefficients specification of utility, in which the utility u_{ij} that consumer i has for product j is

$$u_{ij} = x_j \tilde{\beta}_{1i} - \alpha p_j + \sum_{k \in N(j)} w_{jk} z_k \beta_2 + \xi_j + \epsilon_{ij}, \tag{A11}$$

where the (unobserved to econometricians) consumer-specific taste parameters are $\tilde{\beta}_{1i}$ and ϵ_{ij}. For simplicity, consider the following decomposition for characteristic h:

$$\tilde{\beta}_{1ih} = \beta_{1h} + \sigma_h \zeta_{ih}, \tag{A12}$$

where β_{1h} is the mean level of $\tilde{\beta}_{1ih}$, and ζ_{ih} is normally distributed with mean zero. Therefore, we have

$$u_{ij} = x_j \beta_1 + \sum_{k \in N(j)} w_{jk} z_k \beta_2 - \alpha p_j + \xi_j + \nu_{ij}, \tag{A13}$$

where $\nu_{ij} = \sum_h x_{jh} \sigma_h \zeta_{ih} + \epsilon_{ij}$.

The term ν_{ij} has mean zero, and captures the effects of random taste parameters. The mean utility level of product j is thus expressed in equation (7). Assuming that ϵ_{ij} is i.i.d. distributed with Type-I extreme value distribution with mean zero, we have the market share equation as follows:

$$s_j(\delta_j) = \frac{\exp(\delta_j)}{\sum_k \exp(\delta_k)}, \tag{A14}$$

and the estimating equation (8).

Appendix B Asymptotic Properties

B.1 Consistency

I first discuss the consistency of the new Hausman IV estimator. Since it belongs to the extremum estimators category, consistency is based on two assumptions: Assumption B1 (Uniform Weak Convergence) and B2 (Identifiable Uniqueness). These two assumptions are as follows:

Assumption B1. *(Assumption U-WCON)* $\sup_{\Delta \in D} |Q_N(\Delta) - Q(\Delta)| \to_p 0$ *for some nonstochastic function* $Q(\Delta)$.

"U-WCON" stands for uniform weak convergence. Assumption B1 is a high-level assumption for the asymptotic consistency of an extremum estimator. However, I can show below that such an assumption can be validated for the new Hausman IV proposed above, in the next subsection. The sample and population objective function $Q_N(\Delta)$ and the $Q(\Delta)$ function of the Hausman IV is as follows.

$$\begin{aligned} Q_N^{IV}(\Delta) = (\frac{1}{N}\sum_i (Y_i - X_i\beta_1 - Z_i\beta_2 - \sum_{p=1}^{K^X} W^p X_i \gamma_p^X - \sum_{q=1}^{K^Z} W^q Z_i \gamma_q^Z)' \\ A_N' A_N (\frac{1}{N}\sum_i (Y_i - X_i\beta_1 - Z_i\beta_2 - \sum_{p=1}^{K^X} W^p X_i \gamma_p^X - \sum_{q=1}^{K^Z} W^q Z_i \gamma_q^Z))/2, \end{aligned} \tag{B1}$$

where

$$A_N = \sum_i V_i (V_i' V_i)^{-1} V_i', \tag{B2}$$

where V_i is a vector of instruments (higher-order neighbors' average exogenous characteristics).

$$V_i = (W^{\max\{K^X, K^Z\}+1} Z_i, ..., W^{\max\{K^X, K^Z\}+K^X} Z_i). \tag{B3}$$

Then,

$$Q^{IV}(\Delta) = ||AE((Y_i - X_i\beta_1 - Z_i\beta_2 - \sum_{p=1}^{K^X} W^p X_i \gamma_p^X - \sum_{q=1}^{K^Z} W^q Z_i \gamma_q^Z))||^2/2, \tag{B4}$$

where $A_N \to_p A$.

The new Hausman IV estimator is then given by:

$$\hat{\Delta}_{IV} = \arg\max_{\Delta} Q_N^{IV}(\Delta). \tag{B5}$$

In addition, the result of consistency relies on the other assumption (identifiable uniqueness) as follows:

Assumption B2. *(Assumption ID1) (1)* D_0 *is compact; (2)* $Q(\Delta)$ *is continuous in* D*; (3)* Δ_0 *uniquely minimizes* $Q(\Delta)$ *over* $\Delta \in D$.

According to standard econometric textbooks, Assumption U-WCON and ID1 can be verified with the new Hausman IV estimator. The details of the verification are shown in the next subsection.

Thus, we have the following Proposition B1. The proof of the proposition is presented in the next subsection.

Proposition B1. *With Assumption U-WCON and Assumption ID1, the new Hausman IV estimator satisfies* $\hat{\Delta} \to_p \Delta$.

B.2 Asymptotic Normality

The results of asymptotic normality are also based on two assumptions: Assumption B3 (Assumption CF) and Assumption B4 (Assumption EE2). The Assumption CF is as follows:

Assumption B3. *(Assumption CF) (1)* Δ_0 *is in the interior of the parameter space D; (2)* $Q_N(\Delta)$ *is twice continuously differentiable on some neighborhood* $D_0 \subseteq D$ *of* Δ_0 *with probability one; (3)* $\sqrt{N}\frac{\partial Q_N(\Delta)}{\partial \Delta} \to_d N(0, \Omega_0)$*; (4)* $\sup_{\Delta \in D_0} ||\frac{\partial^2 Q_N(\Delta)}{\partial \Delta \partial \Delta'} - B(\Delta)|| \to_p 0$ *for some non-stochastic* $d \times d$ *matrix-valued function* $B(\Delta)$ *that is continuous at* Δ_0 *and for which* $B_0 = B(\Delta_0)$ *is non-singular.*

Assumption B3 is a high-level assumption for the asymptotic normality of an extremum estimator. However, we can show below that such an assumption can be validated for each estimator proposed, in the next subsection.

For the new Hausman IV estimator, the $Q_N(\Delta)$, $B(\Delta)$ function, and the B_0, Ω_0 matrix is as follows:

$$\begin{aligned} Q_N^{IV}(\Delta) = (\frac{1}{N}\sum_i (Y_i - X_i\beta_1 - Z_i\beta_2 - \sum_{p=1}^{K^X} W^p X_i \gamma_p^X - \sum_{q=1}^{K^Z} W^q Z_i \gamma_q^Z)' \\ A_N' A_N (\frac{1}{N}\sum_i (Y_i - X_i\beta_1 - Z_i\beta_2 - \sum_{p=1}^{K^X} W^p X_i \gamma_p^X - \sum_{q=1}^{K^Z} W^q Z_i \gamma_q^Z)/2, \end{aligned} \tag{B6}$$

where A_N is defined in equation (B10) and,

$$\Omega_0 = \Gamma_0' A' A M_0 A' A \Gamma_0, \tag{B7}$$

$$M_0 = E((Y_i - X_i\beta_1 - Z_i\beta_2 - \sum_{p=1}^{K^X} W^p X_i \gamma_p^X - \sum_{q=1}^{K^Z} W^q Z_i \gamma_q^Z)(Y_i - X_i\beta_1 - Z_i\beta_2 - \sum_{p=1}^{K^X} W^p X_i \gamma_p^X - \sum_{q=1}^{K^Z} W^q Z_i \gamma_q^Z)'). \tag{B8}$$

and here V_i is a vector of instruments (higher-order neighbors' average exogenous characteristics).

$$V_i = (W^{\max\{K^X, K^Z\}+1} Z_i, ..., W^{\max\{K^X, K^Z\}+K^X} Z_i). \tag{B9}$$

In addition,

$$A_N = \sum_i V_i (V_i' V_i)^{-1} V_i', \tag{B10}$$

and,

$$\Gamma_0 = E(\frac{\partial (Y_i - X_i\beta_1 - Z_i\beta_2 - \sum_{p=1}^{K^X} W^p X_i \gamma_p^X - \sum_{q=1}^{K^Z} W^q Z_i \gamma_q^Z)}{\partial \Delta'}), \tag{B11}$$

and,

$$B_0 = \Gamma_0' A' A \Gamma_0. \tag{B12}$$

The asymptotic normality results are also based on the other assumption, Assumption EE2.

Assumption B4. *(Assumption EE2) (1)* $\hat{\Delta} \rightarrow_p \Delta_0$*; (2)* $\frac{\partial Q(\hat{\Delta})}{\partial \Delta} = o_p(N^{-1/2})$.

According to standard econometric textbooks, the Assumption CF and EE2 can be verified for the Hausman IV estimator. The details of the verification are shown in the next subsection. Given these two assumptions, we can derive Proposition B2. The proof of the proposition is presented in the next subsection.

Proposition B2. *Under Assumption CF and Assumption EE2:* $\sqrt{N}(\hat{\Delta} - \Delta_0) \rightarrow_d N(0, B_0^{-1}\Omega_0 B_0^{-1})$.

B.3 Further Details of Asymptotic Properties

B.3.1 Verifying Assumption B1 and B2

The verification of Assumption B1 can be done by using the weak law of large numbers. To verify Assumption B2, for Hausman IV estimator: If A is nonsingular and there exists a unique value $\Delta_0 \in D$ such that $Eh(X, D, \Delta_0) = 0$, then Δ_0 uniquely minimizes $Q^{GMM}(\Delta)$. If not, then Δ_0 is the value that minimizes $Q^{GMM}(\Delta) = ||AE((Y_i - X_i\beta_1 - Z_i\beta_2 - \sum_{p=1}^{K^X} W^p X_i \gamma_p^X - \sum_{q=1}^{K^Z} W^q Z_i \gamma_q^Z))||^2/2$.

B.3.2 Proof of Theorem B1

We first prove the following lemma.

Lemma B1. *If Assumption B2 holds, then Assumption B5 (Assumption ID) holds, which is the following.*

Assumption B5. *(Assumption ID) There exists* $\Delta_0 \in D$ *such that* $\forall \epsilon$, $\inf_{\Delta \notin B(\Delta_0,\epsilon) Q(\Delta) > Q(\Delta_0)}$.

Proof of Lemma B1: By contradiction. □

Given Lemma B1, we can prove Theorem B1.

Proof of Theorem B1: By Assumption B5, given any $\epsilon > 0$, $\exists \delta$ such that $\Delta \notin B(\Delta_0, \epsilon)$ implies $Q(\Delta) - Q(\Delta_0) \geq \delta > 0$. Thus, $Pr(\hat{\Delta} \notin B(\Delta_0, \epsilon)) \leq Pr(Q(\hat{\Delta}) - Q(\Delta_0) \geq \delta) = Pr(Q(\hat{\Delta}) - Q_N(\Delta_0) + Q_N(\Delta_0) - Q(\Delta_0) \geq \delta) \leq Pr(Q(\hat{\Delta}) - Q_N(\Delta_0) + Q_N(\hat{\Delta}) + o_p(1) - Q(\Delta_0) \geq \delta) \leq Pr(2 \sup_{\Delta \in D} |Q(\Delta) - Q_N(\Delta) + o_p(1)| \geq \delta) \rightarrow_p 0$. The second inequality holds because $\hat{\Delta}$ minimizes $Q_N(\Delta)$. The convergence to zero holds by Assumption U-WCON. □

B.3.3 Verifying Assumption B3 and B4

To verify Assumption B3(2) for the new Hausman IV estimator, we use the twice continuous differentiability of the linear functions.

To verify Assumption B3(3) and (4) for the Hausman IV estimator, we use the Central Limit Theorem and Weak Law of Large Numbers.

To verify Assumption B4, we use the result of Theorem B1.

B.3.4 Proof of Theorem B2

Before proving Theorem B2, we first establish Lemma B2.

Lemma B2. *Suppose (1)* $\hat{\Delta} \to_p \Delta_0$*, (2)* $\sup_{\Delta \in B(\Delta_0,\epsilon)} |L_N(\Delta) - L(\Delta)| \to_p 0$ *for some stochastic functions* $L_N(\Delta)$*, some non-stochastic function* $L(\Delta)$*, and some* $\epsilon > 0$*, and (3)* $L(\Delta)$ *is continuous at* Δ_0*. Then,* $L_N(\Delta) \to L(\Delta)$*.*

Proof of Lemma B2: $|L_N(\hat{\Delta}) - L(\Delta_0)| = |L_N(\hat{\Delta}) - L(\hat{\Delta}) + L(\hat{\Delta}) - L(\Delta_0)| \leq |L_N(\hat{\Delta}) - L(\hat{\Delta})| + |L(\hat{\Delta}) - L(\Delta_0)| \leq \sup_{\Delta \in B(\Delta_0,\epsilon)} |L_N(\Delta) - L(\Delta)| + |L(\hat{\Delta}) - L(\Delta_0)| \to_p 0$, where the first inequality holds by the triangle inequality, the second inequality holds with probability 1 because $\hat{\Delta} \in B(\Delta_0, \epsilon)$ with probability 1 by (1), and the convergence to zero holds using (1), (2), (3), and the Slutsky's Theorem. □

Proof of Theorem B2: Using CF(2) and EE(2), element-by-element mean value expansions of $\frac{\partial Q_N(\hat{\Delta})}{\partial \Delta}$ about Δ_0 yield

$$o_p(N^{-1/2}) = \frac{\partial Q_N(\hat{\Delta})}{\partial \Delta} = \frac{\partial Q_N(\Delta_0)}{\partial \Delta} + \frac{\partial^2 Q_N(\Delta^*)}{\partial \Delta \partial \Delta'}(\hat{\Delta} - \Delta_0), \tag{B13}$$

where Δ^* lies between $\hat{\Delta}$ and Δ_0 (thus $\Delta^* \to_p \Delta_0$). By Lemma B2, CF(2), CF(4), and EE2(1),

$$\frac{\partial^2 Q_N(\Delta^*)}{\partial \Delta \partial \Delta'} = B_0 + o_p(1). \tag{B14}$$

From the above two equations, we have

$$o_p(1) = \sqrt{N}\frac{\partial Q_N(\Delta_0)}{\partial \Delta} + (B_0 + o_p(1))\sqrt{N}(\hat{\Delta} - \Delta_0). \tag{B15}$$

Rearranging yields

$$\sqrt{n}(\hat{\Delta} - \Delta_0) = -(B_0 + o_p(1))^{-1}\sqrt{n}\frac{\partial Q_N(\Delta_0)}{\partial \Delta} + o_p(1) \to_d N(0, B_0^{-1}\Omega_0 B_0^{-1}), \tag{B16}$$

using Assumptions CF(3) and CF(4). □