Identification of Social Effects with Endogenous Networks and Covariates: Theory and Simulations

Santiago Pereda-Fernández[1] and Paolo Zacchia[2]

[1]Banca d'Italia

[2]IMT School for Advanced Studies Lucca

July 2018

Abstract

The estimation of spillover and peer effects presents challenges that are still unsolved. In fact, even if separate algebraic identification of the endogenous and exogenous effects is possible, these might be contaminated by the simultaneous dependence of outcomes, covariates and the network structure upon spatially correlated unobservables. In this paper we characterize the identification conditions for consistently estimating all the parameters of a spatially autoregressive or linear-in-means model in presence of linear forms of endogeneity. We show that identification is possible if the network of social interactions is non-overlapping up to three degrees of separation, and the spatial matrix that characterizes the co-dependence of individual covariates and peers' unobservables is known to the econometrician. We propose a GMM approach for the estimation of the model's parameters, and we evaluate its performance through Monte Carlo simulations.

JEL Classification Codes: C21, C31, D85

Keywords: Peer Effects, Networks, Network Endogeneity, Common Shocks

1 Introduction

A sizable body of empirical economic research deals with the analysis of peer effects, network effects, social interactions and – more generally – externalities. Among these studies, the literature about peer effects in education occupies perhaps a more prominent position (Sacerdote, 2001; Calvó-Armengol et al., 2009; De Giorgi et al., 2010; Carrell et al., 2013), but applications in more diverse fields are numerous (Glaeser et al., 1996; Duflo and Saez, 2003; Mas and Moretti, 2009).[1] In the face of growing empirical evidence, econometric analysis has struggled for a while to characterize conditions for providing a more structural interpretation to observed group correlations in outcomes. While important advances have been realized, their relevance is limited to a restricted set of empirical settings, in which the characteristics of individuals as well as their structure of socio-economic interactions are both as good as exogenous.

In order to better characterize the position of the present article in this literature, it is worth to summarize the evolution of economists' understanding of the workhorse framework for many studies about social effects: the "linear-in-means" model. Following Manski (1993), who highlighted the so-called *reflection* problem of simultaneity between group characteristics and group outcomes, econometricians have attempted to individuate settings in which endogenous social responses can be identified separately from the influence of common external factors. In an influential contribution, Bramoullé et al. (2009) express the conditions for identification when social effects are shaped by networked structures of interaction, which is particularly appealing as networks typically provide more realistic descriptions of actual social relationships. Blume et al. (2015) incorporate their identification results – as well as one based on covariance restrictions which builds on Graham (2008) – within a larger framework. Thanks to these and other efforts, it is now well understood that complex patterns of individual dependence if anything make the identification of social effects *easier*. Yet all these analyses maintain the assumption that the model's error term is conditionally independent of the individual covariates and the structure of interactions.

[1] Studies of R&D and knowledge spillovers more generally, which follow the tradition initiated by Jaffe (1986, 1989), are seldom counted among these studies. This is quite a notable omission, since the workhorse econometric frameworks employed in this literature are easily seen as variations of the standard spatial models utilized for the estimation of peer effects. More recent contributions about R&D spillovers include Bloom et al. (2013), Lychagin et al. (2016) and Zacchia (2018). Other related strands of literature include the one about peer effects in scientific production (Azoulay et al., 2010; Waldinger, 2011; Oettl, 2012) and that about learning externalities (Conley and Udry, 2010).

By contrast, in this paper we examine a model of social interactions where both individual characteristics and the network that defines paired relationships are simultaneously dependent on individual unobservables. Our departure point is a Spatially Autoregressive (SAR) model (Cliff and Ord, 1981), of which the linear-in-means model is a special case, and whose econometrics was analyzed in a number of papers (Lee, 2007a,b; Lee et al., 2010; Lin and Lee, 2010; Liu and Lee, 2010; Lee and Liu, 2010). Like other articles from this literature, we derive our empirical model from an explicit theoretical framework; unlike most, ours is based on a Cobb-Douglas utility function, and it can accommodate contexts ranging from peer effects in the classroom to R&D spillovers. We discuss how standard estimates of social effects are inconsistent if unobservables correlate with covariates *and* with peers' unobservables, and we illustrate the related identification problem by showing that the errors' cross-correlation is observationally equivalent to typical exogenous or "contextual" effects often featured in linear-in-means models. This resonates with the critique of peer effect studies that is put forward by Angrist (2014), according to whom the current results in the literature may reflect spurious correlations due to "correlated effects."

The main contribution of our paper is to show that even in presence of endogeneity, under certain conditions the social effects are internally identified without resorting to external instruments. We analyze a scenario where the observable characteristics of agents depend linearly on both their own unobservables and on those of other agents: for example, the learning environment of students may depend on the ability of their peers, or the production inputs of firms might respond to the technical innovations of their competitors. We place no substantive restriction upon the spatial matrices that characterize this form of endogeneity, except that they are known to the econometrician up to a multiplicative parameter – which is identified – that quantifies the extent of endogeneity. The main identifying assumption extends those by Bramoullé et al. (2009), as it requires that the structure of social interactions is non-overlapping up to three degrees of separation in network space. This allows to construct an additional moment for the estimation of the structure of spatial correlation between individual characteristics and individual unobservables. We propose a GMM approach for the joint estimation of both the social effects and the other parameters of the model, and we evaluate its performance through some Monte Carlo simulations.

The one we propose is a novel approach in the literature. In line with a recommendation given by Blume et al. (2015), some scholars (Arduini et al., 2015; Johnsson and

Moon, 2017) develop a control function approach to account for network endogeneity. These contributions embed, within a SAR model, the network formation stage and its estimation, which is based on Graham (2017). Nevertheless, this does not account for simultaneity between individual and peers' characteristics (both observable and unobservable), which can be due to correlated shocks or to structural co-dependence. By contrast, our strategy does not require an explicit model of network formation; in this respect, our method extends some previous work by Zacchia (2018).[2] Obviously, the spatial econometrics literature has examined correlated unobservables at length (Kelejian and Prucha, 1998, 2007, 2010; Kapoor et al., 2007; Drukker et al., 2013); however, individual covariates are usually assumed exogenous in these studies.[3]

It is useful to relate our article to other papers from the literature about peer and network effects. In addition to the cited contribution by Graham (2008), other papers make use of conditional covariance restrictions to achieve the identification of social effects (Glaeser et al., 1996; Moffitt, 2001; Davezies et al., 2009; Pereda Fernández, 2017; Rose, 2017a). Our method also exploits some covariance restrictions, but unlike these papers, their role in identification is to disentangle the autonomous covariance structure of the error term from that of individual covariates, if the two are correlated. Other contributions develop methods for estimating unknown structures of interaction (Manresa, 2017; Rose, 2017b; De Paula et al., 2018) by using penalized estimators such as the LASSO (Tibshirani, 1996). These procedures cannot be extended to our setup as endogeneity is an additional source of cross-correlation in the outcomes. However, they might be adapted for the purpose of recovering the "endogeneity matrices" that relate individual covariates to peers unobservables, if those are unknown. We revisit this observations in the conclusion of the paper while suggesting future lines of work.

The remainder of this paper is organized as follows. Section 2 presents our general analytical framework. Section 3 discusses the conditions for the identification of social effects. Section 4 characterizes the GMM estimator and its asymptotic properties. Section 5 demonstrates its performance in Monte Carlo simulation. Finally, Section 6 concludes the paper. An Appendix provides mathematical proofs of our main results.

[2]Zacchia (2018) analyzes a model of R&D spillovers in which firms' unobservables are correlated in the network of R&D relationships, and are simultaneous to the R&D of connected firms. In order to identify spillover effects, he constructs IVs motivated on the finite empirical spatial correlation of R&D. The framework presented here instead does not restrict the spatial correlation of covariates.

[3]In a recent contribution, Kuersteiner and Prucha (2018) examine a SAR model for panel data in which the interaction matrix is possibly endogenous and covariates are weakly exogenous, and propose an appropriate GMM estimator. In our cross-sectional framework covariates are endogenous.

2 Analytical Framework

In this section we introduce the social interactions game that constitutes the theoretical framework of this paper. We subdivide this section between the description of the model's setup and the discussion of the equilibrium predictions.

2.1 Model's Setup

We consider an abstract setting of social and economic interactions between heterogeneous agents (players) in a network. In order to allow for interdependence between the characteristics of agents and the structure of their connections, we allow nature to randomly draw the *weighted network* $(\mathcal{I}, \mathcal{G})$ that characterizes the social interactions. Here, $\mathcal{I}$ is the set that comprises the N players, who are indexed as $i = 1, \ldots, N$. The N^2-dimensional set $\mathcal{G}$, instead, represents the interaction structure: thus, $g_{ij} \in \mathbb{R}$ denotes the relative strength of the influence exerted by player j on player i (and vice versa). We impose two standard normalizations: $g_{ij} \in [0, 1]$ and $g_{ii} = 0$ for all players $i = 1, \ldots, N$. Otherwise, we force no particular structure of the network: we generally allow for asymmetric, directed networks such that for any pair $(i, j) \in \mathcal{I}^2$, the weight g_{ij} implies no restriction upon the weight g_{ji}, and vice versa.

Every player in $\mathcal{I}$ is typified by two variables (x_i, ε_i). We denominate $x_i \in \mathcal{X}$ the *observable characteristics* of player i, and $\varepsilon_i \in \mathcal{E}$ his or her *unobservables*: this abstract terminology clearly relates to the information which is available to the econometricians who are in search of social externalities. For simplicity we set $\mathcal{X} = \mathcal{E} = \mathbb{R}$, although many, possibly discrete characteristics could easily be accommodated. We assume that the random vector of individual observable characteristics $\boldsymbol{x} = (x_1, \ldots, x_N)$, the random vector of individual unobservables, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_N)$ and the network $\mathcal{G}$ are all randomly drawn from a joint probability distribution $\mathcal{F}(\boldsymbol{x}, \boldsymbol{\varepsilon}, \mathcal{G})$, which is known by all agents. We place no a priori restrictions on the distribution $\mathcal{F}(\cdot)$.

The economic content of the description outlined thus far deserves some further discussion. In social networks, the probability of a connection occurring between any two agents is documented to be correlated with their characteristics. For example, friends usually sort on social background and demographics, while R&D spillovers naturally occur between firms belonging to the same technological class. This result is predicted by many models of random and strategic network formation, and the social mechanism by which similar agents are paired to one another bears the name of homophily. However, it is apparent that many of the characteristics that predict the occurrence (or the relative strength) of connections are unobserved by researchers: for example, student friendships may be sorted on ability; likewise, R&D connections may appear more frequently between firms with shared technologies. In either case, the fact that connected agents share some of their unobservables poses identification problems to the econometrician. Zacchia (2018) discusses in more detail how "common unobservables" and "network endogeneity" are two intertwined issues.[4]

Players maximize the following "twice exponential" utility function:

$$U_i\left(e_1,\ldots,e_N;x_i,\varepsilon_i\right)=\exp\left[y_i\left(e_1,\ldots,e_N;x_i,\varepsilon_i\right)\right]-\exp\left(e_i\right) \tag{1}$$

where y_i is the individual-level *outcome* (denoting, say, grades, or production output) which is determined through the following linear relationship.

$$y_i\left(e_1,\ldots,e_N;x_i,\varepsilon_i\right)=\alpha_0+\gamma_0 x_i+\mu e_i+\nu\sum_{i=1}^{N}g_{ij}e_j+\varepsilon_i \tag{2}$$

The outcomes of individuals depend upon their characteristics (x_i,ε_i) as well as on a costly strategic variable $e_i\in\mathbb{R}$ that we call *effort*: this may represent, for instance, time dedicated to homework or R&D investment. Because of social interactions and externalities, y_i also depends on the effort of all the other players an agent is connected to (possibly negatively). Private and social effects of effort are parametrized as μ and ν, respectively. To make the model realistic, we impose the following restriction.

Assumption 1. Concavity: $\mu\in(0,1)$

This assumption makes *i.* individual output positively dependent on effort, and *ii.* the utility function concave in $\exp(e_i)$, so that choice trade-offs are cogent. As we discuss later, additional restrictions on ν are necessary to ensure equilibrium uniqueness.

Note two differences between this framework and the quadratic utility model that is typical of the peer effects literature (Calvó-Armengol et al., 2009; Blume et al., 2015). First, the model proposed here outlines a clear distinction between individual choice variables and ultimate outcomes, which undoubtedly gives it more generality.

[4]In addition, in an extension of his theoretical framework (which is tailored to the setting of R&D spillovers) he observes that common models of network formation imply little or no cross-correlation of individual variables at three degrees of separation or more, which is in remarkable agreement with the empirical evidence.

Second, while both utility functions are globally concave in their respective strategic variables, in the case of (1) individual characteristics x_i, ability ε_i, individual effort e_i and the effort e_j of connections are complements. In addition to accommodating functional forms such as those that are typical of production functions, this increases the degree of realism of the model even in other social contexts. For example, more skilled or better supported students may benefit relatively more from devoting more time to homework and independent study, either alone or with their friends. Finally, observe that one could easily introduce heterogeneous weights to the benefit and cost components of (1), but this is beyond the point of the present analysis.

We analyze a game of complete information characterized by the following timing.

1. Nature draws $(\boldsymbol{x}, \boldsymbol{\varepsilon}, \mathcal{G})$ from $\mathcal{F}(\cdot)$. Every player observes the result of this draw.
2. Players simultaneously make their effort choices, and utilities are determined accordingly.

By letting the network be generated randomly by nature we abstract from the specifics of the network formation process, as our results do not depend on it. Also note that by assuming complete information we make our analysis more general. As discussed by Zacchia (2018), in fact, incomplete information provides more avenues for the identification of social effects, in the form of implicit restrictions on the cross-correlation of strategic variables.

2.2 Analysis

We analyze the properties of the equilibrium conditional upon a restriction, that we maintain throughout our discussion, about the combined parameter $\beta \equiv \nu (1 - \mu)^{-1}$.

Assumption 2. Non-explosiveness: $|\beta| \cdot \max_{i \in \mathcal{I}} \sum_{j=1}^{N} g_{ij} \in [0, 1)$

This assumption imposes that social effects do not "dominate" the process of outcome generation. In the game, it ensures uniqueness of the equilibrium by ruling out unrealistically "explosive" scenarios. In statistical terms, this assumption makes it possible that the variation of y_i is not predominantly explained by the cross-correlation of outcomes in the network: we find that otherwise, the identification problems discussed in this article are largely moot, since standard estimators would capture the social effects with little bias relatively to the overall variance of the dependent variable. We observe that variations of this hypothesis are often assumed in the literature.

In standard models of peer effects, it is also routinely assumed that the in-degree of agents is constant and normalized to one, as follows.

Assumption 3. Row Normalization: $\bar{g}_i \equiv \sum_{j=1}^{N} g_{ij} = 1$ *for all* $i = 1, \ldots, N$

Under Assumption 3 social effects represent the individual response to the (weighted) average behavior or characteristics of peers. This contrasts with models where social effects are a function of the total intensity of connections. Throughout most of this paper we will maintain Assumption 3, while concentrating on the identification of the combined parameter β. Later we relax this hypothesis and, among the possible extensions of our approach, we discuss the possibility to separately identify μ and ν by exploiting variation in individual in-degree. Incidentally, observe that Assumption 3 implies that no agent is allowed to be "isolated" (disconnected from the network) and that under row normalization, Assumption 2 reduces to $|\beta| \in [0, 1)$.

Under all the hypotheses outlined thus far, the following result is easily obtained.

Proposition 1. Equilibrium. *For all realizations of* $(\boldsymbol{x}, \boldsymbol{\varepsilon}, \mathcal{G})$*, under Assumptions 1-3 there exists a unique equilibrium of the game, which gives rise to an equation for the outcome* y_i *that can be expressed for each player* $i = 1, \ldots, N$ *as follows:*

$$y_i = \alpha + \beta \sum_{j=1}^{N} g_{ij} y_j + \gamma x_i + \varepsilon_i \tag{3}$$

where $\alpha \equiv (1 - \mu)^{-1} [\alpha_0 + (\mu + \nu) \log \mu]$ *and* $\gamma \equiv (1 - \mu)^{-1} \gamma_0$.

Proof. The First Order Condition from utility maximization can be written, for each player $j = 1, \ldots, N$, as:

$$e_j = y_j + \log \mu \tag{4}$$

substituting this expression into (2) results in (3). Moreover, by substituting (2) into (4) and solving for e_j it is easily seen that – under the non-explosiveness condition – the N First Order Conditions together represent a contraction of $(e_1, \ldots, e_N)$ in the $\left(\mathbb{R}^N, \mathfrak{M}\right)$ metric space, where $\mathfrak{M}$ is the max norm. This implies uniqueness. □

Let us examine the reduced form expression (3) that is generated in equilibrium. While it resembles the standard equation of linear in means models from the peer effects literature, it provides some additional insights in relationship with the model.

First, parameter β – corresponding to the *endogenous effect* from the original classification by Manski (1993) – is given here a clear behavioral interpretation. In fact, β is equal to the direct effect of connections' effort ν amplified by a factor representing the equilibrium response of individual effort caused by complementarities: intuitively, students put additional effort while firms increase their R&D investment as they are aware of the interdependencies and expect their connections to behave similarly. This interpretation of β is important, since in many empirical studies of social externalities individual "effort" is not observable by researchers.

The second difference with typical linear-in-means models is that in our model we do not include Manski's *exogenous effect*, that is a structural dependence of individual outcomes on the characteristics x_j of peers (also called "contextual" effects). While we could easily include an additional term in (2) to allow for the exogenous effect, we believe that our choice makes it easier to illustrate the following fact.

Proposition 2. Non-identification of contextual effects: *There exist specific restrictions on $\mathcal{F}(\cdot)$ such that the model is observationally equivalent to the following alternative structure:*

$$y_i = \alpha' + \beta' \sum_{j=1}^{N} g_{ij} y_j + \gamma' x_i + \delta' \sum_{j=1}^{N} g_{ij} x_j + \varepsilon'_i \tag{5}$$

where $\delta' \neq 0$ and the random vector $\boldsymbol{\varepsilon}' = (\varepsilon'_1, \dots, \varepsilon'_N)$ is such that $\mathbb{E}[\boldsymbol{\varepsilon}' | \boldsymbol{x}, \mathcal{G}] = 0$.

Proof. Suppose that $\mathcal{F}(\cdot)$ implies that $\varepsilon_i = \rho \sum_{j=1}^{N} g_{ij} \varepsilon_j + \varepsilon'_i$ and $\mathbb{E}[\varepsilon_j | x_j] = \kappa + \chi x_j$ where $\chi \neq 0$ and $\rho \neq 0$. It is easy to see that under those conditions, models (3) and (5) are observationally equivalent under $\alpha' = \alpha + \rho\kappa$, $\beta' = \beta$, $\gamma' = \gamma$ and $\delta' = \rho\chi$. □

While the particular example that we chose to straightforwardly prove our statement is abstract,[5] it serves to make an important point. If individual unobservables ε_i are correlated in the network – say, because agents form connections by sorting on ability – and, in addition, individual characteristics x_i are also correlated with the unobservables, then "contextual effects" δ' are just a statistical byproduct of these more fundamental structural behavioral patterns. We see this as a cautionary message to researchers aiming to estimate spillover effects in any given economic context: the

[5] In that example ability ε_i follows a first order spatially autoregressive process, which implies that individual unobservables are increasingly dissimilar the farther apart are any two agents in the network (intuitively, a spatial AR(1) process can be approximated as a spatial MA(∞) process).

solution of potential endogeneity problems due to simultaneous unobservables and network formation must precede model specification. Clearly, a similar problem may also affect the main behavioral parameter β of endogenous spillover effects. The rest of this article discusses strategies aimed at disentangling genuine externalities from shared confounders. Throughout most of the exposition we maintain the assumption that individual "effort" is not directly observable by researchers.

3 Identification

In this section we discuss under what conditions it is possible to identify the parameters of model (3) even if individual characteristics and the network are endogenous with respect to the unobservables. Following a description the problem, we illustrate our approach first under simple linear assumptions about the underlying data generation process, and then under more general conditions. At the end of the section we comment on some possible extensions of the proposed methodology.

3.1 SAR models

We find it useful to briefly discuss how the endogeneity problem we are concerned with differs from those of previous analyses. To this end, we re-write the statistical model implied by the structural relationship (3) by making use of compact notation (for illustratory purposes, for now we omit subindices regarding the sample size N).

$$\mathbf{y} = \alpha\boldsymbol{\iota} + \beta\mathbf{G}\mathbf{y} + \gamma\mathbf{x} + \boldsymbol{\varepsilon} \tag{6}$$

Here, $\mathbf{y} = (y_1, \dots, y_N)^{\mathrm{T}}$, $\mathbf{x} = (x_1, \dots, x_N)^{\mathrm{T}}$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_N)^{\mathrm{T}}$ are the realizations of y_i, x_i and ε_i – respectively – stacked over all the agents; $\mathbf{G}$ instead is the *adjacency matrix* with g_{ij} entries. Following the classification of spatial econometric models by Elhorst (2014), we call this a *spatially autoregressive* (SAR) model.[6] Note that under row-normalization of $\mathbf{G}$ (Assumption 3) any SAR model corresponds to the linear-in-means model typical of peer effects studies, but deprived of contextual effects.

The most apparent econometric problem of model (6) is one of simultaneity: since the y_i's of different agents are structurally dependent on one another, the spatially

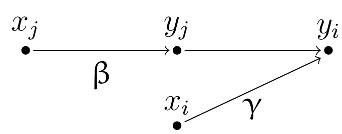
[6] Other authors prefer the denomination "mixed regressive-spatially autoregressive" in order to remark the presence of $\mathbf{x}$ on the right-hand side of (6). Here we opt for a more concise terminology.

autoregressive component $\mathbf{Gy}$ of (6) is correlated with the error term – the so-called *reflection* problem – and thus OLS estimation is inconsistent. There is a vast literature in spatial econometrics, which is not our objective to review here, that concerns the ML estimation of (6) under normality assumptions. Semi-parametric approaches to the estimation of models akin to (6) include IV-2SLS (Kelejian and Prucha, 1998) as well as GMM (Lin and Lee, 2010). The former appears of particular relevance here, as it has been extended to models featuring contextual effects and network fixed effects through the influential contribution by Bramoullé et al. (2009).

To understand why internal identification of (6) is possible, observe that if $\mathbf{G}$ is linearly independent from the identity matrix $\mathbf{I}$, as $(\mathbf{I} - \beta\mathbf{G})$ would also be invertible, the model can be then rewritten in a "reduced form" fashion as:

$$\mathbf{y} = (\mathbf{I} - \beta\mathbf{G})^{-1} (\alpha\iota + \gamma\mathbf{x} + \boldsymbol{\varepsilon}) \simeq \sum_{s=0}^{\infty} \beta^s \mathbf{G}^s (\alpha\iota + \gamma\mathbf{x} + \boldsymbol{\varepsilon}) \tag{7}$$

implying, under $\mathbb{E}[\boldsymbol{\varepsilon}|\,\mathbf{x}] = 0$, the existence of an *infinite* set of instrumental variables of the form $(\mathbf{Gx}, \mathbf{G}^2\mathbf{x}, \dots)$. The intuition behind identification is that it is possible to predict the outcomes $\mathbf{Gy}$ of connected agents through their characteristics $\mathbf{Gx}$. This idea is exemplified Graph 1, which involves variables $(x_i, y_i; x_j, y_j)$ pertaining to any two connected observations (i, j). In the graph, arrows represent the structural relationships between variables that allow to identify the indicated parameter of interest.



Graph 1: Identification of SAR models

Models featuring contextual effects like a term $\delta\mathbf{Gx}$ on the right-hand side of (6) entail the additional complication that clearly $\mathbf{Gx}$ is not excluded from the structural form. However, $\mathbf{G}^2\mathbf{x}$ would then be a relevant instrument for the identification of β: if contextual effects exist, the characteristics of "friends of friends" affect the outcomes of direct peers – easily extending the intuition above – so that β and δ are separately identified. These ideas are best framed as a system of simultaneous equations, which are generally known to be identified so long as enough instrument exist to satisfy both the order and rank conditions. Here it is the structure of networks that naturally gives rise to the appropriate exclusion restrictions, in the form of the characteristic of others agents that have no direct effect on individual outcomes (Rose, 2017b).

These ideas and the related results are all based on the assumptions of exogenous covariates $\mathbf{x}$. In the systematic analysis of the literature by Blume et al. (2015), an equivalent of assumption $\mathbb{E}[\boldsymbol{\varepsilon}|\,\mathbf{x}] = 0$ is central to all results about identification. In SAR and linear-in-means models, the endogeneity of individual characteristics not only prevents the identification of their specific effect on the outcome of interest, but also of social effects themselves, as the x_i's of peers can no longer serve as instruments for the spatially autoregressive term. This suggests that the extent of the problem depends on the breadth of endogeneity in network space – that is, to what extent individual unobservables are correlated with the characteristics of peers, of peers of peers and so forth. In what follows we illustrate under what conditions the internal identification of SAR models is possible even if the x_i's are endogenous. For the sake of exposition, we start from a simplified, semi-formal characterization the problem. In the more general treatment we will introduce our assumptions more rigorously.

3.2 Spatial Linear Endogeneity

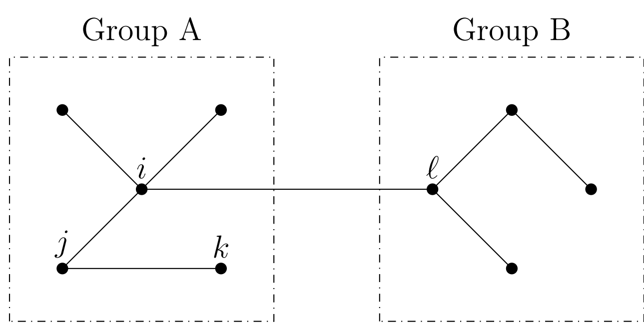
Suppose that the agents' characteristics $\boldsymbol{x}$ are statistically related to the unobservables $\boldsymbol{\varepsilon}$ according to the following simple linear relationship:

$$x_i = \widetilde{x}_i + \xi \sum_{j=1}^{N} c_{ij} \varepsilon_j \tag{8}$$

for $i = 1, \dots, N$. In the above, $\xi \in \mathbb{R}$ is a parameter; $\widetilde{x}_i$ is what we call the *independent component* of the variation of x_i, whose distribution is left unrestricted except for the fact that it is assumed to be continuous ($\widetilde{x}_i \neq \widetilde{x}_j$ almost surely for $i \neq j$) and independent of individual unobservables ($\mathbb{E}[\widetilde{x}_i \varepsilon_j] = 0$ for any i, j); the weights c_{ij} instead, that we call *characteristic weights*, introduce the statistical spatial dependence of interest. Like in the case of the adjacency weights g_{ij}, we impose the normalization $c_{ij} \in [0, 1]$; unlike those, however, we allow for $c_{ii} \neq 0$. We collect all the characteristic weights in the N^2-dimensional set $\mathcal{C}$, that we call the "characteristics structure."

The relationship expressed in (8) can flexibly represent various patterns of interdependence between the socio-economic variables of different agents. In a schooling context, for example, the quality of teachers and the overall resources made available to every child (x_i) may endogenously depend on their preferences and/or abilities (ε_i) of their classmates. This may be induced via an explicit allocation mechanism, if say more motivated students are assigned the best resources or, conversely, more disadvantaged ones are compensated with some extra support. Alternatively, a statistical relationship such as (8) might represent some other types of endogenous self-selection or sorting mechanisms, say driven by the choice of parents, which assigns pupils across classrooms. This type of dependence between the characteristics of different students is made explicit by equation (8): x_i is decomposed between an independent part $\widetilde{x}_i$, which does not reflect any group factors, and an "endogenous" component, which is shaped by a set of characteristic weights $\mathcal{C}$ with a fully overlapping group structure.[7]

We allow the characteristics structure $\mathcal{C}$ to be statistically dependent on the network structure $\mathcal{G}$ that gives rise to social effects, even if their realizations are different. This would introduce an additional source of network endogeneity in our model: if the distribution of $\mathcal{C}$ depends on that of $\mathcal{G}$, and $\boldsymbol{x}$ is generated according to (8), it follows that $\mathcal{G}$ and $(\boldsymbol{x}, \boldsymbol{\varepsilon})$ are statistically dependent in $\mathcal{F}(\cdot)$, even if the connections in $\mathcal{G}$ do not depend directly on the individual characteristics. To substantiate, suppose again that $\mathcal{C}$ describes a group structure like that of different classrooms, while $\mathcal{G}$ is the network actual friendship relationships between pair of students. Typically, friendship links are not transitive; yet it is natural to think that, for any two students i and j, the probability distribution of g_{ij} varies whether they are classmates or not ($c_{ij} \in \{0, 1\}$), since being classmates makes it easier to develop social bonds. This would give rise to scenarios like the one represented in in Graph 2, where friendship links are dense within groups (classrooms) but sparse between groups.



Graph 2: A Cross-Group Friendship Network

[7]In a fully overlapping group structure, if any two agents i and j are connected, they are also either both connected or both disconnected to any third agent k (if $c_{ij} \neq 0$ then $c_{ik} \neq 0 \Leftrightarrow c_{jk} \neq 0$ for all $(i, j, k) \in \mathcal{I}^3$).

The setting we have described is a simple form of endogeneity that would threaten the identification of social effects in the spirit of Angrist's (2014) mentioned critique. In fact, it invalidates standard moments in the spirit of Lee (2007a) that are based on the spatial lags of $\mathbf{x}$. To verify this, collect the terms $\widetilde{x}_i$ and c_{ij} from (8) in the vector $\widetilde{\mathbf{x}} = (\widetilde{x}_1, \dots, \widetilde{x}_N)^{\mathrm{T}}$ and in the *characteristics matrix* $\mathbf{C}$ (of size $N \times N$) respectively, and assume homoscedasticity:

$$\mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\right] = \sigma_0^2 \mathbf{I} \tag{9}$$

thus, for any nonnegative integer q:

$$\begin{aligned}
\mathbb{E}\left[\left(\mathbf{G}^q \mathbf{x}\right)^{\mathrm{T}} \boldsymbol{\varepsilon}\right] &= \mathbb{E}\left[\left(\widetilde{\mathbf{x}} + \xi \mathbf{C}\boldsymbol{\varepsilon}\right)^{\mathrm{T}} \left(\mathbf{G}^q\right)^{\mathrm{T}} \boldsymbol{\varepsilon}\right] \\
&= \xi \cdot \mathbb{E}\left[\boldsymbol{\varepsilon}^{\mathrm{T}} \left(\mathbf{G}^q \mathbf{C}\right)^{\mathrm{T}} \boldsymbol{\varepsilon}\right] \\
&= \xi \cdot \mathrm{Tr}\left(\mathbf{C}\mathbf{G}^q \cdot \mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^{\mathrm{T}}\right]\right) \\
&= \xi \sigma_0^2 \cdot \mathrm{vec}\left(\mathbf{C}\right)^{\mathrm{T}} \mathrm{vec}\left(\mathbf{G}^q\right)
\end{aligned} \tag{10}$$

where the second line exploits the fact that $\mathbb{E}\left[\widetilde{\mathbf{x}}^{\mathrm{T}} \mathbf{A} \boldsymbol{\varepsilon}\right] = 0$ for any conformable matrix $\mathbf{A}$, while the subsequent lines exploit the properties of the trace in relationship with the vectorization operator. Observe that, while the moments in question are nonzero (as expected), we have obtained an explicit expression for their bias. This suggests a natural set of moment conditions to be employed for identification and estimation of the set of combined parameters $\boldsymbol{\vartheta} \equiv (\alpha, \beta, \gamma, \xi^*)$, where $\xi^* \equiv \xi\sigma_0^2$:

$$\mathbb{E}\left[\begin{pmatrix} \boldsymbol{\iota}^{\mathrm{T}} \\ \mathbf{x}^{\mathrm{T}} \\ \mathbf{x}^{\mathrm{T}}\mathbf{G} \\ \mathbf{x}^{\mathrm{T}}\mathbf{G}^2 \end{pmatrix} \left(\mathbf{y} - \alpha\boldsymbol{\iota} - \beta\mathbf{G}\mathbf{y} - \gamma\mathbf{x}\right)\right] - \xi^* \cdot \begin{bmatrix} 0 \\ \mathrm{Tr}\left(\mathbf{C}\right) \\ \mathrm{Tr}\left(\mathbf{C}\mathbf{G}\right) \\ \mathrm{Tr}\left(\mathbf{C}\mathbf{G}^2\right) \end{bmatrix} = \mathbf{0} \tag{11}$$

in fact, the identification of $\boldsymbol{\vartheta}$ is possible under very general conditions.

Proposition 3. Identification under simple linear endogeneity. *Consider the statistical model characterized by equations (6), (8) and (9); and suppose that matrices* $\mathbf{C}$ *and* $\mathbf{G}$ *are observed. If the three matrices* $\mathbf{I}$, $\mathbf{G}$ *and* $\mathbf{G}^2$ *are linearly independent and matrix* $\mathbf{C}$ *overlaps at least partially with any of the three matrices above – in the sense that the traces* $\mathrm{Tr}\left(\mathbf{C}\right)$, $\mathrm{Tr}\left(\mathbf{C}\mathbf{G}\right)$ *and* $\mathrm{Tr}\left(\mathbf{C}\mathbf{G}^2\right)$ *are not simultaneously all zeros – then the combined parameters* $\boldsymbol{\vartheta} \equiv (\alpha, \beta, \gamma, \xi^*)$ *are almost surely identified.*

Proof. The Jacobian matrix $\mathbf{H}(\boldsymbol{\vartheta})$ of the moment conditions (11) can be split as follows. The first row is trivially given by:

$$\mathbf{H}_{1,*}(\boldsymbol{\vartheta}) = -\mathbb{E}\begin{bmatrix} N & \boldsymbol{\iota}^{\mathrm{T}}\mathbf{G}\mathbf{y} & \boldsymbol{\iota}^{\mathrm{T}}\mathbf{x} & 0 \end{bmatrix}$$

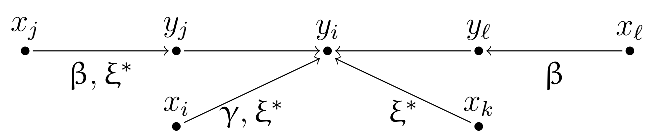
while the block constituted by the second, third and fourth rows can be written as:

$$\mathbf{H}_{2-4,*}(\boldsymbol{\vartheta}) = -\begin{bmatrix} \operatorname{vec}(\mathbf{I})^{\mathrm{T}} \\ \operatorname{vec}(\mathbf{G})^{\mathrm{T}} \\ \operatorname{vec}(\mathbf{G}^2)^{\mathrm{T}} \end{bmatrix} \cdot \mathbb{E}\begin{bmatrix} \operatorname{vec}\left(\boldsymbol{\iota}\mathbf{x}^{\mathrm{T}}\right) & \operatorname{vec}\left(\mathbf{G}\mathbf{y}\mathbf{x}^{\mathrm{T}}\right) & \operatorname{vec}\left(\mathbf{x}\mathbf{x}^{\mathrm{T}}\right) & \operatorname{vec}(\mathbf{C}) \end{bmatrix}$$

by exploiting again the properties of the trace and vectorization operators. Observe that, by the specification of linear endogeneity (8), $\operatorname{vec}\left(\boldsymbol{\iota}\mathbf{x}^{\mathrm{T}}\right)$, $\operatorname{vec}\left(\mathbf{x}\mathbf{x}^{\mathrm{T}}\right)$ and $\operatorname{vec}(\mathbf{C})$ are almost surely linearly independent, even if $\mathbf{C}$ has constant columns. Furthermore, by the definition of a SAR model (6), these vectors would be almost surely linearly independent of $\operatorname{vec}\left(\mathbf{G}\mathbf{y}\mathbf{x}^{\mathrm{T}}\right)$ as well. In this case, matrix $\mathbf{H}(\boldsymbol{\vartheta})$ is singular under only two circumstances: either matrices $\mathbf{I}$, $\mathbf{G}$ and $\mathbf{G}^2$ – and thus their vectorized versions – are linearly dependent; or the traces of matrices $\mathbf{C}$, $\mathbf{CG}$ and $\mathbf{CG}^2$ are all zero. □

This is a very powerful result: it states that if the researcher has some knowledge about the spatial extent of the process which relates the observable characteristics of agents to the unobservables of some others, then the parameters of the SAR model – including the "endogenous" social effect β – can be identified under the same conditions given by Bramoullé et al. (2009): that the network $\mathcal{G}$ is not shaped according to a "fully overlapping" group structure. In addition, it is necessary that the characteristics matrix $\mathbf{C}$ – which defines the spatial extent of endogeneity – overlaps at least partially with the network, but otherwise it is left unrestricted; it is allowed to assume a group structure or to coincide with the adjacency matrix $\mathbf{G}$. This second condition, however, is largely moot, since its violation would prevent the identification of the combined parameter ξ^*, but not of the main parameters of interest (α, β, γ). In fact, if the spatial correlation of individual characteristics is unrelated to the network there is no endogeneity problem, and standard "peers-of-peers" instruments are valid!

We illustrate the intuition behind our identification result in two ways: one graphical and one algebraic-statistical. The graphical one is supported by Graph 3 below, which represents the structural relationships between the x and y variables of four agents (i, j, k, ℓ) that are involved in both a network and group structure like the one depicted in Graph 2. Consider first agents i, j and k, who all belong to Group A. While i and j are direct friends, identification of β and γ cannot proceed like in the case of an "exogenous" SAR (Graph 1) since both own characteristics x_i and peers' characteristics x_j are contaminated by the unobservables. However, the characteristics x_k of agent k, who is a friend of j but not of i, *conditioning on the endogenous effect* only reflect the endogeneity due to the group structure, making it possible to identify ξ^*. Hence, β and γ can be identified residually like in the exogenous case. Consider next agent ℓ, who is a direct friend of i but belongs to a different group: his or her characteristics x_ℓ do not correlate with i's unobservables under the running assumptions, and thus can be exploited in order to straightforwardly identify the social effect β. While useful, this is however redundant for identification, since it is enough that agent i has *at least one* indirect friend, even if from the same group like k.



Graph 3: Identification of SAR models under linear endogeneity

Moving to the second piece of intuition, consider again the reduced form of the SAR model given in (7); under the endogeneity specification (8) it reads:

$$\mathbf{y} = (\mathbf{I} - \beta\mathbf{G})^{-1} \left[\alpha\iota + \gamma\left(\widetilde{\mathbf{x}} + \xi\mathbf{C}\boldsymbol{\varepsilon}\right) + \boldsymbol{\varepsilon}\right] \simeq \sum_{s=0}^{\infty} \beta^s \mathbf{G}^s \left[\alpha\iota + \gamma\widetilde{\mathbf{x}} + (\mathbf{I} + \gamma\xi\mathbf{C})\,\boldsymbol{\varepsilon}\right]$$

clearly, the model would be identified in analogy with the exogenous case if one could observe the independent component $\widetilde{x}_i$ of the observable characteristics x_i. Obviously this is unfeasible by definition; however, if a researcher knows $\mathbf{C}$, these independent components can be indirectly backed up through the following, intuitively appealing nonlinear moments:

$$\mathbb{E}\left[(\mathbf{x} - \xi\mathbf{C}\boldsymbol{\varepsilon})^{\mathrm{T}} \mathbf{G}^q \boldsymbol{\varepsilon}\right] = 0 \tag{12}$$

for $q = 0, 1, 2$ or higher. Note that like in Proposition 3, ξ^* is identified if $\mathbf{C}$ overlaps at least partially with matrices $\mathbf{I}$, $\mathbf{G}$, and $\mathbf{G}^2$, and such overlap presents some variation (which does under the standard linear independence condition). Interestingly, we have simulated an estimation of our model using moment conditions (12); however, this exercise is outperformed by the main simulation which is based on the bias-adjusted moments (11), and that we discuss later in Section 5. Both sets of moments follow from the same data generation process and thus should be equivalent, but the linear ones are clearly computationally more convenient.

3.3 Spatial Linear Endogeneity: General Result

The identification result illustrated by Proposition 3 is restricted to a simple model under a very restrictive (homoscedastic) structure of the error terms. Next, we discuss how that result can be extended to the following, more general model:

$$\mathbf{y} = \alpha\boldsymbol{\iota} + \beta\mathbf{G}\mathbf{y} + \mathbf{X}\boldsymbol{\gamma} + \mathbf{G}\mathbf{X}\boldsymbol{\delta} + \boldsymbol{\varepsilon} \tag{13}$$

where $\mathbf{X}$ is a $N \times K$ data matrix of K observable characteristics (with $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ having full rank), $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_K)$ is the vector of K direct effects associated with each of these characteristics, and $\boldsymbol{\delta} = (\delta_1, \dots, \delta_K)$ are the K related "contextual effects." In Elhorst's taxonomy, this is a standard multivariate "Spatial Durbin Model" (SDM), otherwise known – under row-normalization of $\mathbf{G}$ – as a linear-in-means model. Note that such a model could easily follow from an extension of our theoretical framework, where nature initially draws $(\mathbf{X}, \boldsymbol{\varepsilon}, \mathcal{G})$ from some more general distribution $\mathcal{F}(\cdot)$. To keep the problem interesting we presume that all the observable characteristics $\mathbf{X}$ are potentially endogenous and structurally dependent on $\boldsymbol{\varepsilon}$, or else a single exogenous factor could be enough to identify β under the logic of Graph 1. In addition, we allow for a more general structure of the error term as per the following assumptions.

Assumption 4. Primitive shocks: *there exists a set of N "primitive" i.i.d. shocks $\boldsymbol{\upsilon} \equiv (\upsilon_1, \dots, \upsilon_N)^{\mathrm{T}}$ such that:* (i) $\mathbb{E}[\boldsymbol{\upsilon}] = \mathbf{0}$*;* (ii) $\mathbb{E}\left[\boldsymbol{\upsilon}\boldsymbol{\upsilon}^{\mathrm{T}}\right] = \sigma^2\mathbf{I}$*;* (iii) *for some $d > 0$, $\mathbb{E}\left[|\upsilon_i|^{4+d}\right] < \infty$ for $i = 1, \dots, N$.*

Assumption 5. SARMA Unobservables: *the unobservable characteristics follow a stationary Spatial Autoregressive Moving Average process of order (A, M):*

$$\boldsymbol{\varepsilon} = (\mathbf{I} - \phi_1\mathbf{F}_1 - \phi_2\mathbf{F}_2 - \dots - \phi_A\mathbf{F}_A)^{-1}(\mathbf{I} + \psi_1\mathbf{E} - \psi_2\mathbf{E}_2 - \dots - \psi_M\mathbf{E}_M)\boldsymbol{\upsilon}$$

where $(\mathbf{E}_1, \mathbf{E}_2, \dots, \mathbf{E}_A)$ and $(\mathbf{F}_1, \mathbf{F}_2, \dots, \mathbf{F}_M)$ are two possibly identical sets of linearly independent $N \times N$ matrices; $\mathbf{\Phi}_a \equiv \mathbf{I} - \sum_{a=1}^{a'} \phi_a\mathbf{F}_a$ is invertible for all $a' \leq A$, and the associated parameters are restricted to the unit circle: $\|\boldsymbol{\phi}\|_2 < 1$ and $\|\boldsymbol{\psi}\|_2 < 1$.

Together, these two assumptions characterize the stochastic properties of the error term, which is allowed to have a very general spatial correlation structure expressed in terms of a sequence of "primitive" well-behaved shocks. The spatially autoregressive component of the error term is defined by the sequence of matrices $(\mathbf{F}_1, \ldots, \mathbf{F}_A)$ as well as the set of parameters $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_A)$; the moving average part is encapsulated by matrices $(\mathbf{E}_1, \ldots, \mathbf{E}_M)$ and parameters $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_M)$. Note that both sequences of matrices can be identical and depend on the network structure; a leading case that we allow is $\mathbf{F}_a = \mathbf{G}^a$ for $a \leq A$ and $\mathbf{F}_m = \mathbf{G}^m$ for $m \leq M$. While this specification is quite flexible, we are especially interested in the Spatial Moving Average (SMA) component of the process. If the spatially autocorrelated component is absent (e.g. $\boldsymbol{\phi} = \mathbf{0}$), in fact, a spatial moving average process implies finite spatial autocorrelation in the space under analysis. We find this empirical property to be a good approximation of some real-world stylized facts about variables that are diffused in networks.[8] While our identification results extend to any SARMA process, our estimation framework and Monte Carlo simulation specialize in a simple SMA(1) process, or $(A, M) = (0, 1)$.

The next assumption generalizes expression (8), which characterizes the spatial extent of endogeneity, to the multivariate case. In particular, we associate a different characteristics matrix $\mathbf{C}_k$ to each of the K individual observable characteristic.

Assumption 6. Multivariate Spatial Linear Endogeneity: *each column of* $\mathbf{X}$ *is given, for* $k = 1, \ldots, K$, *by:*

$$\mathbf{X}_{*,k} = \widetilde{\mathbf{x}}_k + \xi_k \mathbf{C}_k \boldsymbol{\upsilon} \tag{14}$$

where $\xi_k \in \mathbb{R}$, $\mathbf{C}_k$ *is an unrestricted characteristics matrix specific of the* k*-th covariate, while* $\widetilde{\mathbf{x}}_k$ *is a random vector with unrestricted but finite mean. In addition, we assume that* $\widetilde{\mathbf{x}}_k$ *has a continuous support, in the sense that for any two observations* $i \neq j$, $\widetilde{x}_{ki} = \widetilde{x}_{kj}$ *has probability zero; and moreover that for any two* $k, k' = 1, \ldots, K$, *the probability limit* $\Xi_{kk'} \equiv \operatorname{plim} N^{-1} \sum_{i=1}^{N} \left(\widetilde{x}_{ki} - \mathbb{E}\left[\widetilde{x}_{ki}\right]\right)\left(\widetilde{x}_{k'i} - \mathbb{E}\left[\widetilde{x}_{k'i}\right]\right)$ *is finite.*

[8] In a study about the health outcomes of children, Christakis and Fowler (2013) find that most variables present a spatial autocorrelation in the space of friendship network up to two degrees of distance. Zacchia (2018) observes the same property for the R&D investment of high-tech firms that are connected through research collaborations. In addition, he argues that this property can follow from an underlying SMA(1) process of technological shock, and that it is a good approximation of a model of network formation driven by a homophily dynamic, where two firms link up with some probability only if their unobservables are similar.

Notice a difference with the simpler one-characteristic case given in (8): in the latter, matrix $\mathbf{C}$ multiplies the error terms ε_i's of the model; in (14), each of the K characteristic matrices $\mathbf{C}_k$ multiplies the "primitive" shocks υ_i's. We believe that this specification better captures a scenario where there are some external, unobservable factors (the primitive "common" shocks) that affect both individual outcomes and their characteristics, in a way that depends on the structure of social interactions as defined in our assumptions. However, our results about identification and estimation would be easy to extend to a setup where the specification of endogeneity in (14) were to involve $\boldsymbol{\varepsilon}$ (which would still follow a generalized SARMA process) instead of $\boldsymbol{\upsilon}$.

We are now ready to state our main result. In what follows, we collect the parameters that characterize the spatial endogeneity with the vector $\boldsymbol{\xi} = (\xi_1, \dots, \xi_K)$.

Theorem 1. General Identification Result. *Under Assumptions 1-6, the parameters* $\boldsymbol{\theta} \equiv (\alpha, \beta, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\xi}, \boldsymbol{\phi}, \boldsymbol{\psi}, \sigma^2)$ *are almost surely identified if the matrices* $\mathbf{I}$, $\mathbf{G}$, $\mathbf{G}^2$ *and* $\mathbf{G}^3$ *are linearly independent; for every* $k = 1, \dots, K$ *it is* $\beta\gamma_k + \delta_k \neq 0$*; and the following three conditions hold simultaneously:*

(a) *the researcher can observe some* $P \geq 1 + A + M$ *matrices* $\{\mathbf{P}_p\}_{p=1}^{P}$ *of size* $N \times N$ *that are all linearly independent of one another;*

(b) *for all appropriate* $(\boldsymbol{\phi}, \boldsymbol{\psi})$ *all matrices in the following set:*

$$\left\{\mathbf{F}_a \left(\mathbf{I} - \phi_1 \mathbf{F}_1 - \cdots - \phi_A \mathbf{F}_A\right)^{-1} \left(\mathbf{I} + \psi_1 \mathbf{E} - \cdots - \psi_M \mathbf{E}_M\right)\right\}_{a=0}^{A}$$

(where $\mathbf{F}_0 = \mathbf{I}$*) are linearly independent of all matrices in the set* $\{\mathbf{E}_m\}_{m=1}^{M}$*;*

(c) *for every* k*, the four traces defined by the following expression for* $q = 0, 1, 2, 3$*:*

$$\operatorname{Tr}\left[\mathbf{C}_k \mathbf{G}^{q-1} \left(\mathbf{I} - \phi_1 \mathbf{F}_1 - \cdots - \phi_A \mathbf{F}_A\right)^{-1} \left(\mathbf{I} + \psi_1 \mathbf{E} - \cdots - \psi_M \mathbf{E}_M\right)\right]$$

are not simultaneously all zeros.

Proof. See the Appendix. The proof is a generalization of Proposition 3. □

Theorem 1 provides a general identification result for linear-in-means models that feature contextual effects, when the observable characteristics of individuals, the error terms and the interaction structure itself are structurally dependent. In addition, we allow for an error term which is allowed to follow a very general stochastic process, and we show that under specific conditions the associated parameters are identified.

The latter is, to the best of our knowledge, a novel result in the spatial econometrics literature, which so far has prevalently examined models whose errors follow simple spatially autoregressive processes.[9] It is instructive to discuss how the various parameters are identified in relation to the requirements of the theorem. We first focus on the "linear" parameters: in particular, the social (endogenous) effect, the contextual (exogenous) effects, and the endogeneity parameters $\boldsymbol{\xi}$; next, we separately elaborate on the various components of the SARMA structure of the error term's variance.

First, observe that $\beta\gamma_k + \delta_k \neq 0$ is a standard condition for the identification of linear-in-means models, as it imposes that social and contextual effects do not cancel out. Next, consider that in our previous simplified analysis, if matrix $\mathbf{G}^3$ is linearly independent from $\mathbf{I}$, $\mathbf{G}$ and $\mathbf{G}^2$, another moment condition like (10) with $s = 3$ can be exploited for identification. In the general case we exploit QK sets of moments of of the kind, for $q = 1, \ldots, Q$, $Q \geq 4$ and $k = 1, \ldots, K$:

$$\mathbb{E}\left[\mathbf{x}_k \mathbf{G}^{q-1} \boldsymbol{\varepsilon}\right] - \lambda_{qk} = 0 \tag{15}$$

where λ_{qk} varies according to the specification of the SARMA process. In an extension of the intuition illustrated through Graph 3, once the endogeneity effect expressed through $\boldsymbol{\xi}$ is netted out, the characteristics of friends of friends $\mathbf{G}^2\mathbf{X}$ would identify the contextual effect $\boldsymbol{\delta}$; while those of third degree indirect friends ("friends of friends of friends") – $\mathbf{G}^3\mathbf{X}$ – identify the social effect β. Note how the theorem's condition (c) is necessary for the identification of $\boldsymbol{\xi}$, as it corresponds with the "not all-zero traces" requirement from Proposition 3. Like that one, this condition is not very interesting, as it only ensures that the endogeneity problem is cogent, in the sense that the spatial correlation of the x_{ik}'s overlaps at least partially with that of the outcomes y_i's.

We consider the variance components next. We identify them through standard covariance restrictions of the kind, for $p = 1, \ldots, P$:

$$\mathbb{E}\left[\boldsymbol{\varepsilon}^{\mathrm{T}} \mathbf{P}_p \boldsymbol{\varepsilon}\right] - \lambda_p = 0 \tag{16}$$

where again λ_p may vary across cases. It appears that requirement (a) of the theorem is necessary in order to rule out collinearity between the P moments; this applies to both elements on the left-hand side of (16). A natural choice for the moment matrices

[9] For example, Kapoor et al. (2007); Kelejian and Prucha (2010); Drukker et al. (2013), among the others, analyze SAR(1) disturbances, while Lee and Liu (2010) consider higher order SAR processes.

is $\mathbf{P}_p = \mathbf{G}^{p-1}$, especially where P is small. Condition (b) instead might seem daunting at first, as it is conceived for quite general situations. It is helpful to evaluate it in a simpler scenario, which might work well in most empirical applications, where the error term follows a first-order SARMA process with $(A, M) = (1, 1)$, and where the network structure accurately describes both components, that is $\mathbf{E}_1 = \mathbf{F}_1 = \mathbf{G}$. In this case, condition (b) simply asks that the adjacency matrix $\mathbf{G}$ and matrix

$$(\mathbf{I} + \phi_1 \mathbf{G})^{-1} (\mathbf{I} + \psi_1 \mathbf{G}) \simeq \sum_{s=0}^{\infty} \phi_1^s \mathbf{G}^{s+1} (\mathbf{I} + \psi_1 \mathbf{G})$$

are linearly independent, which is typically automatically verified for those non-fully transitive, sparse networks that we have in mind as possible empirical applications.

3.4 Extensions

Finally, we analyze two simple extensions of our framework. First, we show how it can accommodate multiple networks and the relative fixed effects. Next, we discuss how the primitive parameters μ and ν from our analytical framework, which are combined in β, can be identified under certain conditions.

Network-level fixed effects

The use of the third power of $\mathbf{G}$ bears some analogies with the scenario analyzed by Bramoullé et al. (2009), where the adjacency matrix represents a set of disconnected networks, to each of which is associated a separate fixed effect, and where the use of indirect connections of third degree is necessary once such fixed effects are partialled out. The difference is that here, it is the endogeneity expressed in (14) which is being removed first. The following corollary is consequent to this one last observation.

Corollary 1. *If the model of interest is:*

$$\mathbf{y} = \mathbf{D}\boldsymbol{\alpha}^* + \beta \mathbf{G}\mathbf{y} + \mathbf{X}\boldsymbol{\gamma} + \mathbf{G}\mathbf{X}\boldsymbol{\delta} + \boldsymbol{\varepsilon} \tag{17}$$

where $\mathbf{D}$ represents a set of D dummy variables, each for a separate component of the network $\mathcal{G}$, and $\boldsymbol{\alpha}^* = (\alpha_1, \ldots, \alpha_D)$ is a vector of associated fixed effects, the parameters $\boldsymbol{\theta} \equiv (\boldsymbol{\alpha}^*, \beta, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\xi}, \boldsymbol{\phi}, \boldsymbol{\psi}, \sigma^2)$ are identified if, in addition to the conditions expressed in Theorem 1, also matrix $\mathbf{G}^4$ is linearly independent of matrices $\mathbf{I}$, $\mathbf{G}$, $\mathbf{G}^2$ and $\mathbf{G}^3$.

Proof. This follows straightforwardly from "network differencing" equation (17) by pre-multiplying the data $(\mathbf{X}, \mathbf{y})$ by $\mathbf{I} - \mathbf{G}$ as in Bramoullé et al. (2009). The identification of the differenced model would follow as per our previous analysis with $\alpha = 0$; the resulting moments are a function of $\mathbf{G}^4$ which thus must be linearly independent of its lower powers. The fixed effects $\boldsymbol{\alpha}^*$ are residually identified as a subnetwork-specific set of intercepts. □

Identification of μ and ν

In our framework, parameter β represents a composite equilibrium effect: it encloses the direct effect of peers' effort, ν, amplified by the equilibrium response of individual effort, $(1-\mu)^{-1}$. Because of Assumption 3 (row-normalization of **G**) the two parameters μ and ν disappear from the reduced form equilibrium equation. However, note that when this hypothesis is dropped, under our framework (6) would become:

$$\mathbf{y} = (\alpha - \zeta)\,\boldsymbol{\iota} + \beta\mathbf{G}\mathbf{y} + \gamma\mathbf{x} + \zeta\bar{\mathbf{g}} + \boldsymbol{\varepsilon} \tag{18}$$

where $\zeta \equiv (1-\mu)^{-1}\nu\log\mu$ and $\bar{\mathbf{g}} \equiv \mathbf{G}\boldsymbol{\iota}$ is the vector of individual in-degrees (the overall strength of all one individual's connections, such that $\bar{g}_i = \sum_{j=1}^{N} g_{ij}$). Since $\exp(\zeta/\beta) = \mu$, if the observable characteristics x_i's and the network $\mathcal{G}$ are exogenous the primitive parameters μ and ν are separately identified in (18). The intuition is straightforward: the variation in individual in-degree $\bar{\mathbf{g}}$ conveys additional information about the overall strength of direct spillovers (expressed by the parameter ν).[10] An individual with more friends or a firm with more connections is likely to enjoy more beneficial externalities. While row-normalization is routinely assumed in studies of peer effects, we find the latter to be a realistic hypothesis.[11]

In our framework, μ and ν are separately identified also under a mildly restrictive instance of endogeneity.

Corollary 2. *Under the conditions expressed by Theorem 1 but Assumption 3, if* $\bar{\mathbf{g}}$ *is linearly independent of the unit vector* $\boldsymbol{\iota}$ *or any other covariate* $\mathbf{X}_{\cdot,k}$ *and, in addition,* $\mathbb{E}\left[\bar{\mathbf{g}}^{\mathrm{T}}\boldsymbol{\varepsilon}\right] = 0$ *holds, then parameters* μ *and* ν *are separately identified.*

[10] Note that the exact relationship between β, μ and ν depends on functional form assumptions of our model, but the intuition is more general.

[11] If individual "effort" e_i is observable, an alternative route for the separate identification of μ and ν would be based on the structural "production function" (2): this is the approach taken in studies of R&D spillovers, since researchers can typically observe the R&D expenditures of firms.

Proof. Re-define the moments from the proof of Theorem 1 in terms of the residual $\boldsymbol{\varepsilon}(\boldsymbol{\theta}, \zeta) = \mathbf{y} - (\alpha - \zeta)\boldsymbol{\iota} - \beta\mathbf{G}\mathbf{y} - \mathbf{X}\boldsymbol{\gamma} - \mathbf{G}\mathbf{X}\boldsymbol{\delta} - \zeta\bar{\mathbf{g}}$, and add $\mathbb{E}\left[\bar{\mathbf{g}}^{\mathrm{T}}\boldsymbol{\varepsilon}(\boldsymbol{\theta}, \zeta)\right] = 0$ to the set. Clearly, this does not affect the full rank properties of the moments' Jacobian. □

Essentially, if the observable characteristics are endogenous as per Assumption 6, *but the intensity of individual connections is independent of individual unobservables*, μ and $\boldsymbol{\nu}$ can be separately identified by adding the additional regressor $\bar{g}_i$. Note that some form of statistical dependence of the adjacency matrix $\mathbf{G}$ on the characteristics matrices $\mathbf{C}_k$ is still allowed. Scenarios where the identifying assumption are violated are obvious: a very skilled pupil or a very successful firm may find themselves with more or more intense connections. In future work, it would be interesting to examine under what conditions μ and $\boldsymbol{\nu}$ can be separately identified even if the hypothesis in question fails.

4 Estimation

The moment conditions that support our main identification results lend themselves naturally to GMM estimation. In this section we describe how the estimation framework introduced by Lee (2007a) can be adapted to our assumed forms of endogeneity. In doing so, we specialize – as mentioned earlier – to a simple stochastic process that governs our the error term: a spatial moving average of first degree. This facilitates the asymptotic analysis, although the results can be extended to any SARMA process.

Assumption 7. SMA(1) Unobservables: $\boldsymbol{\phi} = \mathbf{0}$ *and* $\psi_m = 0$ *for* $m \geq 2$.

In what follows, we write $\psi = \psi_1$ and $\boldsymbol{\theta} = (\alpha, \beta, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\xi}, \psi, \sigma^2)$. We also denote the true parameter values as $\boldsymbol{\theta}_0$, we introduce N subscripts, and we define the following matrices for $q = 1, \dots, Q$.

$$\mathbf{Q}_{q,N} \equiv \mathbf{X}_N^{\mathrm{T}}\mathbf{G}_N^{q-1}$$

Our GMM estimator is based on a set of $1 + QK + P$ moments conditions, with $Q \geq 4$ and $P \geq 2$:

$$\mathbb{E}\left[\mathbf{m}_N(\boldsymbol{\theta}_0)\right] - \boldsymbol{\lambda}_N(\boldsymbol{\theta}_0) = \mathbf{0} \tag{19}$$

where, given $\boldsymbol{\varepsilon}_N(\boldsymbol{\theta}) = \mathbf{y}_N - \alpha\boldsymbol{\iota}_N - \beta\mathbf{G}_N\mathbf{y}_N - \mathbf{X}_N\boldsymbol{\gamma} - \mathbf{G}_N\mathbf{X}_N\boldsymbol{\delta}$, it is:

$$\mathbf{m}_N(\boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{\iota}^{\mathrm{T}}\boldsymbol{\varepsilon}_N(\boldsymbol{\theta}) & \cdots & \boldsymbol{\varepsilon}_N^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{Q}_{q,N}^{\mathrm{T}} & \cdots & \boldsymbol{\varepsilon}_N^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{P}_{p,N}\boldsymbol{\varepsilon}_N(\boldsymbol{\theta}) & \cdots \end{bmatrix}^{\mathrm{T}}$$

for $q = 1, \ldots, Q$ and $p = 1, \ldots, P$. As for vector $\boldsymbol{\lambda}_N(\boldsymbol{\theta})$, its first element is given by $\lambda_{1,N}(\boldsymbol{\theta}) = 0$, while the others are:

$$\lambda_{1+qk,N}(\boldsymbol{\theta}) \equiv \sigma^2 \xi_k \mathrm{Tr}\left[\mathbf{C}_{k,N}^{\mathrm{T}} \mathbf{G}_N^{q-1} (\mathbf{I}_N + \psi \mathbf{E}_N)\right]$$

for $q = 1, \ldots, Q$ and $k = 1, \ldots, K$; and:

$$\lambda_{1+QK+p,N}(\boldsymbol{\theta}) \equiv \sigma^2 \mathrm{Tr}\left[(\mathbf{I}_N + \psi \mathbf{E}_N)^{\mathrm{T}} \mathbf{P}_{p,N} (\mathbf{I}_N + \psi \mathbf{E}_N)\right]$$

for $p = 1, \ldots, P$. For some $\boldsymbol{\theta}$, the sample moments are, simply:

$$\overline{\mathbf{m}}_N(\boldsymbol{\theta}) \equiv \frac{1}{N}\left[\mathbf{m}_N(\boldsymbol{\theta}) - \boldsymbol{\lambda}_N(\boldsymbol{\theta})\right] \tag{20}$$

while our GMM estimator $\widehat{\boldsymbol{\theta}}_{GMM}$ is the usual minimizer in the parameter space $\boldsymbol{\Theta}$:

$$\widehat{\boldsymbol{\theta}}_{GMM} = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \overline{\mathbf{m}}_N^{\mathrm{T}}(\boldsymbol{\theta}) \mathbf{W}_N \overline{\mathbf{m}}_N(\boldsymbol{\theta}) \tag{21}$$

where $\mathbf{W}_N$ is a weighting matrix. We derive the asymptotic properties of the estimator under the following additional assumptions.

Assumption 8. Bounded Parameter Space: $\boldsymbol{\Theta}$ *is bounded.*

Assumption 9. Probability Limits of the Covariates: *the independent component of x_{ik} are such that $N^{-1}\sum_{i=1}^{N}(\widetilde{x}_{ik} - \mathbb{E}[\widetilde{x}_{ik}]) = o_P(1)$ for all $k = 1, \ldots, K$.*

Assumptions 8 and 9 are regularity conditions that are necessary to ensure consistency of the GMM estimator.

Assumption 10. Bounded Characteristics: *matrix $\mathbf{C}_{k,N}$ is bounded by $\overline{C}_k < \infty$, that is $\sum_{j=1}^{N} \mathrm{c}_{k,ij} < \overline{C}_k$ for $i = 1, \ldots, N$, for all $k = 1, ..., K$.*

Assumption 11. Bounded Adjacencies: *the network's adjacency matrix $\mathbf{G}_N$ and its corresponding Leontiev inverse $(\mathbf{I}_N - \beta_0 \mathbf{G}_N)^{-1}$ are uniformly bounded in both row and column sums in absolute value.*

Assumption 12. Bounded Moment Matrices: *all the matrices $(\mathbf{Q}_{1,N}, \ldots, \mathbf{Q}_{Q,N})$ and $(\mathbf{P}_{1,N}, \ldots, \mathbf{P}_{P,N})$ used in the moment conditions are all uniformly bounded in both row and column sums in absolute value.*

Assumptions 10-12 all ensure that the moments in question have finite variance. Note that Assumptions 11-12 have their analogues in Lee (2007a), while Assumption 10 is specific to our framework. Also observe that bounded adjacencies are implied by the row normalization of $\mathbf{G}$. Yet Assumption 11 may be useful when row normalization is dropped, e.g. if interest falls on the separate identification of μ and ν.

The asymptotic properties of our GMM estimator are standard.

Theorem 2. Asymptotics of the GMM estimator. *Under Assumptions 1-12, $\widehat{\theta}_{GMM}$ is a consistent estimator of θ_0 and has the following limiting distribution:*

$$\sqrt{N}\left(\widehat{\theta}_{GMM}-\theta_0\right)\xrightarrow{d}\mathcal{N}\left(\mathbf{0},\left[\mathbf{J}_0^{\mathrm{T}}\mathbf{W}_0\mathbf{J}_0\right]^{-1}\mathbf{J}_0^{\mathrm{T}}\mathbf{W}_0\mathbf{\Omega}_0\mathbf{W}_0\mathbf{J}_0\left[\mathbf{J}_0^{\mathrm{T}}\mathbf{A}_0\mathbf{J}_0\right]^{-1}\right)$$

where: (i) $\mathbf{\Omega}_0\equiv\operatorname{plim}\frac{1}{N}\mathbb{V}\mathrm{ar}\left[\boldsymbol{m}_N\left(\theta_0\right)\right]$*;* (ii) $\mathbf{J}_0\equiv\operatorname{plim}\frac{\partial}{\partial\theta^{\mathrm{T}}}\overline{\mathbf{m}}_N\left(\theta_0\right)$*; and finally* (iii) $\mathbf{W}_0\equiv\operatorname{plim}\mathbf{W}_N$.

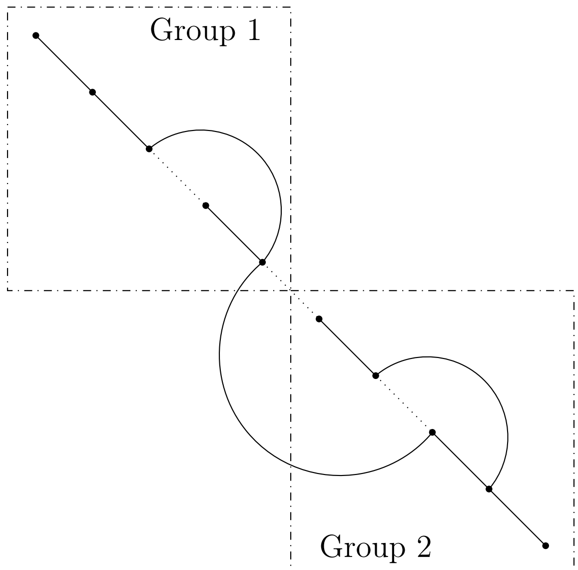
Proof. See the Appendix. The proof is based on the results by Lee (2007a), which in turn rely on White (1996) as well as Kelejian and Prucha (2001). □

The choice of the optimal weighting matrix $\mathbf{W}_N$ is informed by the same considerations advanced by Lee (2007a), to whom we refer for the details (we use a parallel notation for the moment matrices $\mathbf{Q}_{q,N}$ and $\mathbf{P}_{p,N}$ for ease of comparison). Extending this result to more general SARMA processes or to heteroscedastic primitive shocks is conceptually straightforward but analytically tedious.

5 Monte Carlo

We find it useful to evaluate the performance of our GMM estimator through Monte Carlo simulations. In particular, we simulate a minimal d.g.p.: the SAR model (6) with one covariate and no contextual effects ($\delta=0$), combined with the simple setup of linear endogeneity given in (8). We also let the error term to follow a simple first order spatial moving average process based on $\mathbf{G}$: $\boldsymbol{\varepsilon}=(\mathbf{I}+\psi\mathbf{G})\,\boldsymbol{\upsilon}$. In all simulations we set $N=500$; moreover we construct a homogeneous, block-diagonal characteristics matrix $\mathbf{C}$ which – in our baseline case – is composed by 50 "groups" of size 10. We generate a new matrix $\mathbf{G}$ in each repetition of every simulation, in order to minimize the dependence of our results from a specific network matrix. Specifically, each matrix $\mathbf{G}$ is randomly generated through the 'small-world' algorithm by Watts and Strogatz

(1998); by this procedure, all observations are first ordered along a line and connected to an even number of φ neighbors; next, links are reshuffled with some probability π (connections are unweighted, that is $g_{ij} \in \{0, 1\}$). Given that the initial ordering of observations corresponds with the one used for defining the characteristics matrix, **G** and **C** are guaranteed to have some degree of overlap, although not a complete one. We represent this through the following graphical example.



Graph 4: Partial overlap of **C** and **G**: Example

In Graph 4, 10 nodes are ordered along a line, and split in two symmetrical groups – each of size 5 – which characterize **C**. Through a small-world algorithm with $\varphi = 2$, all nodes are connected in the network with their immediate neighbors on the line, but three links are eventually reshuffled so that the resulting matrix **G** is irregular.

In our baseline simulation, we set the following parameters:[12]

$$(\alpha_0, \beta_0, \gamma_0, \xi_0, \psi_0, \sigma_0) = (.25, .4, .5, .1, .25, .05)$$

note that ψ_0 amounts to five times the standard deviation of the primitive shocks υ_i, which results in substantial endogeneity. In addition, we set $\varphi = 2$ and $\pi = 0.25$ in the network-generation algorithm. Over 1,000 repetitions, we estimate our model with equally-weighted moment conditions of order $Q = 3$ and $P = 2$, where $\mathbf{P}_1 = \mathbf{I}$ and $\mathbf{P}_1 = \mathbf{G}$. We also compare our estimates of (α, β, γ) with those obtained through

[12]Furthermore, we set $\mathbb{V}\text{ar}\left[\widetilde{x}_i\right] = 0.09$, but we are not interested in estimating this parameter.

OLS as well as through an IV estimator where $\mathbf{Gx}$ is used as an instrument for $\mathbf{Gy}$. Finally, we repeat the simulation by tuning certain parameters differently relative to the baseline. The results are reported in Tables 1 and 2.

Table 1: Monte Carlo Simulations (part one)

	Baseline			$\beta = 0.50$		
	PFZ	IV	OLS	PFZ	IV	OLS
α	0.256	0.082	0.044	0.254	0.081	0.041
	(0.040)	(0.018)	(0.015)	(0.038)	(0.020)	(0.015)
β	0.385	0.802	0.894	0.492	0.839	0.918
	(0.095)	(0.044)	(0.035)	(0.075)	(0.039)	(0.029)
γ	0.496	0.288	0.229	0.497	0.292	0.232
	(0.050)	(0.034)	(0.029)	(0.050)	(0.035)	(0.029)
ξ	0.100	–	–	0.100	–	–
	(0.013)			(0.013)		
ψ	0.270	–	–	0.260	–	–
	(0.010)			(0.090)		
σ	0.050	–	–	0.050	–	–
	(0.005)			(0.004)		

	$\gamma = 0.2$			$\psi = 0$		
	PFZ	IV	OLS	PFZ	IV	OLS
α	0.252	-0.144	-0.048	0.253	0.046	0.036
	(0.065)	(0.056)	(0.017)	(0.036)	(0.024)	(0.019)
β	0.395	1.347	1.115	0.392	0.891	0.913
	(0.156)	(0.133)	(0.042)	(0.086)	(0.057)	(0.046)
γ	0.188	-0.133	-0.040	0.499	0.190	0.176
	(0.042)	(0.064)	(0.027)	(0.053)	(0.046)	(0.038)
ξ	0.100	–	–	0.996	–	–
	(0.013)			(0.107)		
ψ	0.267	–	–	0.037	–	–
	(0.170)			(0.048)		
σ	0.050	–	–	0.050	–	–
	(0.005)			(0.004)		

Note. Every column reports the median and the standard deviation (in parentheses) of the relevant parameter estimates across 1000 repetitions. 'PFZ' indicates our proposed procedure, 'IV' the estimator obtained by instrumenting $\mathbf{Gy}$ with $\mathbf{Gx}$, while 'OLS' is self-explanatory.

Table 2: Monte Carlo Simulations (part two)

	$\xi = 0$			Group Size: 5		
	PFZ	IV	OLS	PFZ	IV	OLS
α	0.254	0.251	0.104	0.251	0.211	0.179
	(0.044)	(0.028)	(0.018)	(0.016)	(0.009)	(0.011)
β	0.389	0.398	0.749	0.397	0.494	0.571
	(0.106)	(0.067)	(0.044)	(0.038)	(0.022)	(0.026)
γ	0.499	0.498	0.303	0.497	0.292	0.232
	(0.055)	(0.043)	(0.037)	(0.050)	(0.035)	(0.029)
ξ	0.004	–	–	0.101	–	–
	(0.005)			(0.026)		
ψ	0.259	–	–	0.245	–	–
	(0.134)			(0.053)		
σ^2	0.050	–	–	0.050	–	–
	(0.002)			(0.002)		

	$\varphi = 4$			$\pi = 0.9$		
	PFZ	IV	OLS	PFZ	IV	OLS
α	0.263	0.027	-0.012	0.242	0.116	0.099
	(0.056)	(0.037)	(0.026)	(0.038)	(0.016)	(0.013)
β	0.368	0.935	1.029	0.420	0.723	0.762
	(0.135)	(0.088)	(0.061)	(0.092)	(0.037)	(0.032)
γ	0.506	0.226	0.173	0.491	0.418	0.402
	(0.060)	(0.058)	(0.044)	(0.034)	(0.025)	(0.023)
ξ	0.102	–	–	0.093	–	–
	(0.016)			(0.021)		
ψ	0.273	–	–	0.228	–	–
	(0.140)			(0.144)		
σ^2	0.050	–	–	0.049	–	–
	(0.003)			(0.004)		

Note. See the notes for Table 1.

In our baseline simulations, our proposed estimator appears to be quite accurate. While it slightly underestimates β (on average) it contrasts with both OLS and IV estimators, which estimate β about twice as large. We obtain similar results when we set different values of β or γ, or when the primitive shocks v_i and the error terms ε_i coincide ($\psi = 0$). If we silence the characteristics matrix channel ($\xi = 0$) IV becomes consistent; however, it behaves similarly as our proposed GMM estimator. The more interesting implications are obtained by altering the parameters that define matrices **C** and **G**. By halving the size of groups in the characteristics matrix, endogeneity is reduced; however, our GMM method still provides accurate estimates, unlike IV or OLS. Increasing the density of **G** (by setting $\varphi = 4$) does not seem to significantly affect the simulated estimates; however, increasing the randomness of links ($\pi = 0.9$) results in β to be slightly overestimated (instead of underestimated) on average. To summarize, it appears that our GMM method – while consistent and preferable to the standard IV estimator – is as usual biased in small samples, in a way that depends on the characteristics of the underlying networks.

6 Conclusion

In this paper we have shown that, under certain configurations of the underlying socio-economic relationships that determine the characteristics and relevant outcomes of economic agents, it is possible to identify and estimate peer or social effects within a standard spatial econometric framework, even if the right-hand side characteristics are themselves endogenous. The requirements for identification are quite general: it suffices that the network that characterizes social interactions is not fully-overlapping in only a slightly stronger sense than the identification conditions by Bramoullé et al. (2009), and that the spatial dimension of endogeneity (the dependence of individual covariates on peers' unobservables) is known by the econometrician. We believe that this approach can be applied to schooling settings where the friendship relationships of students are endogenous, while individual covariates are endogenously dependent on peers' unobservables in a way that can be determined by well-defined social groups, such as classrooms. In future work, we plan on implementing our GMM methodology to such an empirical application; in addition, we aim at extending it to more general functional forms of the underlying cross-correlations that give rise to endogeneity.

Our contribution is relevant for the developing literature about the econometrics of social effects, which is currently focused on ways to address the problem of network endogeneity. Our approach is more general than that, since it allows for both network and covariates' endogeneity – which, in fact, are two sides of the same coin. Paradoxically, our work also speaks to the recent literature about the identification of interaction networks when these are unknown, under the assumption that individual covariates are exogenous with respect to the error term. It may appear that our contribution goes in the opposite direction, as the application of our proposed estimator requires the researcher's knowledge of the underlying patterns of cross-correlation between observable and unobservable characteristics. However, we find that in future work it would be interesting to analyze how these two approaches can be interacted, say if some form of endogeneity is present in the relevant application but the researcher has access to only partial information about its actual spatial scope.

References

Angrist, Joshua D. (2014) "The perils of peer effects," *Labour Economics*, Vol. 30, pp. 98–108.

Arduini, Tiziano, Eleonora Patacchini, and Edoardo Rainone (2015) "Parametric and Semiparametric IV Estimation of Network Models with Selectivity." Mimeo.

Azoulay, Pierre, Joshua S. Graff Zivin, and Jialan Wang (2010) "Superstar Extinction," *The Quarterly Journal of Economics*, Vol. 125, No. 2, pp. 549–589.

Bloom, Nicholas, Mark Schankerman, and John Van Reenen (2013) "Identifying Technology Spillovers and Product Market Rivalry," *Econometrica*, Vol. 81, No. 4, pp. 1347–1393.

Blume, Lawrence E., William A. Brock, Steven N Durlauf, and Rajshri Jayaraman (2015) "Linear Social Interactions Models," *Journal of Political Economy*, Vol. 123, No. 2, pp. 444–496.

Bramoullé, Yann, Habiba Djebbari, and Bernard Fortin (2009) "Identification of peer effects through social networks," *Journal of Econometrics*, Vol. 150, No. 1, pp. 41–55.

Calvó-Armengol, Antoni, Eleonora Patacchini, and Yves Zenou (2009) "Peer Effects and Social Networks in Education," *The Review of Economic Studies*, Vol. 76, No. 4, pp. 1239–1267.

Carrell, Scott E., Bruce I. Sacerdote, and James E. West (2013) "From Natural Variation to Optimal Policy? The Importance of Endogenous Peer Group Formation," *Econometrica*, Vol. 81, No. 3, pp. 855–882.

Christakis, Nicholas A. and James H. Fowler (2013) "Social contagion theory: examining dynamic social networks and human behavior," *Statistics in medicine*, Vol. 32, No. 4, pp. 556–577.

Cliff, Andrew D. and John Keith Ord (1981) *Spatial Processes: Models and Applications*: Pion London.

Conley, Timothy G. and Christopher R. Udry (2010) "Learning about a new technology: Pineapple in Ghana," *The American Economic Review*, Vol. 100, No. 1, pp. 35–69.

Davezies, Laurent, Xavier D'Haultfoeuille, and Denis Fougère (2009) "Identification of peer effects using group size variation," *The Econometrics Journal*, Vol. 12, No. 3, pp. 397–413.

De Giorgi, Giacomo, Michele Pellizzari, and Silvia Redaelli (2010) "Identification of Social Interactions through Partially Overlapping Peer Groups," *American Economic Journal: Applied Economics*, Vol. 2, No. 2, pp. 241–275.

De Paula, Áureo, Imran Rasul, and Pedro Sousa (2018) "Recovering Social Networks from Outcome Data: Identification, Simulations and an Application." Mimeo.

Drukker, David M., Peter Egger, and Ingmar R. Prucha (2013) "On Two-Step Estimation of a Spatial Autoregressive Model with Autoregressive Disturbances and Endogenous Regressors," *Econometric Reviews*, Vol. 32, No. 5-6, pp. 686–733.

Duflo, Esther and Emmanuel Saez (2003) "The Role of Information and Social Interactions in Retirement Plan Decisions: Evidence from a Randomized Experiment," *The Quarterly Journal of Economics*, Vol. 118, No. 3, pp. 815–842.

Elhorst, Paul J. (2014) *Spatial Econometrics: From Cross-Sectional Data to Spatial Panels*: Springer.

Glaeser, Edward L., Bruce I. Sacerdote, and José A. Scheinkman (1996) "Crime and Social Interactions," *The Quarterly Journal of Economics*, Vol. 11, No. 2, pp. 507–48.

Graham, Bryan S. (2008) "Identifying Social Interactions Through Conditional Variance Restrictions," *Econometrica*, Vol. 76, No. 3, pp. 643–660.

——— (2017) "An Econometric Model of Network Formation with Degree Heterogeneity," *Econometrica*, Vol. 85, No. 4, pp. 1033–1063.

Jaffe, Adam B. (1986) "Technological opportunity and spillovers of R&D: evidence from firms' patents, profits and market value," *American Economic Review*, Vol. 76, No. 5, pp. 984–1001.

——— (1989) "Characterizing the technological position of firms, with application to quantifying technological opportunity and research spillovers," *Research Policy*, Vol. 18, No. 2, pp. 87–97.

Johnsson, Ida and Hyungsik Roger M. Moon (2017) “Estimation of Peer Effects in Endogenous Social Networks: Control Function Approach.” Mimeo.

Kapoor, Mudit, Harry H. Kelejian, and Ingmar R. Prucha (2007) “Panel data models with spatially correlated error components,” *Journal of Econometrics*, Vol. 140, No. 1, pp. 97–130.

Kelejian, Harry H. and Ingmar R. Prucha (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*, Vol. 17, No. 1, pp. 99–121.

——— (2001) “On the asymptotic distribution of the Moran I test statistic with applications,” *Journal of Econometrics*, Vol. 104, No. 2, pp. 219–257.

——— (2007) “HAC estimation in a spatial framework,” *Journal of Econometrics*, Vol. 140, No. 1, pp. 131–154.

——— (2010) “Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances,” *Journal of Econometrics*, Vol. 157, No. 1, pp. 53–67.

Kuersteiner, Guido M. and Ingmar Prucha (2018) “Dynamic Spatial Panel Models: Networks, Common Shocks, and Sequential Exogeneity.” Mimeo.

Lee, Lung-fei (2007a) “GMM and 2SLS estimation of mixed regressive, spatial autoregressive models,” *Journal of Econometrics*, Vol. 137, No. 2, pp. 489–514.

——— (2007b) “Identification and estimation of econometric models with group interactions, contextual factors and fixed effects,” *Journal of Econometrics*, Vol. 140, No. 2, pp. 333–374.

Lee, Lung-fei, Xiaodong Liu, and Xu Lin (2010) “Specification and estimation of social interaction models with network structures,” *The Econometrics Journal*, Vol. 13, No. 2, pp. 145–176.

Lee, Lung-fei and Xiaodong Liu (2010) “Efficient GMM estimation of high order spatial autoregressive models with autoregressive disturbances,” *Econometric Theory*, Vol. 26, No. 1, pp. 187–230.

Lin, Xu and Lung-fei Lee (2010) “GMM estimation of spatial autoregressive models with unknown heteroskedasticity,” *Journal of Econometrics*, Vol. 157, No. 1, pp. 34–52.

Liu, Xiaodong and Lung-fei Lee (2010) “GMM estimation of social interaction models with centrality,” *Journal of Econometrics*, Vol. 157, No. 1, pp. 34–52.

Lychagin, Sergey, Joris Pinkse, Margaret E. Slade, and John Van Reenen (2016) "Spillovers in Space: Does Geography Matter?" *The Journal of Industrial Economics*, Vol. 64, No. 2, pp. 295–335.

Manresa, Elena (2017) "Estimating the Structure of Social Interactions Using Panel Data." Mimeo.

Manski, Charles F. (1993) "Identification of Endogenous Social Effects: The Reflection Problem," *The Review of Economic Studies*, Vol. 60, No. 3, pp. 531–542.

Mas, Alexandre and Enrico Moretti (2009) "Peers at Work," *American Economic Review*, Vol. 99, No. 1, pp. 112–145.

Moffitt, Robert A. (2001) "Policy Interventions, Low-Level Equilibria And Social Interactions," in Steven Durlauf and Peyton Young eds. *Social Dynamics*: MIT Press.

Oettl, Alexander (2012) "Reconceptualizing Stars: Scientist Helpfulness and Peer Performance," *Management Science*, Vol. 56, No. 6, pp. 1122–1140.

Pereda Fernández, Santiago (2017) "Social Spillovers in the Classroom: Identification, Estimation and Policy Analysis," *Economica*, Vol. 84, No. 336, pp. 712–747.

Rose, Christiern D. (2017a) "Identification of peer effects through social networks using variance restrictions," *The Econometrics Journal*, Vol. 20, No. 3, pp. S47–S60.

——— (2017b) "Identification of Spillover Effects using Panel Data." Mimeo.

Sacerdote, Bruce I. (2001) "Peer Effects with Random Assignment: Results for Dartmouth Roommates," *The Quarterly Journal of Economics*, Vol. 116, No. 2, pp. 681–704.

Tibshirani, Robert (1996) "Regression Shrinkage and Selection via the LASSO," *Journal of the Royal Statistical Society*, Vol. 58, No. 1, pp. 267–288.

Waldinger, Fabian (2011) "Peer Effects in Science: Evidence from the Dismissal of Scientists in Nazi Germany," *Review of Economics Studies*, Vol. 72, No. 2, pp. 838–861.

Watts, Duncan J. and Steven H. Strogatz (1998) "Collective dynamics of 'small-world' networks," *Nature*, Vol. 393, No. 6684, pp. 440–442.

White, Halbert L. (1996) *Estimation, Inference and Specification Analysis*: Cambridge University Press.

Zacchia, Paolo (2018) "Knowledge Spillovers through Networks of Scientists." Mimeo.

Appendix – Mathematical Proofs

Proof of Theorem 1

We begin with a preliminary observation. We want to show that under the Theorem's condition (b), the matrix that characterizes the SARMA structure of the error:

$$\mathbf{\Psi}_{AM}\left(\boldsymbol{\phi},\boldsymbol{\psi}\right)\equiv\left(\mathbf{I}-\phi_1\mathbf{F}_1-\cdots-\phi_A\mathbf{F}_A\right)^{-1}\left(\mathbf{I}+\psi_1\mathbf{E}_1+\cdots+\phi_M\mathbf{E}_M\right)$$

and all its derivatives with respect to $(\boldsymbol{\phi},\boldsymbol{\psi})$ are linearly independent. To this end, recall the following matrix defined in Assumption 5:

$$\mathbf{\Phi}_A\equiv\left(\mathbf{I}-\phi_1\mathbf{F}_1-\cdots-\phi_A\mathbf{F}_A\right)$$

and write it as $\mathbf{\Phi}_A\left(\boldsymbol{\phi}\right)$. By the rules for the derivatives of inverted matrices, we can write, for $a=1,\dots,A$:

$$\frac{\partial\mathbf{\Psi}_{A,M}\left(\boldsymbol{\phi},\boldsymbol{\psi}\right)}{\partial\phi_a}=-\mathbf{\Phi}_A^{-1}\left(\boldsymbol{\phi}\right)\cdot\mathbf{F}_a\cdot\mathbf{\Phi}_A^{-1}\left(\boldsymbol{\phi}\right)\cdot\left(\mathbf{I}+\psi_1\mathbf{E}_1+\cdots+\phi_M\mathbf{E}_M\right)$$

and, for $m=1,\dots,M$:

$$\frac{\partial\mathbf{\Psi}_{A,M}\left(\boldsymbol{\phi},\boldsymbol{\psi}\right)}{\partial\psi_m}=\mathbf{\Phi}_A^{-1}\left(\boldsymbol{\phi}\right)\cdot\mathbf{E}_m$$

it is straightforward to see that under the hypotheses made in Assumption 5 about the $\mathbf{F}_a$ and $\mathbf{E}_m$ matrices, all the derivatives with respect to $\boldsymbol{\phi}$ are linearly independent of one another, and so are all the derivatives with respect to $\boldsymbol{\psi}$. In order to demonstrate the desired result, it must be verified that the matrices in these two sets are linearly independent of one another too, as well as on matrix $\mathbf{\Psi}_{A,M}\left(\boldsymbol{\phi},\boldsymbol{\psi}\right)$ itself. Clearly, this is the case if condition (b) holds. This requirement must be verified on a case-by-case basis; however, the usual expansion for generalized Leontiev inverses, which converges so long as $\left\|\boldsymbol{\phi}\right\|<1$ and that reads:

$$\mathbf{\Phi}_A^{-1}\left(\boldsymbol{\phi}\right)=\left(\mathbf{I}-\phi_1\mathbf{F}_1-\cdots-\phi_A\mathbf{F}_A\right)^{-1}\simeq\sum_{s=1}^{\infty}\left(\phi_1\mathbf{F}_1+\cdots+\phi_A\mathbf{F}_A\right)^s$$

can also come to rescue in selected cases. To illustrate, consider the relatively simple circumstance where $(A,M)=(1,1)$ and $\mathbf{F}_1=\mathbf{E}_1=\mathbf{G}$. In this case, condition (b) requires that the adjacency matrix $\mathbf{G}$ is linearly independent of $\sum_{s=0}^{\infty}\phi_1^s\mathbf{G}^{s+1}\left(\mathbf{I}+\psi_1\mathbf{G}\right)$ which is the usually the case for typical intransitive, sparse networks. Finally, observe that if matrix $\mathbf{\Psi}_{A,M}\left(\boldsymbol{\phi},\boldsymbol{\psi}\right)$ as well as all its derivatives are linearly independent, so are their vectorized versions following the application of the $\operatorname{vec}\left(\cdot\right)$ operator. In fact, we use for convenience the vectorized versions in the ensuing demonstration.

We now move to our main identification proof. Let

$$\boldsymbol{\varepsilon}(\boldsymbol{\theta}) \equiv \mathbf{y} - \alpha\boldsymbol{\iota} - \beta\mathbf{G}\mathbf{y} - \mathbf{X}\boldsymbol{\gamma} - \mathbf{G}\mathbf{X}\boldsymbol{\delta}$$

and, for notational consistency, write the previous SARMA matrix $\boldsymbol{\Psi}_{A,M}(\boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$. Consider the following set of moments:

$$\begin{aligned} \mathbb{E}\left[\boldsymbol{\iota}^{\mathrm{T}}\boldsymbol{\varepsilon}(\boldsymbol{\theta})\right] &= 0 \\ \mathbb{E}\left[\mathbf{x}_k^{\mathrm{T}}\mathbf{G}^{q-1}\boldsymbol{\varepsilon}(\boldsymbol{\theta})\right] - \lambda_{1,qk}(\boldsymbol{\theta}) &= 0 \qquad \text{for } q = 1, \dots, Q \text{ and } k = 1, \dots, K \\ \mathbb{E}\left[\boldsymbol{\varepsilon}(\boldsymbol{\theta})\mathbf{P}^p\boldsymbol{\varepsilon}(\boldsymbol{\theta})\right] - \lambda_{2,p}(\boldsymbol{\theta}) &= 0 \qquad \text{for } p = 1, \dots, P \end{aligned}$$

where:

$$\begin{aligned} \lambda_{1,qk}(\boldsymbol{\theta}) &\equiv \sigma^2\xi_k \mathrm{Tr}\left(\mathbf{C}_k^{\mathrm{T}}\mathbf{G}^{q-1} \cdot \boldsymbol{\Psi}_{A,M}(\boldsymbol{\theta})\right) \\ \lambda_{2,p}(\boldsymbol{\theta}) &\equiv \sigma^2 \mathrm{Tr}\left(\boldsymbol{\Psi}_{A,M}^{\mathrm{T}}(\boldsymbol{\theta}) \cdot \mathbf{P}^p \cdot \boldsymbol{\Psi}_{A,M}(\boldsymbol{\theta})\right) \end{aligned}$$

think of them as being stacked vertically in the vector $\mathbb{E}\left[\mathbf{m}(\boldsymbol{\theta}) - \boldsymbol{\lambda}(\boldsymbol{\theta})\right] = \mathbf{0}$ of length $1 + QK + P$. We evaluate a just-identified case with $Q = 4$ and $P = 1 + A + M$; to this end, partition the Jacobian matrix of the moments in four blocks, by splitting it horizontally after the first $1 + (Q-1)K$ rows and vertically after the first $1 + QK$ columns:

$$\mathbf{H}(\boldsymbol{\theta}) = \begin{bmatrix} \mathbf{H}_{11}(\boldsymbol{\theta}) & \mathbf{H}_{12}(\boldsymbol{\theta}) \\ \mathbf{H}_{21}(\boldsymbol{\theta}) & \mathbf{H}_{22}(\boldsymbol{\theta}) \end{bmatrix}$$

we will analyze each of these blocks in sequence.

The upper-left block is standard:

$$\mathbf{H}_{11}(\boldsymbol{\theta}) = -\mathbb{E}\begin{bmatrix} N & \boldsymbol{\iota}^{\mathrm{T}}\mathbf{G}\mathbf{y} & \boldsymbol{\iota}^{\mathrm{T}}\mathbf{X} & \boldsymbol{\iota}^{\mathrm{T}}\mathbf{G}\mathbf{X} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{x}_k^{\mathrm{T}}\mathbf{G}^{q-1}\boldsymbol{\iota} & \mathbf{x}_k^{\mathrm{T}}\mathbf{G}^{q}\mathbf{y} & \mathbf{x}_k^{\mathrm{T}}\mathbf{G}^{q-1}\mathbf{X} & \mathbf{x}_k^{\mathrm{T}}\mathbf{G}^{q}\mathbf{X} \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

for $k = 1, \dots, K$ and $q = 1, \dots, Q$. This block has full column and full row ranks if the data $(\mathbf{X}, \mathbf{y})$ are not collinear and the four matrices $\mathbf{I}$, $\mathbf{G}$, $\mathbf{G}^2$ and $\mathbf{G}^3$ are linearly independent. The lower-left block is also quite regular; for $p = 1, \dots, P$:

$$\mathbf{H}_{21}(\boldsymbol{\theta}) = -2 \cdot \mathbb{E}\begin{bmatrix} \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{\varepsilon}^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{P}^p\boldsymbol{\iota} & \boldsymbol{\varepsilon}^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{P}^p\mathbf{G}\mathbf{y} & \boldsymbol{\varepsilon}^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{P}^p\mathbf{X} & \boldsymbol{\varepsilon}^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{P}^p\mathbf{G}\mathbf{X} \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

which also has full row and column rank when the data are not collinear and the $\mathbf{P}^p$ matrices are linearly independent (condition (a) in the text). Note that for a given $\boldsymbol{\theta}$, in this case the rows of the upper-left block $\mathbf{H}_{11}(\boldsymbol{\theta})$ are also linearly independent of those in the lower-left block $\mathbf{H}_{21}(\boldsymbol{\theta})$, as standard when linear moments are juxtaposed to second order moments.

The upper-right block compares with the last column of the Jacobian matrix from Proposition 3:

$$\mathbf{H}_{12}(\boldsymbol{\theta}) = \begin{bmatrix} \mathbf{0}^{\mathrm{T}} & \mathbf{0}^{\mathrm{T}} & \mathbf{0}^{\mathrm{T}} & 0 \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial \lambda_{1,qk}(\boldsymbol{\theta})}{\partial \boldsymbol{\xi}^{\mathrm{T}}} & \frac{\partial \lambda_{1,qk}(\boldsymbol{\theta})}{\partial \boldsymbol{\phi}^{\mathrm{T}}} & \frac{\partial \lambda_{1,qk}(\boldsymbol{\theta})}{\partial \boldsymbol{\psi}^{\mathrm{T}}} & \frac{\partial \lambda_{1,qk}(\boldsymbol{\theta})}{\partial \sigma^2} \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

for $k = 1, \dots, K$ and $q = 1, \dots, Q$; the derivatives in each row are as follows.

$$\begin{aligned} \frac{\partial \lambda_{1,qk}(\boldsymbol{\theta})}{\xi_{k'}} &= \sigma^2 \operatorname{vec}\left(\mathbf{C}_k^{\mathrm{T}}\right)^{\mathrm{T}} \left(\mathbf{I} \otimes \mathbf{G}^{q-1}\right) \operatorname{vec}\left(\boldsymbol{\Psi}_{A,M}(\boldsymbol{\theta})\right) \cdot \mathbb{1}\left[k = k'\right] \quad && \text{for } k' = 1, \dots, K \\ \frac{\partial \lambda_{1,qk}(\boldsymbol{\theta})}{\phi_a} &= \sigma^2 \xi_k \operatorname{vec}\left(\mathbf{C}_k^{\mathrm{T}}\right)^{\mathrm{T}} \left(\mathbf{I} \otimes \mathbf{G}^{q-1}\right) \operatorname{vec}\left(\frac{\partial \boldsymbol{\Psi}_{A,M}(\boldsymbol{\theta})}{\partial \phi_a}\right) && \text{for } a = 1, \dots, A \\ \frac{\partial \lambda_{1,qk}(\boldsymbol{\theta})}{\psi_m} &= \sigma^2 \xi_k \operatorname{vec}\left(\mathbf{C}_k^{\mathrm{T}}\right)^{\mathrm{T}} \left(\mathbf{I} \otimes \mathbf{G}^{q-1}\right) \operatorname{vec}\left(\frac{\partial \boldsymbol{\Psi}_{A,M}(\boldsymbol{\theta})}{\partial \psi_m}\right) && \text{for } m = 1, \dots, M \\ \frac{\partial \lambda_{1,qk}(\boldsymbol{\theta})}{\sigma^2} &= \xi_k \operatorname{vec}\left(\mathbf{C}_k^{\mathrm{T}}\right)^{\mathrm{T}} \left(\mathbf{I} \otimes \mathbf{G}^{q-1}\right) \operatorname{vec}\left(\boldsymbol{\Psi}_{A,M}(\boldsymbol{\theta})\right) \end{aligned}$$

Notice that some rows of $\mathbf{H}_{12}(\boldsymbol{\theta})$ might be linearly dependent if any pair of matrices $\mathbf{C}_k$ coincide, but this does not affect our main argument. In fact, we focus on column rank for the two right blocks as it relates with the identification of the $(\boldsymbol{\xi}, \boldsymbol{\phi}, \boldsymbol{\psi}, \sigma^2)$ parameters. Thus, it is helpful to consider the lower-right block as well:

$$\mathbf{H}_{22}(\boldsymbol{\theta}) = \begin{bmatrix} \vdots & \vdots & \vdots & \vdots \\ \mathbf{0}^{\mathrm{T}} & \frac{\partial \lambda_{2,p}(\boldsymbol{\theta})}{\partial \boldsymbol{\phi}^{\mathrm{T}}} & \frac{\partial \lambda_{2,p}(\boldsymbol{\theta})}{\partial \boldsymbol{\psi}^{\mathrm{T}}} & \frac{\partial \lambda_{2,p}(\boldsymbol{\theta})}{\partial \sigma^2} \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

for $p = 1, \dots, P$, with the following derivatives.

$$\begin{aligned} \frac{\partial \lambda_{2,p}(\boldsymbol{\theta})}{\phi_a} &= 2\sigma^2 \operatorname{vec}\left(\boldsymbol{\Psi}_{A,M}^{\mathrm{T}}(\boldsymbol{\theta})\right)^{\mathrm{T}} \left(\mathbf{I} \otimes \mathbf{P}^p\right) \operatorname{vec}\left(\frac{\partial \boldsymbol{\Psi}_{A,M}(\boldsymbol{\theta})}{\partial \phi_a}\right) \quad && \text{for } a = 1, \dots, A \\ \frac{\partial \lambda_{2,p}(\boldsymbol{\theta})}{\psi_m} &= 2\sigma^2 \operatorname{vec}\left(\boldsymbol{\Psi}_{A,M}^{\mathrm{T}}(\boldsymbol{\theta})\right)^{\mathrm{T}} \left(\mathbf{I} \otimes \mathbf{P}^p\right) \operatorname{vec}\left(\frac{\partial \boldsymbol{\Psi}_{A,M}(\boldsymbol{\theta})}{\partial \psi_m}\right) && \text{for } m = 1, \dots, M \\ \frac{\partial \lambda_{2,p}(\boldsymbol{\theta})}{\sigma^2} &= \operatorname{vec}\left(\boldsymbol{\Psi}_{A,M}^{\mathrm{T}}(\boldsymbol{\theta})\right)^{\mathrm{T}} \left(\mathbf{I} \otimes \mathbf{P}^p\right) \operatorname{vec}\left(\boldsymbol{\Psi}_{A,M}(\boldsymbol{\theta})\right) \end{aligned}$$

Observe how the submatrix of $\mathbf{H}(\boldsymbol{\theta})$ formed by the two blocks $\mathbf{H}_{12}(\boldsymbol{\theta})$ and $\mathbf{H}_{22}(\boldsymbol{\theta})$ has full column rank. In fact, the previous observation about matrix $\boldsymbol{\Psi}_{A,M}(\boldsymbol{\phi}, \boldsymbol{\psi})$ implies that its vectorized version and that of its derivatives with respect to $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ are all linearly independent; in addition, the linear independence of all the columns that contain the derivatives of $\lambda_{1,qk}(\boldsymbol{\theta})$ and $\lambda_{2,p}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\xi}$ and σ^2 is guaranteed by the sparsity structure of the first K columns of both blocks. The only case where full column rank would fail is when any of the first K columns of $\mathbf{H}_{12}(\boldsymbol{\theta})$, those that contain the derivatives of $\lambda_{1,qk}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\xi}$, features all zeros. However, this possibility is ruled out by condition (c) in the statement of Theorem. Like its analog for Proposition 3 this is a moot requirement, as its violation would imply that there is no meaningful endogeneity for some covariates of interest.

It remains to show that the the entire Jacobian $\mathbf{H}(\boldsymbol{\theta})$ is overall nonsingular. Full row rank is guaranteed by the previous observation that the rows of the left blocks of $\mathbf{H}_{11}(\boldsymbol{\theta})$ and $\mathbf{H}_{21}(\boldsymbol{\theta})$ are linearly independent. Full column rank instead might fail if any of the columns of the upper-right block $\mathbf{H}_{12}(\boldsymbol{\theta})$, which contain the derivatives of $\lambda_{1,qk}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\xi}$, is collinear with some other column of the corresponding upper-left block $\mathbf{H}_{11}(\boldsymbol{\theta})$. This is generally a strong requirement: to illustrate, suppose that $\boldsymbol{\Psi}_{A,M}(\boldsymbol{\theta}) = \mathbf{I}$; in analogy with Proposition 3, full column rank fails if, for some k', vector $\operatorname{vec}(\mathbf{C}_{k'})$ is perfectly collinear with any of $\operatorname{vec}\left(\boldsymbol{\iota}\mathbf{x}_{k'}^{\mathrm{T}}\right)$, $\operatorname{vec}\left(\mathbf{x}_k\mathbf{x}_{k'}^{\mathrm{T}}\right)$, $\operatorname{vec}\left(\mathbf{x}_k\mathbf{G}\mathbf{x}_{k'}^{\mathrm{T}}\right)$, $\operatorname{vec}\left(\mathbf{G}\mathbf{y}\mathbf{x}_{k'}^{\mathrm{T}}\right)$, for $k = 1, \ldots, K$. In words, some characteristic matrix $\mathbf{C}_k$ must perfectly predict a specific configuration of some subset of the data. However, the variation of each of the covariates $\mathbf{x}_k$ (and thus ultimately of the outcome $\mathbf{y}$) is largely determined by the independent component $\widetilde{\mathbf{x}}_k$, which takes values on a continuous support. Since all the characteristic matrices are non-stochastic, it follows that such an instance of perfect multicollinearity has probability zero. Therefore, the Jacobian matrix $\mathbf{H}(\boldsymbol{\theta})$ is almost surely invertible and the parameter set $\boldsymbol{\theta}$ is almost always identified.

Proof of Theorem 2

In this proof, we denote by $\mathbf{x}_{k,N}^{*}$ the k-th column of $\mathbf{X}_N$ for $k = 1, \ldots, K$; its expected value $\mathbb{E}\left[\mathbf{x}_{k,N}\right] = \mathbb{E}\left[\widetilde{\mathbf{x}}_{k,N}^{*}\right]$, where $\widetilde{\mathbf{x}}_{k,N}$ is defined in Assumption 6, corresponds with the k-th column of $\mathbb{E}\left[\mathbf{X}_N\right]$. We also write the unconditional expected value of $\mathbf{y}_N$ as follows.

$$\mathbb{E}\left[\mathbf{y}_N\right] = \left(\mathbf{I}_N - \beta_0\mathbf{G}_N\right)^{-1}\left(\alpha_0\boldsymbol{\iota}_N + \mathbb{E}\left[\mathbf{X}_N\right]\boldsymbol{\gamma}_0 + \mathbf{G}_N\mathbb{E}\left[\mathbf{X}_N\right]\boldsymbol{\delta}_0\right)$$

We also introduce some additional auxiliary notation; to begin, the matrix:

$$\widetilde{\mathbf{G}}_N(\beta) \equiv \mathbf{G}_N\left(\mathbf{I}_N - \beta\mathbf{G}_N\right)^{-1}$$

is immediately helpful towards the definition of the following vectors:

$$\begin{aligned}
\mathbf{d}_N(\boldsymbol{\theta}) &\equiv (\alpha_0 - \alpha)\,\boldsymbol{\iota}_N + (\beta_0 - \beta)\,\mathbf{G}_N\mathbb{E}\left[\mathbf{y}_N\right] + \mathbb{E}\left[\mathbf{X}_N\right](\boldsymbol{\gamma}_0 - \boldsymbol{\gamma}) + \mathbf{G}_N\mathbb{E}\left[\mathbf{X}_N\right](\boldsymbol{\delta}_0 - \boldsymbol{\delta}) \\
\mathbf{e}_N(\boldsymbol{\theta}) &\equiv \boldsymbol{\varepsilon}_N + \left(\mathbf{X}_N - \mathbb{E}\left[\mathbf{X}_N\right]\right)(\boldsymbol{\gamma}_0 - \boldsymbol{\gamma}) + \mathbf{G}_N\left(\mathbf{X}_N - \mathbb{E}\left[\mathbf{X}_N\right]\right)(\boldsymbol{\delta}_0 - \boldsymbol{\delta}) \\
&\quad + (\beta_0 - \beta)\,\widetilde{\mathbf{G}}_N(\beta_0)\left[\boldsymbol{\varepsilon}_N + \left(\mathbf{X}_N - \mathbb{E}\left[\mathbf{X}_N\right]\right)\boldsymbol{\gamma}_0 + \mathbf{G}_N\left(\mathbf{X}_N - \mathbb{E}\left[\mathbf{X}_N\right]\right)\boldsymbol{\delta}_0\right]
\end{aligned}$$

note that $\boldsymbol{\varepsilon}_N(\boldsymbol{\theta}) = \mathbf{d}_N(\boldsymbol{\theta}) + \mathbf{e}_N(\boldsymbol{\theta})$; furthermore, the following K matrices will be helpful throughout:

$$\boldsymbol{\Gamma}_{k,0}(\boldsymbol{\theta}) \equiv \left[(\gamma_{k,0} - \gamma_k)\mathbf{I}_N + (\delta_{k,0} - \delta_k)\mathbf{G}_N + (\beta_0 - \beta)\widetilde{\mathbf{G}}_N(\beta_0)(\gamma_{k,0}\mathbf{I}_N + \delta_{k,0}\mathbf{G}_N)\right]$$

where $k = 1, \dots, K$. Finally, observe that the GMM weighting matrix $\mathbf{W}_N$ can be written as:

$$\mathbf{W}_N = \mathbf{A}_N^{\mathrm{T}}\mathbf{A}_N$$

where $\mathbf{A}_N$ is a square matrix of dimension $1 + QK + P$ and such that $\mathbf{A}_N \xrightarrow{p} \mathbf{A}_0$ and $\operatorname{rank}(\mathbf{A}_N) \geq \dim|\boldsymbol{\theta}|$, where $\mathbf{A}_0^{\mathrm{T}}\mathbf{A}_0 = \mathbf{W}_0$. This implies that the vector $\mathbf{A}_N\mathbf{m}_N(\boldsymbol{\theta})$ can be decomposed as:

$$\begin{aligned} \frac{1}{N}\mathbf{A}_N\mathbf{m}_N(\boldsymbol{\theta}) = \frac{1}{N}a_{1,N}\boldsymbol{\iota}_N^{\mathrm{T}} + \frac{1}{N}&\left[\sum_{q=1}^{Q}\sum_{k=1}^{K} a_{1+(q-1)K+k,N}\mathbf{q}_{qk,N}\right. \\ &\left. + \sum_{p=1}^{P} a_{1+QK+p,N}\boldsymbol{\varepsilon}_N^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{P}_{p,N}\right]\boldsymbol{\varepsilon}_N(\boldsymbol{\theta}) \end{aligned} \tag{A.1}$$

where the $1 + QK + P$ elements written as $a_{\cdot,N}$ are appropriate combinations of the elements of $\mathbf{A}_N$. Our main proof of consistency is based on this decomposition; later we refer to the "first" and the "second" element of (A.1) as the two summations laid out within brackets on the first and second line of the above, respectively.

Before we get to said proof, one last preparatory step is useful. We later further decompose the elements of (A.1) into smaller bits as linear functions of some auxiliary vectors and matrices. It helpful to introduce these arrays immediately, so to not break the flow of our main argument later. They are: (i) some $K(1+K)$ matrices, which are written as $\mathbf{R}_{k,N}^{*}(\boldsymbol{\theta})$ and $\mathbf{R}_{kk',N}^{*}(\boldsymbol{\theta})$, and are indexed by $k, k' = 1, \dots, K$:

$$\mathbf{R}_{k,N}^{*}(\boldsymbol{\theta}) \equiv \sum_{q=1}^{Q}\mathbf{G}_N^{q-1}a_{1+qk,N}\left[\mathbf{I}_N + (\beta_0 - \beta)\widetilde{\mathbf{G}}_N(\beta_0)\right]$$

$$\mathbf{R}_{kk',N}^{*}(\boldsymbol{\theta}) \equiv \sum_{q=1}^{Q}\mathbf{G}_N^{q-1}a_{1+qk,N}\boldsymbol{\Gamma}_{k,0}(\boldsymbol{\theta})$$

(ii) a set of $K+1$ vectors written as $\mathbf{l}_{0,N}^{**}(\boldsymbol{\theta})$ and as $\mathbf{l}_{k,N}^{**}(\boldsymbol{\theta})$ for $k = 1, \dots, K$;

$$\mathbf{l}_{0,N}^{**}(\boldsymbol{\theta}) \equiv \mathbf{d}_N^{\mathrm{T}}(\boldsymbol{\theta})\sum_{p=1}^{P}a_{1+QK+p,N}\mathbf{P}_{p,N}\left[\mathbf{I}_N + (\beta_0 - \beta)\widetilde{\mathbf{G}}_N(\beta_0)\right]$$

$$\mathbf{l}_{k,N}^{**}(\boldsymbol{\theta}) \equiv \mathbf{d}_N^{\mathrm{T}}(\boldsymbol{\theta})\sum_{p=1}^{P}a_{1+QK+p,N}\mathbf{P}_{p,N}\boldsymbol{\Gamma}_{k,0}(\boldsymbol{\theta})$$

(iii) another set of $1+K+K^2$ matrices, written as $\mathbf{R}^{**}_{0,N}(\boldsymbol{\theta})$, $\mathbf{R}^{**}_{k,N}(\boldsymbol{\theta})$ and $\mathbf{R}^{**}_{kk',N}(\boldsymbol{\theta})$, and indexed by $k,k'=1,\dots,K$.

$$\mathbf{R}^{**}_{0,N}(\boldsymbol{\theta}) \equiv \left(\mathbf{I}_N + (\beta_0-\beta)\widetilde{\mathbf{G}}^{\mathrm{T}}_N(\beta_0)\right)\sum_{p=1}^{P} a_{1+QK+p,N}\mathbf{P}_{p,N}\left(\mathbf{I}_N + (\beta_0-\beta)\widetilde{\mathbf{G}}_N(\beta_0)\right)$$

$$\mathbf{R}^{**}_{k,N}(\boldsymbol{\theta}) \equiv \left(\mathbf{I}_N + (\beta_0-\beta)\widetilde{\mathbf{G}}^{\mathrm{T}}_N(\beta_0)\right)\sum_{p=1}^{P} a_{1+QK+p,N}\left(\mathbf{P}_{p,N}+\mathbf{P}^{\mathrm{T}}_{p,N}\right)\boldsymbol{\Gamma}_{k,0}(\boldsymbol{\theta})$$

$$\mathbf{R}^{**}_{kk',N}(\boldsymbol{\theta}) \equiv \boldsymbol{\Gamma}^{\mathrm{T}}_{k',0}(\boldsymbol{\theta})\left[\sum_{p=1}^{P} a_{1+QK+p,N}\mathbf{P}_{p,N}\right]\boldsymbol{\Gamma}_{k,0}(\boldsymbol{\theta})$$

We now proceed to our main argument. In order to establish consistency of $\widehat{\boldsymbol{\theta}}_{GMM}$, it is necessary to show uniform convergence in probability for all the elements that comprise the vector $\mathbf{A}_N\mathbf{m}_N(\boldsymbol{\theta})$. Consider the first element in brackets in (A.1):

$$\sum_{q=1}^{Q}\sum_{k=1}^{K} a_{1+(q-1)K+k,N}\mathbf{q}_{qk,N}\boldsymbol{\varepsilon}_N(\boldsymbol{\theta}) = \underbrace{\sum_{q=1}^{Q}\sum_{k=1}^{K} a_{1+(q-1)K+k,N}\left(\mathbf{G}^{q-1}_N\mathbf{x}^*_{k,N}\right)^{\mathrm{T}}\mathbf{d}_N(\boldsymbol{\theta})}_{\equiv l^*_N(\boldsymbol{\theta})} + \underbrace{\sum_{q=1}^{Q}\sum_{k=1}^{K} a_{1+(q-1)K+q,N}\left(\mathbf{G}^{q-1}_N\mathbf{x}^*_{k,N}\right)^{\mathrm{T}}\mathbf{e}_N(\boldsymbol{\theta})}_{\equiv r^*_N(\boldsymbol{\theta})}$$

where $l^*_N(\boldsymbol{\theta})$ is given by:

$$\frac{1}{N}l^*_N(\boldsymbol{\theta}) = \frac{1}{N}\sum_{q=1}^{Q}\sum_{k=1}^{K} a_{1+(q-1)K+k,N}\left(\mathbf{G}^{q-1}_N\mathbb{E}\left[\mathbf{x}^*_{k,N}\right]\right)^{\mathrm{T}}\mathbf{d}^{\mathrm{T}}_N(\boldsymbol{\theta}) + o_P(1)$$

while $r^*_N(\boldsymbol{\theta})$ can be expressed as a function of the $\mathbf{R}^*_{k,N}(\boldsymbol{\theta})$ and $\mathbf{R}^*_{kk',N}(\boldsymbol{\theta})$ matrices defined above (note: the second line continues on the next page).

$$\begin{aligned}\frac{1}{N}r^*_N(\boldsymbol{\theta}) &= \frac{1}{N}\sum_{k=1}^{K}\left(\mathbf{x}^*_{k,N}-\mathbb{E}\left[\mathbf{x}^*_{k,N}\right]\right)^{\mathrm{T}}\mathbf{R}^*_{k,N}(\boldsymbol{\theta})\left(\mathbf{x}^*_{k,N}-\mathbb{E}\left[\mathbf{x}^*_{k,N}\right]\right)\\ &\quad+\frac{1}{N}\sum_{k=1}^{K}\sum_{k'=1}^{K}\left(\mathbf{x}^*_{k',N}-\mathbb{E}\left[\mathbf{x}^*_{k',N}\right]\right)^{\mathrm{T}}\mathbf{R}^*_{k,k',N}(\boldsymbol{\theta})\boldsymbol{\varepsilon}_N\\ &= \sigma^2_0\frac{1}{N}\sum_{k=1}^{K}\xi_{0,k}\mathrm{Tr}\left(\mathbf{C}^{\mathrm{T}}_{k,N}\mathbf{R}^*_{k,N}(\boldsymbol{\theta})\left(\mathbf{I}_N+\psi_0\mathbf{E}_N\right)\right)\end{aligned}$$

$$+ \frac{1}{N} \sum_{k=1}^{K} \sum_{k'=1}^{K} \xi_{0,k} \xi_{0,k'} \operatorname{Tr} \left(\mathbf{C}_{k,N}^{\mathrm{T}} \mathbf{R}_{k,k',N}^{*} \left(\boldsymbol{\theta} \right) \mathbf{C}_{k,N} \right)$$

$$+ \frac{1}{N} \sum_{k=1}^{K} \sum_{k'=1}^{K} \Xi_{kk'} \cdot \operatorname{Tr} \left(\mathbf{R}_{k,k',N}^{*} \left(\boldsymbol{\theta} \right) \right) + o_P \left(1 \right)$$

Similarly, the second term in brackets in (A.1) can be decomposed as:

$$\sum_{p=1}^{P} a_{1+QK+p,N} \boldsymbol{\varepsilon}_N^{\mathrm{T}} \left(\boldsymbol{\theta} \right) \mathbf{P}_{p,N} \boldsymbol{\varepsilon}_N \left(\boldsymbol{\theta} \right) = \sum_{p=1}^{P} a_{1+QK+p,N} \mathbf{d}_N^{\mathrm{T}} \left(\boldsymbol{\theta} \right) \mathbf{P}_{p,N} \mathbf{d}_N \left(\boldsymbol{\theta} \right)$$

$$+ 2 \underbrace{\sum_{p=1}^{P} a_{1+QK+p,N} \mathbf{d}_N^{\mathrm{T}} \left(\boldsymbol{\theta} \right) \mathbf{P}_{p,N} \mathbf{e}_N \left(\boldsymbol{\theta} \right)}_{\equiv l_N^{**} \left(\boldsymbol{\theta} \right)} + \underbrace{\sum_{p=1}^{P} a_{1+QK+p,N} \mathbf{e}_N^{\mathrm{T}} \left(\boldsymbol{\theta} \right) \mathbf{P}_{p,N} \mathbf{e}_N \left(\boldsymbol{\theta} \right)}_{\equiv r_N^{**} \left(\boldsymbol{\theta} \right)}$$

where $l_N^{**} \left(\boldsymbol{\theta} \right)$ is written in terms of $\mathbf{l}_{0,N}^{**} \left(\boldsymbol{\theta} \right)$ and $\mathbf{l}_{k,N}^{**} \left(\boldsymbol{\theta} \right)$:

$$\frac{1}{N} l_N^{**} \left(\boldsymbol{\theta} \right) = \frac{1}{N} \mathbf{l}_{0,N} \left(\boldsymbol{\theta} \right) \boldsymbol{\varepsilon}_N + \frac{1}{N} \sum_{k=1}^{K} \mathbf{l}_{k,N} \left(\boldsymbol{\theta} \right) \left(\mathbf{x}_{k,N}^{*} - \mathbb{E} \left[\mathbf{x}_{k,N}^{*} \right] \right) = o_P \left(1 \right)$$

while the term $r_N^{**} \left(\boldsymbol{\theta} \right)$ can be related to $\mathbf{R}_{0,N}^{**} \left(\boldsymbol{\theta} \right)$, $\mathbf{R}_{k,N}^{**} \left(\boldsymbol{\theta} \right)$ and $\mathbf{R}_{kk',N}^{**} \left(\boldsymbol{\theta} \right)$.

$$\begin{aligned} \frac{1}{N} r_N^{**} \left(\boldsymbol{\theta} \right) = & \frac{1}{N} \boldsymbol{\varepsilon}_N^{\mathrm{T}} \mathbf{R}_{0,N}^{**} \left(\boldsymbol{\theta} \right) \boldsymbol{\varepsilon}_N + \frac{1}{N} \sum_{k=1}^{K} \boldsymbol{\varepsilon}_N^{\mathrm{T}} \mathbf{R}_{k,N}^{**} \left(\boldsymbol{\theta} \right) \left(\mathbf{x}_{k,N}^{*} - \mathbb{E} \left[\mathbf{x}_{k,N}^{*} \right] \right) \\ & + \frac{1}{N} \sum_{k'=1}^{K} \left(\mathbf{x}_{k',N}^{*} - \mathbb{E} \left[\mathbf{x}_{k',N}^{*} \right] \right)^{\mathrm{T}} \mathbf{R}_{k,k',N}^{**} \left(\boldsymbol{\theta} \right) \left(\mathbf{x}_{k,N}^{*} - \mathbb{E} \left[\mathbf{x}_{k,N}^{*} \right] \right) \\ = & \, \sigma_0^2 \frac{1}{N} \operatorname{Tr} \left(\left(\mathbf{I}_N + \psi_0 \mathbf{E}_N \right)^{\mathrm{T}} \mathbf{R}_{0,N}^{**} \left(\boldsymbol{\theta} \right) \left(\mathbf{I}_N + \psi_0 \mathbf{E}_N \right) \right) \\ & + \frac{1}{N} \sum_{k=1}^{K} \sigma_0^2 \xi_{0,k} \operatorname{Tr} \left(\left(\mathbf{I}_N + \psi_0 \mathbf{E}_N \right)^{\mathrm{T}} \mathbf{R}_{k,N}^{**} \left(\boldsymbol{\theta} \right) \mathbf{C}_{k,N} \right) \\ & + \frac{1}{N} \sum_{k=1}^{K} \sum_{k'=1}^{K} \sigma_0^2 \xi_{0,k} \xi_{0,k'} \operatorname{Tr} \left(\mathbf{C}_{k,N}^{\mathrm{T}} \mathbf{R}_{k,k',N}^{**} \left(\boldsymbol{\theta} \right) \mathbf{C}_{k,N} \right) \\ & + \frac{1}{N} \sum_{k=1}^{K} \sum_{k'=1}^{K} \Xi_{kk'} \cdot \operatorname{Tr} \left(\mathbf{R}_{k,k',N}^{**} \left(\boldsymbol{\theta} \right) \right) + o_P \left(1 \right) \end{aligned}$$

Note that $N^{-1} \boldsymbol{\iota}^{\mathrm{T}} \left\{ \mathbf{x}_{k,N}^{*} - \mathbb{E} \left[\mathbf{x}_{k,N}^{*} \right] \right\} = o_P \left(1 \right)$ for $k = 1, \ldots, K$ uniformly in $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ by Lemmas A.3 and A.4 in Lee (2007a). Since $\boldsymbol{\Theta}$ is bounded and all the terms $l_N^{*} \left(\boldsymbol{\theta} \right)$,

$r_N^*(\boldsymbol{\theta})$, $l_N^{**}(\boldsymbol{\theta})$ and $r_N^{**}(\boldsymbol{\theta})$ can be expressed as appropriate functions of the relevant parameters, uniform convergence follows. Since $\mathbf{m}_N(\boldsymbol{\theta})$ is also quadratic in $\boldsymbol{\theta}$ and $\boldsymbol{\Theta}$ is bounded, then $\mathbb{E}[\mathbf{m}_N(\boldsymbol{\theta})]$ is uniformly equicontinuous in $\boldsymbol{\Theta}$. This result, along with the identification conditions, implies that the identification uniqueness condition for $\mathbb{E}\left[\mathbf{m}_N^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{A}_N^{\mathrm{T}}\mathbf{A}_N\mathbf{m}_N(\boldsymbol{\theta})\right]$ is satisfied. Thus, the consistency of the GMM estimator follows from standard arguments (White, 1996).

It remains to show that $\widehat{\boldsymbol{\theta}}_{GMM}$ is also asymptotically normal. The usual application of the Mean Value Theorem to the First Order Conditions of the GMM problem gives:

$$\sqrt{N}\left(\widehat{\boldsymbol{\theta}}_{GMM}-\boldsymbol{\theta}_0\right)=-\left[\mathbf{J}_N^{\mathrm{T}}\left(\widehat{\boldsymbol{\theta}}_{GMM}\right)\mathbf{W}_N\mathbf{J}_N\left(\overline{\boldsymbol{\theta}}\right)\right]^{-1}\mathbf{J}_N^{\mathrm{T}}\left(\widehat{\boldsymbol{\theta}}_{GMM}\right)\mathbf{W}_N\sqrt{N}\mathbf{m}_N(\boldsymbol{\theta}_0)$$

where $\mathbf{J}_N(\boldsymbol{\theta})=\frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{m}_N(\boldsymbol{\theta})$. By Theorem 1 in Kelejian and Prucha (2001):

$$\sqrt{N}\mathbf{A}_N\mathbf{m}_N(\boldsymbol{\theta}_0)\xrightarrow{d}\mathcal{N}\left(\mathbf{0},\mathbf{A}_0\boldsymbol{\Omega}_0\mathbf{A}_0^{\mathrm{T}}\right)\tag{A.2}$$

hence the main result would follow if $\mathbf{J}_N\left(\widehat{\boldsymbol{\theta}}_{GMM}\right)=\mathbf{J}_0+o_P(1)$. Note that:

$$\begin{aligned}\mathbf{J}_N(\boldsymbol{\theta})=&-\frac{1}{N}\begin{bmatrix}\mathbf{Q}_{1,N}\\ \vdots\\ \mathbf{Q}_{Q,N}\\ 2\boldsymbol{\varepsilon}_{\boldsymbol{N}}^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{P}_{1,N}\\ \vdots\\ 2\boldsymbol{\varepsilon}_{\boldsymbol{N}}^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{P}_{P,N}\end{bmatrix}\begin{bmatrix}\boldsymbol{\iota}_N & \mathbf{G}_N\mathbf{y}_N & \mathbf{X}_N & \mathbf{G}_N\mathbf{X}_N & \mathbf{0}_N & \mathbf{0}_N & \mathbf{0}_N\end{bmatrix}\\&+\frac{1}{N}\frac{\partial\boldsymbol{\lambda}_N(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^{\mathrm{T}}}\end{aligned}$$

where $\mathbf{0}_N$ is shorthand for an N-dimensional vector of zeros. Leaving $\frac{\partial}{\partial\boldsymbol{\theta}^{\mathrm{T}}}\boldsymbol{\lambda}_N(\boldsymbol{\theta})$ aside for the moment, we focus on a submatrix of the first term on the right-hand side, that is the last P rows of the second column. This vector comprises the derivatives of the P second-order moments with respect to β; the analysis of the rest of the matrix is just a simpler case. By Lemmas A.3 and A.4 in Lee (2007a), one can write every p-th element of said subvector, for $p=1,2,\dots,P$, as:

$$\frac{1}{N}\boldsymbol{\varepsilon}_N^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{P}_{p,N}\widetilde{\mathbf{G}}_N(\beta_0)\left(\alpha_0\boldsymbol{\iota}+\mathbf{X}_N\boldsymbol{\gamma}_0+\mathbf{G}_N\mathbf{X}_N\boldsymbol{\delta}_0+\boldsymbol{\varepsilon}_N\right)=b_{p,N}+v_{p,N}+t_{p,N}+f_{p,N}$$

where $\mathbf{y}_N(\boldsymbol{\theta}_0)\equiv(\mathbf{I}_N-\beta_0\mathbf{G}_N)^{-1}\left(\alpha_0\boldsymbol{\iota}_N+\mathbf{X}_N\boldsymbol{\gamma}_0+\mathbf{G}_N\mathbf{X}_N\boldsymbol{\delta}_0+\boldsymbol{\varepsilon}_N\right)$; the terms on the right-hand side are given by:

$$b_{p,N}=\frac{1}{N}\mathbf{d}_N^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{P}_{p,N}\mathbf{G}_N\mathbf{y}_N(\boldsymbol{\theta}_0)=\frac{1}{N}\mathbf{d}_N^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{P}_{p,N}\mathbf{G}_N\mathbb{E}[\mathbf{y}_N]+o_P(1)$$

and:

$$v_{p,N} = \frac{1}{N}\boldsymbol{\varepsilon}_N^{\mathrm{T}}\mathbf{P}_{p,N}\mathbf{G}_N\mathbf{y}_N(\boldsymbol{\theta}_0) = \sigma_0^2\frac{1}{N}\mathrm{Tr}\left[\mathbf{P}_{p,N}\widetilde{\mathbf{G}}_N(\beta_0)\right]$$
$$+\sum_{k=1}^{K}\frac{1}{N}\sigma_0^2\mathrm{Tr}\left[(\mathbf{I}_N+\psi_0\mathbf{E}_N)^{\mathrm{T}}\mathbf{P}_{p,N}\widetilde{\mathbf{G}}_N(\beta_0)(\gamma_{k,0}\mathbf{C}_{k,N}+\delta_{k,0}\mathbf{G}_N\mathbf{C}_{k,N})\,\xi_{k,0}\right]+o_P(1)$$

and:

$$\begin{aligned}t_{p,N} &= \frac{1}{N}(\beta_0-\beta)\,\boldsymbol{\varepsilon}_N^{\mathrm{T}}\widetilde{\mathbf{G}}_N(\beta_0)^{\mathrm{T}}\mathbf{P}_{p,N}\mathbf{G}_N\mathbf{y}_N(\boldsymbol{\theta}_0)\\ &= \sigma_0^2(\beta_0-\beta)\frac{1}{N}\Bigg\{\mathrm{Tr}\left(\widetilde{\mathbf{G}}_N^{\mathrm{T}}(\beta_0)\mathbf{P}_{p,N}\widetilde{\mathbf{G}}_N(\beta_0)\right)+\sum_{k=1}^{K}\mathrm{Tr}\Big[(\mathbf{I}_N+\psi_0\mathbf{E}_N)^{\mathrm{T}}\\ &\qquad\cdot\widetilde{\mathbf{G}}_N^{\mathrm{T}}(\beta_0)\mathbf{P}_{p,N}\widetilde{\mathbf{G}}_N(\beta_0)(\gamma_{k,0}\mathbf{C}_{k,N}+\delta_{k,0}\mathbf{G}_N\mathbf{C}_{k,N})\,\xi_{k,0}\Big]\Bigg\}+o_P(1)\end{aligned}$$

and:

$$\begin{aligned}f_{p,N} &= \frac{1}{N}\sum_{k=1}^{K}\left(\mathbf{x}_{k',N}^{*}-\mathbb{E}\left[\mathbf{x}_{k',N}^{*}\right]\right)^{\mathrm{T}}\boldsymbol{\Gamma}_{k,0}^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{P}_{p,N}\mathbf{G}_N\mathbf{y}_N(\boldsymbol{\theta}_0)\\ &= \frac{1}{N}\sum_{k=1}^{K}\xi_{k,0}\mathbf{C}_{k,N}^{\mathrm{T}}\boldsymbol{\Gamma}_{k,0}^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{P}_{p,N}\widetilde{\mathbf{G}}_N(\beta_0)\left(\mathbf{I}_N\sigma_0^2+\sum_{k'=1}^{K}(\gamma_{k',0}\mathbf{I}_N+\delta_{k',0}\mathbf{G}_N)\mathbf{C}_{k,N}\right)\\ &\qquad+\frac{1}{N}\sum_{k=1}^{K}\sum_{k'=1}^{K}\Xi_{kk'}\cdot\mathbf{P}_{p,N}\widetilde{\mathbf{G}}_N(\beta_0)(\gamma_{k',0}\mathbf{I}_N+\delta_{k',0}\mathbf{G}_N)+o_P(1)\end{aligned}$$

all the probability limits above imply uniform convergence for any $\boldsymbol{\theta}\in\boldsymbol{\Theta}$. Evaluating these terms at $\boldsymbol{\theta}=\boldsymbol{\theta}_0$, implies $d_{p,N}=v_{p,N}=t_{p,N}=f_{p,N}=o_P(1)$ as $\mathbf{d}_N(\boldsymbol{\theta}_0)=\mathbf{0}$. Collecting these results together gives:

$$\begin{aligned}\frac{1}{N}\boldsymbol{\varepsilon}^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{P}_{p,N}\mathbf{G}_N\mathbf{y}_N(\boldsymbol{\theta}_0) &= \sigma_0^2\mathrm{Tr}\left[\mathbf{P}_{p,N}\widetilde{\mathbf{G}}_N(\beta_0)\right]\\ &\qquad+\sigma_0^2\sum_{k=1}^{K}\xi_{k,0}\mathrm{Tr}\left[(\mathbf{I}_N+\psi_0\mathbf{E}_N)^{\mathrm{T}}\mathbf{P}_{p,N}\widetilde{\mathbf{G}}_N(\beta_0)\mathbf{G}_N\right]+o_P(1)\end{aligned}$$

furthermore, some tedious analysis reveals that $\frac{1}{N}\frac{\partial}{\partial\boldsymbol{\theta}^{\mathrm{T}}}\boldsymbol{\lambda}_N(\boldsymbol{\theta})=\frac{1}{N}\frac{\partial}{\partial\boldsymbol{\theta}^{\mathrm{T}}}\boldsymbol{\lambda}(\boldsymbol{\theta}_0)+o_P(1)$; thus, the $P\times 1$ submatrix of $\mathbf{J}_N(\boldsymbol{\theta})$ under examination has the desired properties. Finally, consistency of $\widehat{\boldsymbol{\theta}}_{GMM}$ also straightforwardly implies that $\mathbf{J}_N\left(\overline{\boldsymbol{\theta}}\right)\xrightarrow{p}\mathbf{J}_0$. These considerations, together with equation A.2, yield the desired result through the usual application of Slutsky's theorem.