ESSAYS IN SPATIAL PANEL DATA ECONOMETRICS

By

Steven Wu-Chaves

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Economics—Doctor of Philosophy

2024

ABSTRACT

Chapter 1: Robust inference in short linear panels with fixed effects with endogenous covariates in a spatial setting

In this chapter, I propose a simple way to obtain robust standard errors in linear panels in a spatial context with endogenous covariates where the number of time periods is small relative to the cross sectional dimension. The method is based on applying a Spatial HAC to an average of moment conditions across time to obtain a covariance estimator that is robust to both spatial and serial correlation (HACSC). I also present a control function approach (CF) alternative to estimate the parameters and extend the HACSC estimator to this case, where the standard errors require an adjustment to account for the sampling variability induced by the first stage estimation. In addition, I derive the Fixed Effects-Random Effects equivalence under a Correlated Random Effects framework in the presence of a spatial lag of the dependent variable to obtain a fully-robust Hausman-type test using the HACSC estimator. I run a Monte Carlo experiment and show that the HACSC estimator is robust to strong patterns of serial and spatial correlation. Furthermore, I also find that whenever the CF assumptions hold, the CF approach is more efficient than Two-Stage Least Squares. Finally, I estimate the effect of school district spending on the performance of fourth-grade students in Michigan, allowing for spillovers across districts. I find that the expenditure from neighboring districts has a positive and non-negligible impact on test passing rates.

Chapter 2: Estimation of models with spatial panels and missing observations in the covariates

Missing data problems are more serious en spatial models with spillover effects as the efficiency loss induced by using estimators that only use the complete cases is larger. In this paper I present a GMM estimator that uses the information on both the complete and incomplete observations for models with spatial spillover effects and missing data on the potentially endogenous variables to obtain potential efficiency gains. I also derive the Fixed Effects and Random Effects equivalence for spatial panels with missing data and I also develop an alternative GMM estimator in this Correlated Random Effects framework. The Monte-Carlo simulations show significant efficiency gains of the

GMM estimator compared to estimators that only use the complete cases.

Chapter 3: Estimation of models with multiple fixed effects and endogenous variables: a correlated random effects approach

The inclusion of multiple individual heterogeneities and time effects, more commonly referred as “fixed effects,” is a common practice in panel data. A common approach to deal with these is to estimate the model using the fixed effects estimator by applying the within transformation, which has the disadvantage of removing all the variables that are constant across one of the dimensions of the data set. An alternative method to estimate the model is the correlated random effects approach using the Mundlak device, which restricts the dependence between the heterogeneities and the covariates in a particular way. In this paper, I show that the fixed effects estimates can be recovered using the Mundlak approach in models with three sets of heterogeneities and in the presence of endogenous variables. Furthermore, I prove that this equivalence can be obtained using two different sets of covariates.

To my parents.

ACKNOWLEDGEMENTS

First, I am deeply thankful to my parents for their continued support. I am also grateful to Lucia, without whose encouragement and help I would not have been able to complete this journey. To the rest of my family and furry friends, thank you for their love and support that helped me in every step of this process.

Second I would like to extend a sincere thank you to my committee chair, Jeffrey Wooldridge, for all his advice, guidance and support. I would also want to thank Kyoo il Kim, Tim Vogelsang and Guo Chen for serving on my committee and providing excellent feedback and support. I also want to recognize Lisa Cook, Richard Baillie, Susan Zhu, Carl Davidson, Leslie Papke, Hugo Freeman, Soren Anderson, Scott Imberman, Todd Elder and Steven Haider for their helped and opportunities they provided me during the program. I am also grateful to Jay Feight and Lori Jean Nichols for their assistance as I navigated the program.

Finally, there are several class classmates that I also want to thank. First and foremost, to Minkyu Kim for his friendship, help and support over these years. To Salem Rogers, Raghav Rakesh, Chia-Hung Kuo, Benjamin Miller and Soo Jeong Lee, thank you for all the moments and conversations we shared together.

TABLE OF CONTENTS

CHAPTER 1

ROBUST INFERENCE IN SHORT LINEAR PANELS WITH FIXED EFFECTS WITH ENDOGENOUS COVARIATES IN A SPATIAL SETTING

1.1 Introduction

The assumption of independent data is widespread in empirical economics since it simplifies many of the estimation methods. However, in many fields such as international trade, urban economics, public policy or even network analysis, this assumption might not hold since the outcome variable of an individual might be affected by other observations' actions, which leads to (spatially) dependent data. Furthermore, many of the tools used to develop the asymptotic theory behind popular econometric methods such as the Central Limit Theorem and Law of Large Numbers often rely on independent and identically distributed (i.i.d.) data. This facilitates both the estimation and inference, but if this assumption is violated, then the latter becomes more difficult even if the parameters are estimated consistently.

Additionally, the increasing availability of data sets over time has increased the popularity of panel methods in recent years as they allow to incorporate time effects and to estimate richer models. Nevertheless, they also introduce complications because the presence of unobserved heterogeneity could generate inconsistency problems both in the parameters and standard errors if it is not properly handled. When combining both spatially dependent observations and panel data, inference becomes more challenging since the error term can be both serially and spatially correlated.

To address the spatial correlation, the literature in the field has usually resorted to assume and model a particular structure of the error term, as it was common to do with time series data. However, since the seminal work of White (1980), the common practice nowadays in the latter is to use standard errors that are robust to general forms of heteroskedasticity and autocorrelation (HAC). This procedure has been extended to the spatial framework (SHAC) by Conley (1999) and Kelejian and Prucha (2007) in a cross sectional setting. However, to the best of my knowledge and surprisingly enough, this has not been extended to the panel case where the time dimension is fixed and the number of units of observation goes to infinity, even in the linear case.[1] Admittedly, there are many cases in which the time dimension is also large, however there are also instances where the number of observations across time is considerably smaller compared to the cross sectional dimension.

This generates issues because ignoring the serial correlation could still generate biased standard errors, even if the associated covariance matrix is robust to spatial correlation. Indeed, some of the estimators that have been proposed in the literature and that have been implemented in software packages only make the standard errors robust to one of these dimensions. For example, Stata is a very popular statistical analysis package and one of the few routines for panel data in a spatial context corrects the standard errors for spatial correlation, but assumes serial independence of the error terms. The main purpose of this paper is to propose a simple way to obtain robust standard errors in a linear panel that are robust to heteroskedasticity and both to spatial and serial correlation (HACSC), without imposing any structure on the time dimension and using a Fixed Effects framework with endogenous covariates. I also extend this procedure to the case of the control function approach, where the computation of standard errors is more difficult due to the presence of a generated regressor.

HAC estimators have been extensively used in the time series literature since they avoid having to model the error term structurally, which can lead to inconsistency issues if that process is misspecified. Newey and West (1987) were the first to extend White's estimator to allow for general forms of heteroskedasticity and autocorrelation. In the panel case, Arellano (1987) introduced the panel clustered standard errors, which are robust to heteroskedasticity and autocorrelation but require that the observations between clusters to be uncorrelated.

In spatial panels, multiple authors have made important contributions to the field, extending many of the methods developed in the time series literature. For example, Driscoll and Kraay (1998) presented how to deal with spatially dependent panel data in a GMM context by averaging the moment conditions in the cross section dimension index, N. Their approach relies on holding

[1] Perhaps one of the reasons is that econometricians assume that it is obvious what to do, but many methods make strong assumptions in the time dimension like serial independence.

fixed N and letting time dimension $T \to \infty$. Vogelsang (2012) develops asymptotic theory for linear spatial panels with fixed effects in a fixed-b framework by averaging HAC estimators and by computing the HAC for averages as in Driscoll and Kraay (1998). In this case, the asymptotics rely again in $T \to \infty$ and allowing N to remain fixed or to grow. In a similar context Kim and Sun (2013) proposed a bivariate kernel HACSC estimator, which requires that both the cross section and time dimensions to go to infinity. Bester et al. (2011) suggested a cluster covariance matrix that is applicable when the data is dependent in the context of time series, spatial and panel data. More recently, Müller and Watson (2022a) introduced a new methodology to construct confidence intervals based on population principal components with the property that the resulting interval will have a coverage probability of 95% for a set of spatial patterns in a cross sectional setting. Müller and Watson (2022b) extended this framework to spatial panels to cover estimation techniques like difference-in-difference setups.

At the cross sectional level, Conley (1999) was the first to develop a Spatial HAC (SHAC) estimator in a GMM context. His approach is based on the assumption that the data generating process is spatially stationary. When working with dependent data and allowing $N \to \infty$, it is common to assume some sort of weak dependence mechanism, analogous to the time series literature, so that the influence of one observation on other units diminishes as the distance between them increases. In this case, Conley assumes that the data is spatially $\alpha-$mixing. Bester et al. (2016) provide a fixed-b analysis of Conley's SHAC estimator.

Kelejian and Prucha (2007) relax the spatial stationarity assumption and model the spatial dependence in terms of a weighting matrix, arguing that having a different number of neighbors, as it is common in empirical work, violates the assumption. In this respect, the notion of assigning weights to different units based on their distance to a particular point has been used in many fields. For example, in urban economics, McMillen (1996) used locally weighted regressions to estimate the value of land in Chicago, where each observation is given a specific weight based on its distance to the central business district. In the same spirit, in the geography literature the geographically weighted regression uses a very similar concept to model the idea that there might be spatial variability for models involving geo-referenced data (Wheeler & Tiefelsdorf, 2005). It is important to note that their Kelejian and Prucha's SHAC estimator is based on consistent estimates of the error terms, but they do not provide any parameter estimation framework.

Kim and Sun (2011) generalize this estimator to allow general linear and nonlinear models using moment conditions. Conley and Molinari (2007) performed a Monte Carlo study in which they compared the performance of multiple covariance estimators with dependent data in the context of locations measured with error and they concluded that non parametric estimators work better compared to parametric ones such as GMM and maximum likelihood estimators. In this paper, I follow Driskoll and Kraay's approach, but instead of averaging the moment conditions over the cross sectional dimension, I average the moment conditions over time and construct a GMM estimator and then apply Kelejian and Prucha's SHAC over the corresponding residuals. By doing this, I avoid imposing any assumptions over the serial correlation and hence, construct a covariance estimator that is robust to both serial and spatial correlation.

Beyond testing the statistical significance of the effect of a covariate on the response variable, robust inference is also important when trying to choose the correct specification of a model. More specifically, the correlated random effects (CRE) approach has been very popular in recent years because it is a simple way to test between Random Effects (RE) and Fixed Effects (FE) specifications and it allows to include time constant variables as noted by Joshi and Wooldridge (2019). Furthermore, we can obtain the FE coefficients of the time varying variables by including the time average of these on the right hand side of the equation in a Pooled OLS or RE regression, a result attributed to Mundlak (1978). Debarsy (2012) was the first to extend the Mundlak approach to the spatial setting. More recently, Li and Yang (2020) showed that when the model includes a structurally modeled error term (which involves maximum likelihood estimation), the equivalence holds conditional on the parameter associated with the error term, however, the equivalence breaks unconditionally, i.e., when this parameter has to be estimated jointly with the rest of parameters. In this paper, I show that the result holds in a specific setting; namely, if the model does not include a structurally modeled error term.

One of the additional advantages of not imposing a particular spatial structure on the error term is that some estimation methods become readily available such as Two Stage Least Squares (2SLS) or a Control Function (CF) approach (Blundell & Powell, 2003) whenever the researcher suspects an endogenous variable is in the model. In fact, adding a spatial lag of the response variable as a covariate yields the spatial autoregressive model (SAR), a very popular model in this literature. However, Kelejian and Prucha (1998) showed that this term induces an endogeneity problem, which is why the researcher has to resort to an Instrumental Variable (IV) procedure. In terms of the estimation of parameters, both 2SLS and the CF approach require the availability of instruments, however one important difference is that the latter imposes additional assumptions and is therefore less robust than 2SLS. On the other hand, if the assumptions hold, the CF allows to deal with the endogeneity in a more parsimonious way if multiple functions[2] of the endogenous variable appear on the right hand side of the equation and is probably more efficient (Wooldridge, 2010). Note that this parsimony is relevant in the spatial case since it is common to include spillover effects in the models and therefore, the likelihood of having multiple functions of a variable increases in this context.

In a spatial setup, Basile (2009) and Basile et al. (2014) extended the CF to additive nonparametric models. In terms of inference, Basile et al. (2014) recommends to use bootstrap to obtain confidence intervals, a practice that is common even in the i.i.d. case. However, as pointed out by Kunsch (1989), the independence assumption plays a critical role on the validity of the bootstrap, so besides the computational cost, in a spatial context this is not a trivial procedure due to the dependence between observations. Intuitively, if we just randomly sample the data in a time series setting at each bootstrap repetition, the serial correlation structure would be lost and a similar issue occurs in the spatial case. This is why different bootstrap methods have been proposed in the time series literature (see Politis and White (2004) for a brief overview), nevertheless their extension to the spatial case is not straightforward due to the absence of a natural ordering of the

[2] A well known result in the literature is that 2SLS and the CF give the same numerical coefficients if only one function of the endogenous variable is in the model. This carries over to the spatial case under the settings outlined at the beginning of the paragraph.

observations. Given this, it might be desirable to obtain a closed-form formula for the covariance matrix when the empirical researcher is working with parametric linear models with panel data in a spatial context. This paper tries to fill this hole in the literature by adjusting the HACSC estimator to the CF setting. This adjustment is necessary because in addition to deal with the spatial and serial correlation, it is necessary to take into account the sampling error induced by the first stage estimation.

The rest of the paper is organized a follows. Section 1.2 discusses the model and the assumptions used to obtain the estimator of the covariance matrix. Section 1.3 presents the HACSC estimator and its asymptotic properties. Section 1.4 derives the FE and RE equivalence using the correlated random effects approach in a spatial context. Section 1.5 presents an additional application of the HACSC estimator under a Feasible GLS context. Section 1.6 presents the control function approach and a discussion of the additional assumptions imposed in this context. Section 1.7 contains a set of Monte Carlo experiments and Section 1.8 shows an empirical application of the HACSC estimator using data from the Michigan education system. Section 1.9 concludes.

1.2 Model

1.2.1 Estimation of the parameters

Consider the following model[3]:

$$\begin{aligned} y_{it} &= x_{1it}\beta_1 + x_{2it}\beta_2 + W_i X_{1t}\gamma_1 + W_i X_{2t}\gamma_2 + \lambda W_i y_t + c_i + u_{it} \\ &= x_{it}\beta + W_i X_t\gamma + \lambda W_i y_t + c_i + u_{it}, \quad i = 1 \ldots N, \; t = 1 \ldots T \qquad (1.1) \end{aligned}$$

where y_{it} is the dependent variable, x_{1it} is a $1 \times (k_1 + 1)$, vector of explanatory exogenous variables (including an intercept), x_{2it} is a $(1 \times k_2)$ vector of endogenous variables. The sense in which x_{1it} is exogenous will be clarified below. W_i is the i-th row of the $N \times N$ time invariant weighting matrix W, whose diagonal elements are zero, X_{1t} and X_{2t} are the $N \times k_1$ and $N \times k_2$ matrices of exogenous and endogenous covariates, respectively, for all observations at time t, y_t is the vector

[3] The model includes a spatial lag of the dependent variable on the right hand side for the sake of generality and because this is a widely spread practice in the spatial literature. Nevertheless, it is important to emphasize that its inclusion precludes the interpretation of (1.1) as a conditional mean function and also complicates the interpretation of the coefficients. As such, in some sections of the paper this variable will be omitted.

of dependent variables at time t, c_i is the individual heterogeneity and u_{it} is the idiosyncratic error. Hence β, γ and λ are the parameters of interest and they are of dimension $(k+1) \times 1$, $k \times 1$ and 1×1 respectively. Throughout the rest of the paper, I assume that $N \longrightarrow \infty$ while T remains fixed.

We assume that there exist a set of instruments z_{2it} for x_{2it} of dimension $l \geq k_2$ (so that $W_i Z_{2t}$ are the instruments for $W_i X_{2t}$). As previously shown by Kelejian and Prucha (1998), the inclusion of a spatial lag of the dependent variable on the right hand side also induces an endogeneity issue for which we also need instruments. Kelejian et al. (2004) and Lee (2003) determined that the optimal set of instruments for this variable is a sequence of the form $W^j X_t$, for $j = 1...s$, $s \in \mathbb{N}$ (in this case, we would only include higher power spatial lags of X_{1t}). If we let $w^j_{rit} \equiv W^j_i X_{rt}$, $r = 1, 2$, and $\mathfrak{z}_{2it} \equiv W_i Z_{2t}$, $A_{it} \equiv (x_{1it} \quad x_{2it} \quad w_{1it} \quad w_{2it} \quad W_i y_t)$ and $\theta \equiv (\beta'_1 \quad \beta'_2 \quad \gamma'_1 \quad \gamma'_2 \quad \lambda)'$, then the model can be written more compactly as:

$$y_{it} = A_{it}\theta + c_i + u_{it} \tag{1.2}$$

Since we are not assuming a particular structure for the error term, we can estimate the parameters of (1.2) with the Fixed Effects 2SLS estimator. To do so, we can apply the within transformation to all the variables, so let $\ddot{y}_{it} = y_{it} - \bar{y}_i$, where $\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}$ and similarly for the independent variables and the instruments. Then we can use Pooled 2SLS to the transformed model

$$\ddot{y}_{it} = \ddot{A}_{it}\theta + \ddot{u}_{it} \tag{1.3}$$

using the instruments $\ddot{Z}_{it} = (\ddot{x}_{1it} \quad \ddot{w}_{1it} \quad \ddot{z}_{2it} \quad \ddot{\mathfrak{z}}_{2it} \quad \ddot{w}^2_{1it} \quad \ddot{w}^3_{1it} \ldots \ddot{w}^s_{1it})$. Note that all the individual unobserved effects have been removed. To obtain consistent parameters, we need the following orthogonality condition:

$$\mathbb{E}(\ddot{Z}'_{it}\ddot{u}_{it}) = \mathbb{E}[g_{it}(Z_{it}, \theta)] = 0, \quad t = 1 \ldots T \tag{1.4}$$

which is implied by the stronger strict exogeneity condition:

$$\mathbb{E}(u_{it}|Z) = \mathbb{E}(u_{it}|Z, W) = 0$$

where Z is the $NT \times [(s+1)k_1 + 2l + 1]$ matrix of exogenous variables for all cross sectional units and all time periods. We note that in this spatial setting, this condition is stronger than in the non-spatial case because here we are conditioning the expected value of u_{it} with respect to all other units and not only i's independent variables (see Wooldridge (2010), pp. 301 for more details).

The $g_{it}(Z_{it}, \theta)$ function is of dimension $(s+1)k_1 + 2l + 1 = r$, hence for each i, there are $T \times r$ moment conditions. Under this framework, we could use many more moment conditions because our strict exogeneity assumption implies orthogonality conditions for each pair of time periods and cross sectional units [i.e. $\mathbb{E}(\ddot{Z}_{it}\ddot{u}_{js})$, $i, j = 1 \ldots N$ and $t, s = 1 \ldots T$], however we will only use the conditions implied by the FE estimator. Using a similar idea as Driscoll and Kraay (1998), for each observation i we can average these moment conditions over time[4], so let:

$$g_i(Z_i, \theta) = \frac{1}{T} \sum_{t=1}^{T} g_{it}(Z_{it}, \theta) \tag{1.5}$$

From this, one can construct a GMM estimator, which will be defined as follows:

$$\hat{\theta} = \min_{\theta \in \Theta} \left[\frac{1}{N} \sum_{i=1}^{N} g_i(Z_i, \theta) \right]' \hat{\Omega} \left[\frac{1}{N} \sum_{i=1}^{N} g_i(Z_i, \theta) \right] \tag{1.6}$$

where $\hat{\Omega}$ is a $r \times r$ positive definite, symmetric, weighting matrix. Admittedly, as noted above we could estimate θ by running Pooled 2SLS on (1.3), however, the GMM framework allows for more generality. For instance, averaging the moment conditions over time for each observation can be done in other setups different than fixed effects. Furthermore, this averaging might not be the most efficient approach, but obtaining the optimal GMM in a two-step procedure might provide some efficiency gains with respect to Pooled 2SLS.

1.2.2 Assumptions

The consistency and normality of this estimator can be obtained from a Uniform Law of Large Numbers (ULLN) and Central Limit Theorem (CLT) derived by Nazgul and Prucha (2009) for non-stationary random fields in a possibly uneven lattice. Before stating their assumptions, we need some definitions. Let $D \subset \mathbb{R}^d$, $d \geq 1$ be an uneven lattice and let $\rho(i, j) = \max_{1 \leq k \leq d} |j_k - i_k|$ and

[4] Note however that Driscoll and Kraay's case is based on having N fixed ant $T \longrightarrow \infty$ and they average across i for all t.

$|i| = \max_{1 \leq k \leq d} |i_k|$, where i_k denotes the k-th component of i, be a metric and norm, respectively, of $\mathbb{R}^d$. The minimum distance between two subsets E, F of D is defined as $\rho(E, F) = \inf[\rho(i, j) : i \in E \text{ and } j \in F]$ and let $|E|$ denote the cardinality of a subset $E \in D$. Other definitions used throughout this section can be found in the Appendix.

We now state the assumptions required to obtain the consistency and asymptotic normality of $\hat{\theta}$. We note that the N subscript in the random fields and scalars of the assumptions are to explicitly indicate that the ULLN and CLT can accommodate for triangular arrays, which are common in the spatial literature and particularly in Cliff-Ord type models. However, for notation simplicity, it will be suppressed in many sections for the remainder of the paper.

Assumption 1

The lattice $D \subset \mathbb{R}^d, d \geq 1$ is infinite countable and there exists a distance ρ_0 such that $\rho(i, j) \geq \rho_0 \; \forall i, j \in D$. Without loss of generality, suppose that $\rho_0 > 1$.

Assumption 1 provides the necessary structure to the lattice. Note that the existence of the distance is essential in order to obtain non parametric estimators of the covariance matrix and it is analogous to the time difference between observations in the time series literature. Furthermore, it is possible that the distance *observed* by the researcher, between two observations i and j, $\rho^*(i, j)$, is measured with error. Note that the existence and availability of this distance measure is not trivial, even in the leading case of a geographical region. As shown in Figure 1.1, there are instances in which using the linear distance between many pairs of points in that territory would not represent the *real* burden to arrive from one location to another (e.g. driving), while there are other cases in which this measure would be appropriate (e.g. pollution).

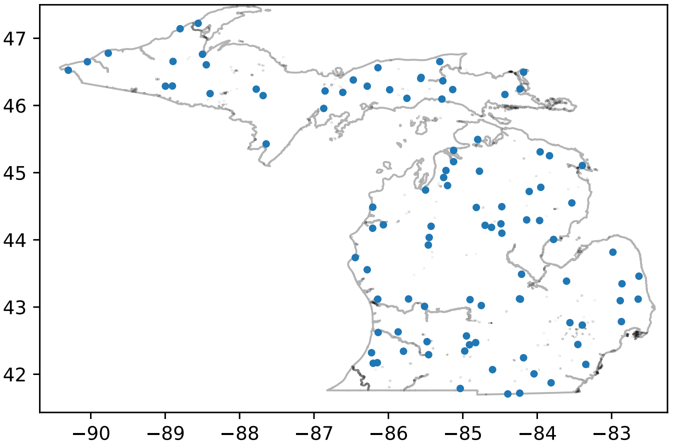


Figure 1.1 Points in an irregular geographic region.

Now we state conditions related to the the $g_i(\cdot)$ functions and $Z_{i,N}$, where $Z_{i,N}$ represents an α-mixing random field such that $i \in D$. At this point, it is important to note that since we are working with panel data and time averages for estimation purposes, the random field considered in the assumptions is the one constructed with the time averages for each observation.

Assumption 2 (Uniform L_2 integrability)

There is an array of positive real constants $\{c_{i,N}\}$ such that

$$\lim_{k\to\infty} \sup_{N} \sup_{i\in D_N} \mathbb{E}\left[|Z_{i,N}/c_{i,N}|^2 \mathbb{1}\left(|Z_{i,N}/c_{i,N}| > k\right)\right] = 0$$

Where $\mathbb{1}(\cdot)$ denotes an indicator function. Note that Assumption 2 allows for the possibility of asymptotic unbounded second moments, however for the remainder of the paper we will focus on the case of bounded moments, in which case we can set $c_{i,N} = 1 \;\; \forall i$. The next assumption put some restrictions on the α coefficients of the random field.

Assumption 3 (α-mixing)

Let $\bar{Q}^k_{i,N} \coloneqq Q_{|X_{i,N}/c_{i,N}|\mathbb{1}(|Z_{i,N}/c_{i,N}|>k)}$ denote the upper tail quantile function of $|Z_{i,N}/c_{i,N}|\mathbb{1}(|Z_{i,N}/c_{i,N}| > k)$ and recall that $\alpha_{\text{inv}}(u)$ is the inverse function of $\bar{\alpha}_{1,1}(m)$ as in the definition specified in the Appendix. The α-mixing coefficients satisfy:

1. $\lim\limits_{k\to\infty} \sup\limits_{N} \sup\limits_{i\in D_N} \int_0^1 \alpha^d_{\text{inv}}(u) \left[\bar{Q}^{(k)}_{i,N}(u)\right]^2 du = 0.$

2. $\sum\limits_{m=1}^{\infty} m^{d-1}\bar{\alpha}_{k,h}(m) < \infty$ for $k + h \leq 4$.

3. $\bar{\alpha}_{1,\infty}(m) = O_p(m^{-p-\varepsilon})$ for some $\varepsilon > 0$.

Under Assumptions 2 and 3.2 with $k = h = 1$ and letting $\{D_N\}$ be a sequence of finite subsets of D that satisfies Assumption 1 such that $|D_N| \to \infty$ as $N \to \infty$, a direct application of Theorem 3 in Nazgul and Prucha (2009) leads to the conclusion that

$$\frac{1}{|D_N|}\sum_{i\in D} Z_{i,N} - \mathbb{E}(Z_{i,N}) \xrightarrow{p} 0$$

Note that one could relax Assumption 2 to L_1 uniform integrability for the theorem to hold, nevertheless, the below CLT requires L_2 uniform integrability. In order to apply this pointwise WLLN to the $g_i(\cdot, \theta)$ functions, we assume that these satisfy the regularity conditions specified in Assumption A.1 presented in the Appendix. Given the fact that any measurable function of an α-mixing process is α-mixing, the $g_i(Z_{i,N}, \theta)$ also satisfy a pointwise WLLN, i.e.

$$\frac{1}{|D_N|}\sum_{i\in D} g_i(Z_{i,N}, \theta) - \mathbb{E}[g_i(Z_{i,N}, \theta)] \xrightarrow{p} 0 \tag{1.7}$$

With this Weak Law of Large Numbers, in order for the above GMM estimator to be consistent, we need an Uniform LLN for which we need the additional regularity conditions on the $g_i(\cdot, \cdot)$ functions stated in Assumption A.2. Under these assumptions, we have the following proposition, which is a special case of Theorem 2 in Nazgul and Prucha (2009).

Proposition 1. *Let $\{D_N\}$ be a sequence of finite subsets of D that satisfies Assumption 1 such that $|D_N| \to \infty$ as $N \to \infty$ and let $Q_N(\theta) = \frac{1}{|D_N|}\sum_{i\in D_N} g_i(Z_{i,N}, \theta)$. Suppose (Θ, ν) is a compact metric space and consider a sequence of real valued functions $\{g_i(Z_{i,N}, \theta) : i \in D_N, N \in \mathbb{N}\}$ satisfying Assumption 2 and that for all θ in Θ, these functions satisfy the WLLN in* (1.7)*. Then*

$$\sup_{\theta\in\Theta} |Q_N(\theta) - \mathbb{E}[Q_N(\theta)]| \xrightarrow{p} 0$$

With these tools at hand, define the following functions:

$$Q_N(\theta) \equiv \left[\frac{1}{N}\sum_{i=1}^{N} g_i(Z_i, \theta)\right]' \hat{\Omega} \left[\frac{1}{N}\sum_{i=1}^{N} g_i(Z_i, \theta)\right]$$

$$Q(\theta_0) \equiv \mathbb{E}[g_i(Z_{i,N}, \theta_0)]' \, \Omega_0 \, \mathbb{E}[g_i(Z_{i,N}, \theta_0)]$$

And suppose that $\hat{\Omega} \xrightarrow{p} \Omega_0$, where Ω_0 is a positive definite matrix. Recalling that $\mathbb{E}[g_i(Z_i, \theta)] = 0$ only when $\theta = \theta_0$, the true population value, the following proposition summarize the conditions under which the GMM estimator will be consistent:

Proposition 2. *Suppose that all the conditions of Proposition 1 hold. Additionally, assume that* (i) $g_i(Z_i, \cdot)$ *are continuous for all* $\theta \in \Theta$, (ii) $\hat{\Omega} \xrightarrow{p} \Omega_0$, *an* $r \times r$ *positive definite matrix and* (iii) θ_0 *is the only vector for which the moment condition in* (1.4) *holds. Then* $Q_N(\hat{\theta})$ *converges uniformly to* $Q(\theta_0)$ *and* $\hat{\theta} \xrightarrow{p} \theta_0$, *the unique minimizer of* $Q(\theta)$.

Note that since $\frac{1}{N}\sum_{i=1}^{N} g_i(Z_i, \theta)$ satisfies the ULLN of Proposition 1 and $\hat{\Omega} \xrightarrow{p} \Omega_0$, the proof of this proposition follows from Theorem 4.1.1 in Amemiya (1985). To obtain the asymptotic distribution of $\hat{\theta}$, we assume the following condition, which guaranties that the sum is not dominated by any term.

Assumption 4

If we define $\tilde{\sigma}_n^2 = \mathrm{Var}(S_n)$ and $S_n = \sum_{i \in D_N} Z_{i,N}$. Then the following condition is satisfied:

$$\liminf_{n \to \infty} |D_N|^{-1} \tilde{\sigma}_n^2 > 0$$

Under this assumption, Theorem 1 in Nazgul and Prucha (2009) ensures the asymptotic normality of the random variables Z_i.

Proposition 3. *Let* $\{D_N\}$ *be a sequence of finite subsets of D that satisfies Assumption 1 such that* $|D_N| \to \infty$ *as* $N \to \infty$ *and let* $\{Z_i : i \in D_N, n \in \mathbb{N}\}$ *be a sequence of zero mean real-valued random variables that satisfy Assumption 2. Furthermore, assume that the random field is* α*-mixing satisfying Assumption 5. Then,*

$$\tilde{\sigma}_n^{-1} S_n \xrightarrow{d} N(0, 1)$$

Once again, the previous proposition applies directly to the underlying random fields, however, we need a result to for the $g_i(Z_{i,N}, \theta)$ functions. Assuming that the latter satisfy the standard regularity conditions of Assumption A.3 , the first order conditions for the GMM estimator are

$$\left[\frac{1}{N}\sum_{i=1}^{N} \nabla_\theta g_i(Z_i, \hat{\theta})\right]' \hat{\Omega} \left[\frac{1}{N}\sum_{i=1}^{N} g_i(Z_i, \hat{\theta})\right] = 0 \tag{1.8}$$

Taking a mean value expansion of the last term around θ_0 yields the following expression:

$$g_i(\hat{\theta}) = g_i(\theta_0) + \nabla_\theta g_i(\tilde{\theta})(\hat{\theta} - \theta_0) + \text{ remainder} \tag{1.9}$$

for $\tilde{\theta}$ between $\hat{\theta}$ and θ_0 element-wise and where I suppressed the dependence of g_i on Z_i for notation simplicity. Replacing (1.9) in (1.8) yields:

$$\begin{aligned}\sqrt{N}(\hat{\theta} - \theta_0) = -&\left\{\left[\frac{1}{N}\sum_{i=1}^{N} \nabla_\theta g_i(\hat{\theta})\right]' \hat{\Omega} \left[\frac{1}{N}\sum_{i=1}^{N} \nabla_\theta g_i(, \tilde{\theta})\right]\right\}^{-1} \\ &\left[\frac{1}{N}\sum_{i=1}^{N} \nabla_\theta g_i(\hat{\theta})\right]' \hat{\Omega} \left[\frac{1}{\sqrt{N}}\sum_{i=1}^{N} g_i(\theta_0)\right] + \text{remainder}\end{aligned} \tag{1.10}$$

Noting again that the $\nabla_\theta g_i(\theta)$ preserve the mixing conditions, then

$$\frac{1}{N}\sum_{i=1}^{N} \nabla_\theta g_i(\hat{\theta}) \xrightarrow{p} \mathbb{E}\left[\nabla_\theta g_i(\theta_0)\right]$$

by the WLLN above. Since $g_i(\theta)$ is continuously differentiable, by Slutzky's Theorem, the first term of (1.10) converges in probability to

$$\left\{[\mathbb{E}(\nabla_\theta g_i(\theta_0)]' \, \Omega_0 \, [\mathbb{E}(\nabla_\theta g_i(\theta_0)]\right\}^{-1}$$

Furthermore, by the CLT,

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} g_i(\theta_0) = O_p(1)$$

Therefore, taking the probability limit of (1.10), we obtain

$$\begin{aligned}\sqrt{N}(\hat{\theta} - \theta_0) &= -\left\{\mathbb{E}\left[\nabla_\theta g_i(\theta_0)\right]' \Omega_0 \mathbb{E}\left[\nabla_\theta g_i(\theta_0)\right]\right\}^{-1} \\ &\qquad \mathbb{E}\left[\nabla_\theta g_i(\theta_0)\right]' \Omega_0 \left[\frac{1}{\sqrt{N}}\sum_{i=1}^{N} g_i(\theta_0)\right] + o_p(1) \\ &\xrightarrow{d} N(0, C'\Sigma C)\end{aligned} \tag{1.11}$$

where

$$C' = \left\{\mathbb{E}\left[\nabla_\theta g_i(\theta_0)\right]' \Omega_0 \mathbb{E}\left[\nabla_\theta g_i(\theta_0)\right]\right\}^{-1} \mathbb{E}\left[\nabla_\theta g_i(\theta_0)\right]' \Omega_0$$

and

$$\Sigma = \mathbb{E}[g_i(\theta_0) g_i(\theta_0)'] = \mathrm{Var}[g_i(\theta_0)] \tag{1.12}$$

Note that for the cases considered in this paper, C is just a matrix of data, so we do not need to estimate it. On the other hand, we need an estimator of the variance of the moment conditions, which we present in the next section. From an empirical implementation point of view, it is important to note that this GMM framework includes the simple estimators mentioned at the beginning of the section as special cases. For example, in the case of A_{it} containing only exogenous variables, then the GMM reduces to the same solution as estimating (1.3) with Pooled OLS. If A_{it} has some endogenous variables like in the model (1.1), and assuming that we have a set of instruments Z_{it}, then the Fixed Effects 2SLS can be obtained from the GMM estimator by setting $\hat{\Omega} = \ddot{Z}'\ddot{Z}$, where Z is the stacked $NT \times r$ matrix of instruments. Furthermore, we would need that the well known matrices of these estimators are of full column rank and to converge in probability to non-singular finite matrices.

Another empirical consideration is the specification of the weighting matrix W since in the model, the dependence of the outcome variable on other observations is generated by this matrix. In practice, there exist different ways to specify W. For example, one could assign weights as the inverse of the distance between two observations and set to zero the weights after a threshold or use a k-neighbors scheme. When dealing with geographic units, one could assign an equal weight for all the units j that share a border with unit i (rook type) or if they share an edge or a vertex (queen type) like in Figure XX, or even assign an equal weight to all other units in the sample (see LeSage and Pace (2009) for a discussion on weighting matrices).

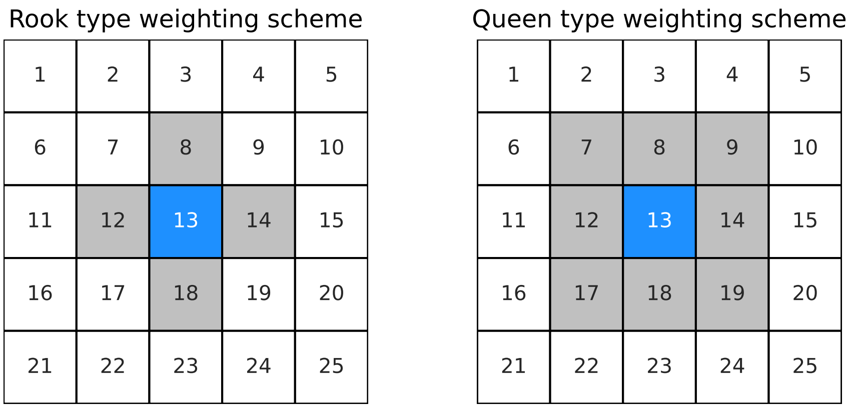


Figure 1.2 Rook and queen type weighting schemes. On the rook type scheme, if W is row normalized, only units 8, 12, 14 and 18 will receive a weight of $\frac{1}{4}$ in row 13. Analogously, if a queen type scheme is used, units 7, 8, 9, 12, 14, 17, 18 and 19 will have a weight of $\frac{1}{8}$ in row 13 of W.

Nonetheless, some of these specifications might violate the assumptions stated in this section. In particular, recall that we are working with an α-mixing random field, which implies that the dependence between the observations decays as they are farther apart. In this respect, it is clear that assigning an equal weight to all other observations violates this assumption. In a similar fashion, a k-neighbors pattern might not satisfy the α-mixing condition in cases where there are isolated units (e.g. a unit located alone in an island). Note that these restrictions to W also apply in cases where the distance measure is of economic nature or derived from a network perspective (e.g. degree of centrality).

1.3 The HACSC estimator

To obtain robust standard errors, recall that because for each observation we took the time average of their corresponding moment conditions, essentially we are working with a cross sectional problem. The idea is therefore to apply Kelejian and Prucha's (2007) estimator of the covariance matrix in this context, for which we need consistent estimates of the error terms. Analogous to the time series literature, their estimator requires a kernel function $K(\cdot)$, which will provide weights to the covariance terms entering the sums. In principle, only the covariance of observations that are close relative to some distance measure will receive a positive weight, while observations that are far away will receive a weight of zero. In other words, this function will operationalize the weak dependence assumption between observations to the error terms. Note however that this kernel will provide weights at the cross sectional dimension and not across time. To fix ideas, the researcher will need to choose a distance ρ_b such that $\rho_b \to \infty$ as $N \to \infty$ that will play the role of the truncation lag in a time series context. The next assumption imposes additional restrictions on the kernel function.

Assumption 5

The kernel $K : \mathbb{R} \to [-1, 1]$, satisfies the following conditions:

1. $K(0) = 1$
2. $K(x) = K(-x)$
3. $K(x) = 0$ for $x > 1$
4. $|K(x) - 1| \leq c_K |x|^{\alpha_K}$, $|x| \leq 1$ for some $\alpha_K \geq 1$ and $0 < c_K < \infty$.

As pointed out by Kelejian and Prucha (2007), Assumption 5 is satisfied by many kernels such as the rectangular kernel, Bartlett, the triangular kernel, among others. The next assumption imposes some structure for the error terms.

Assumption 6

The $N \times 1$ vector of errors is generated as follows:

$$u = R\varepsilon \tag{1.13}$$

where the ε is a $N \times 1$ vector of i.i.d. random variables with mean 0, variance of 1 and $\mathbb{E}[|\varepsilon|^q] < \infty$ for $q \geq 4$ and the R is a $N \times N$ non-singular unknown matrix whose row and column sums are uniformly bounded.

In light of Assumption 6, recall that although theoretically we are working with a cross sectional problem because we took the time average of the moment conditions, the underlying structure of the data is a panel. In this sense, (1.13) can also be seen as an average, so for each i, we have:

$$u_i = \begin{pmatrix} u_{i1} \\ u_{i2} \\ \vdots \\ u_{iT} \end{pmatrix} \tag{1.14}$$

where it t-th row of u_i is:

$$u_{i,t} = \sum_{s=1}^{t} R_{i,s}\varepsilon_s$$

and $R_{i,s}$ is the i-th row of R_s, a matrix with similar properties than R defined above, at time s. This implies that in each time period, the disturbances will depend on other unit's disturbances, past own values of disturbances, and past values of other unit's disturbances. In other words, this structure allows for *both* spatial correlation and serial correlation, "spatial serial" correlation and heteroskedasticity. Nevertheless, the uniform boundedness condition for R guaranties that the correlation between units is restricted at the cross sectional dimension, analogous to the time series case. Given the distance ρ_b, we can denote with v_i the number of pseudo-neighbors for i:

$$v_i = \sum_{j=1}^{N} \mathbb{1}[\rho^*(i,j) \leq \rho_b]$$

and let $v = \max_i v_i$. In words, v_i denotes the number of units j that are at a distance less than ρ_b from unit i. The following assumption is related to v.

Assumption 7

The random variable v satisfies the following conditions:

1. $\mathbb{E}(v^2) = o_p(N^{2\tau})$, where $\tau \leq \left(\frac{1}{2}\right)\frac{q-2}{q-1}$ and q is defined in Assumption 6.

2. $\sum_{j=1}^{N} |\sigma_{ij}|\rho(i,j)^{\alpha_S} \leq c_S$, for $\alpha_S \geq 1$ and $0 < c_S < \infty$ and σ_{ij} is the (i,j)-th element of Σ (defined below).

Assumption 7 plays a role in terms of limiting the degree of correlation between units, as well as ensuring that the estimator of the covariance matrix is consistent given the fact that we are using residuals instead of errors to estimate it. Assumptions 6 and 9 provide an identification condition and bound the measurement error of the distance, respectively.

Assumption 8

The matrix of exogenous variables, $\ddot{Z}$, has full column rank and its elements are uniformly bounded in absolute value by the finite constant $0 < c_Z < 0$. For a fixed and finite T, the matrices:

1. $\lim\limits_{N\to\infty} (NT)^{-1}\ddot{Z}'\ddot{Z} = Q_{ZZ}$.
2. $\lim\limits_{N\to\infty} (NT)^{-1}\ddot{Z}'RR'\ddot{Z} = Q_{ZRRZ}$.
3. $\operatorname*{plim}\limits_{N\to\infty}(NT)^{-1}\ddot{Z}'\ddot{Z} = Q_{ZZ}$.

are finite and non-singular. Furthermore, the matrix $\operatorname*{plim}\limits_{N\to\infty}(NT)^{-1}\ddot{Z}'\ddot{A} = Q_{ZA}$ has full column rank $2k$. Similarly, the diagonal elements of W are zero and all of its elements are uniformly bounded by a finite constant $0 < c_W < \infty$.

Assumption 9

The distance measure used by the empirical researcher $\rho^*(\cdot,\cdot)$ is potentially measured with error, i.e.

$$\rho^*(i,j) = \rho(i,j) + e_{ij} \geq 0$$

where $e_{ij} = e_{ji}$ denotes the measurement errors which are bounded in absolute value by the finite constant $0 < c_e < \infty$. Furthermore, $\{e_{ij}\}$ is independent of $\{\varepsilon_i\}$.

We need an additional assumption to account for the fact that we are using residuals instead of the actual error terms. This condition is provided in Assumption A.4 and should be satisfied by most $N^{\frac{1}{2}}$-consistent estimators. An extensive discussion of this and the previous assumptions is provided by Kelejian and Prucha (2007).

Note that given equations (1.4) and (1.5) and the matrix Σ specified in (1.12), we have the following:

$$\mathbb{E}[g_i(\theta_0)g_i(\theta_0)'] = \mathbb{E}\left[\ddot{Z}_i'\ddot{u}_i\ddot{u}_i'\ddot{Z}_i\right] \tag{1.15}$$

Because all the analysis is conditional on Z and W and by applying the Law of Iterated Expectations, from (1.15) and Assumption 6 we get that $\mathbb{E}(uu') = RR' = \Sigma$, where u is the $N \times 1$ vector of stacked error terms. In practical terms and recalling that $g_i(\cdot, \cdot)$ was defined as an average over time, we can estimate (1.15) by replacing the error terms by their residual counterparts and the expected value by an average applying the WLLN. Therefore, for the proposed estimator $\hat{\Sigma}$, its (r, s)-th element can be obtained as follows:

$$\hat{\Sigma}_{rs} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{t=1}^{T} \sum_{l=1}^{T} \ddot{Z}_{it,r} \ddot{Z}_{jl,s} \hat{\ddot{u}}_{it} \hat{\ddot{u}}_{jl} K \left[\frac{\rho^*(i,j)}{\rho_b} \right] \tag{1.16}$$

where $\ddot{Z}_{it,r}$ is the value of the covariate r for observation i at time t, while its population counterpart is given by the following expression:

$$\Sigma_{rs} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{t=1}^{T} \sum_{l=1}^{T} \ddot{Z}_{it,r} \ddot{Z}_{jl,l} \sigma_{it,jl} \tag{1.17}$$

The following proposition establishes the consistency of $\hat{\Sigma}$.

Proposition 4. *Consider the model in* (1.1) *and Assumptions 5-9 and 4. Suppose that the* (r, s)*-th elements of* Σ *and* $\hat{\Sigma}$ *are given by* (1.17) *and* (1.16) *respectively. Then* $\hat{\Sigma} \xrightarrow{p} \Sigma$.

Given the fact that we have assumed that T is fixed from the beginning, the proof of this proposition is virtually the same as in Kelejian and Prucha (2007). Note that we can re-write (1.16) as follows:

$$\begin{aligned} \hat{\Sigma}_{rs} = \frac{1}{NT} \Bigg\{ & \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{l=1}^{T} \ddot{Z}_{it,r} \ddot{Z}_{il,s} \hat{\ddot{u}}_{it} \hat{\ddot{u}}_{is} \cdot K\,[0] \\ & + \sum_{i=1}^{N} \sum_{i \neq j}^{N} \sum_{t=1}^{T} \sum_{l=1}^{T} \ddot{Z}_{it,r} \ddot{Z}_{il,s} \hat{\ddot{u}}_{it} \hat{\ddot{u}}_{js} K \left[\frac{\rho^*(i,j)}{\rho_b} \right] \Bigg\} \end{aligned} \tag{1.18}$$

The first term of (1.18) makes it clear that there are no restrictions imposed on the serial correlation for a particular observation, as the terms are not being down-weighted.

1.4 Correlated Random Effects

A direct application of the HACSC proposed in the previous section is related to the Correlated Random Effects (CRE) context. One of the most popular method applied in a panel setting is the fixed effects estimator since it allows the unobserved heterogeneity c_i to be arbitrarily correlated with the explanatory variables in the model. On the other side of the spectrum, the random effects approach imposes no correlation between c_i and the independent variables. A typical task that the empirical researcher must face is to choose between these two specifications, for which the literature has suggested multiple approaches. One of these is the CRE framework, which imposes restrictions on the distribution of the individual heterogeneity conditional on the regressors (Wooldridge, 2010).

One option is to follow Mundlak (1978) suggestion, which assumes that c_i can be modeled as a linear function of the averages of the time varying independent variables. More specifically, consider the following model:

$$y_{it} = x_{it}\beta + c_i + u_{it}, \quad i = 1 \dots N, \; t = 1 \dots T \tag{1.19}$$

Assuming that the x_i's are time varying, Mundlak considered the following specification:

$$c_i = \eta + \bar{x}_i\delta + e_i \tag{1.20}$$

where e_i is uncorrelated with $\bar{x}_i$ by assumption. Replacing (1.20) in (1.19) yields:

$$y_{it} = x_{it}\beta + \bar{x}_i\delta + e_i + u_{it}, \quad i = 1 \dots N, \; t = 1 \dots T \tag{1.21}$$

Mundlak (1978) showed that estimating β in (1.21) by pooled OLS (POLS) or random effects yields the same β than estimating (1.19) by fixed effects. In addition, we can perform a Hausman-type test using this equation by testing $\delta = 0$ to determine the suitability of one estimator versus the other one. It turns out that this FE-RE equivalence carries over the spatial setting under a particular setting, namely a model such as in equation (1.1), i.e. no autoregressive process of the error term u (Li and Yang (2020) showed that the equivalence breaks if we try to model structurally). Furthermore, this result carries over to the case of endogenous variables, which is a common issue in empirical work.

More concretely, consider the model in (1.1) and using the same notation, the Fixed Effects Two Stage Least Squares (FE2SLS) coefficients can be obtained by running Pooled 2SLS on the following equation:

$$\ddot{y}_{it} = \ddot{x}_{1it}\beta_1 + \ddot{x}_{2it}\beta_2 + \ddot{w}_{1it}\gamma_1 + \ddot{w}_{2it}\gamma_2 + \rho W_i \ddot{y}_t \tag{1.22}$$

using the instrumental variables $(\ddot{z}_{2it} \quad \ddot{\mathfrak{z}}_{2it} \quad \ddot{w}^2_{1it} \quad \ddot{w}^3_{1it} \ldots \ddot{w}^s_{1it})$, $s \in \mathbb{N}$. Then, it can be shown that running Pooled 2SLS on:

$$\begin{aligned}
y_{it} - \eta\bar{y}_i = {} & (x_{1it} - \eta\bar{x}_{1i})\beta_1 + (x_{2it} - \eta\bar{x}_{2i})\beta_2 + (w_{1it} - \eta\bar{w}_{1i})\gamma_1 + (w_{2it} - \eta\bar{w}_{2i})\gamma_2 \\
& + \rho W_i(y_t - \eta\bar{y}) + (1-\eta)\bar{x}_{1i}\delta_1 + (1-\eta)\bar{z}_{2i}\delta_2 + (1-\eta)\bar{w}_{1i}\lambda_1 \\
& + (1-\eta)\bar{\mathfrak{z}}_{2i}\lambda_2 + (1-\eta)\sum_{j=2}^{s}\bar{w}^j_{1i}\zeta_j \qquad (1.23)
\end{aligned}$$

using IV's:

$$\begin{aligned}
& [(z_{2it} - \eta\bar{z}_{2i}) \quad (\mathfrak{z}_{2it} - \eta\bar{\mathfrak{z}}_{2i}) \quad (w^2_{1it} - \eta\bar{w}^2_{1i}) \ldots \ (w^s_{1it} - \eta\bar{w}^s_{1i}) \\
& (1-\eta)\bar{\mathfrak{z}}_{2i} \quad (1-\eta)W_i\bar{Z}_2 \quad (1-\eta)\bar{w}^2_{1i} \ldots (1-\eta)w^s_{1i}]
\end{aligned}$$

yields the same $(\beta_1 \ \ \beta_2 \ \ \gamma_1 \ \ \gamma_2 \ \ \rho)$ as in (3.1) and where $\eta = 1 - \left[\sigma^2_u/(\sigma^2_u + T\sigma^2_c)\right]^{1/2}$ is assumed to be known. The following proposition summarizes this result.

Proposition 5. *Suppose* $\tilde{\Gamma} = (\tilde{\beta}_2 \ \ \tilde{\beta}_2 \ \ \tilde{\gamma}_1 \ \ \tilde{\gamma}_2 \ \ \tilde{\rho})$ *is the coefficient vector obtained by running Pooled 2SLS to equation* (1.23). *Then* $\tilde{\Gamma} = \hat{\Gamma}_{FE2SLS}$, *the coefficient vector obtained by running Pooled 2SLS to equation* (3.1).

The proof of this proposition can be found in the Appendix. Note that we have included the time averages of the instruments in (1.23), but this might introduce some distortions in the sense that the dimension of the z's might be larger than the original dimension of the x_2's. In practice, this will impact the degrees of freedom employed to perform the hypothesis testing to choose between FE and RE. Although when the cross sectional dimension is large this might not matter, in small samples this could have a significant impact in the statistical significance of the coefficients.

It is important to note that this FE-RE equivalence is an algebraic result, and as it turns out, one can obtain the FE coefficients of $(\beta \ \ \gamma \ \ \rho)$ in (1.23) by replacing the average of the instruments by the time averages of the predicted values of a regression of the endogenous variables on all of the exogenous variables, i.e.

$$\begin{aligned} y_{it} - \eta\bar{y}_i = {} & (x_{1it} - \eta\bar{x}_{1i})\beta_1 + (\hat{x}_{2it} - \eta\hat{\bar{x}}_{2i})\beta_2 + (w_{1it} - \eta\bar{w}_{1i})\gamma_1 + (\hat{w}_{2it} - \eta\hat{\bar{w}}_{2i})\gamma_2 \\ & + \rho W_i(\hat{y}_t - \eta\hat{\bar{y}}) + (1-\eta)\bar{x}_{1i}\delta_1 + (1-\eta)\hat{\bar{x}}_{2i}\delta_2 + (1-\eta)\bar{w}_{1i}\lambda_1 \\ & + (1-\eta)\hat{\bar{w}}_{2i}\lambda_2 + (1-\eta)W_i\hat{\bar{y}}\zeta_1 \qquad (1.24) \end{aligned}$$

This will "correct" the degrees of freedom issue mentioned above, at the expense of making the asymptotic theory harder since we have to take into account that we are using the predicted values instead of the original instrument averages. Proposition 6 summarizes this result and is proved in the Appendix.

Proposition 6. *Suppose* $\check{\Gamma} = (\check{\beta}_1 \quad \check{\beta}_2 \quad \check{\gamma}_1 \quad \check{\gamma}_2 \quad \check{\rho})$ *is the coefficient vector obtained by running Pooled OLS to equation* (1.24)*, where the* $\hat{}$ *represent the linear projections of the endogenous variables on the exogenous covariates. Then* $\check{\Gamma} = \hat{\Gamma}_{FE2SLS}$*, the coefficient vector obtained by running Pooled 2SLS to equation* (3.1)*.*

Once the researcher estimates the coefficients of (1.23) or (1.24), the next natural step is to test the hypothesis $\Xi = (\delta \quad \lambda \quad \zeta) = 0$ [here ζ denotes either $(\zeta_2 \ldots \zeta_s)$ in (1.23) or ζ_1 in (1.24)] to decide between FE and RE specifications. Even if model (1.1) does not have an explicit functional form for the error term, the u_{it} could still be serially or spatially correlated, therefore, we can use the HACSC estimator proposed in section 1.3 to conduct a fully robust Hausman-type test in a simple way. Specifically, one would need to get the Wald statistic as $\mathcal{W} = (\mathbf{R}\Xi)'(\mathbf{R}\hat{\Sigma}\mathbf{R}')^{-1}(\mathbf{R}\Xi)$, where $\mathbf{R}$ includes the set of restrictions on the coefficients, Ξ is the full set of coefficients estimated and $\hat{\Sigma}$ is the estimated HACSC robust covariance matrix.

1.5 Feasible GLS

As previously stated and analogous to the time series literature, it is common practice in empirical work to assume a particular structure of the error term in a spatial context. In particular, consider the following model:

$$\begin{aligned} y_t &= X_t\beta + v_t \\ v_t &= \rho W v_t + \varepsilon_t \\ \varepsilon_t &= c + u_t \end{aligned} \tag{1.25}$$

where y_t is a $N \times 1$ vector, X_t is a $N \times k$ matrix of covariates, c denotes the vector of individual heterogeneity and u_t is a vector of idiosyncratic errors at time t. In this model, X_t may contain spatial lags of the independent variables. In what follows, the conditioning on both X_t and W of all the analysis is implicit. By stacking the equations by time period, the model can be rewritten as follows:

$$\begin{aligned} y &= X\beta + v \\ v &= (\mathbf{I}_T \otimes \rho W)v + \varepsilon \\ \varepsilon &= (e_t \otimes \mathbf{I}_N)c + u \end{aligned} \tag{1.26}$$

where e_t represents a $T \times 1$ vector of ones. At this point, the researcher needs to make an assumption about the orthogonality condition between the independent variables and the composite error term and more specifically, the vector c. A typical choice is to assume that all the explanatory variables X are exogenous with respect to both vectors c and u, with each element of these being i.i.d. with zero mean and finite variances σ_c^2 and σ_u^2 respectively, and both vectors being independent from each other. Note that this working assumption is stronger than the one required to obtain the consistency of the fixed effects estimator described in previous sections (as in the rest of the paper, I assume that T is fixed and $N \to \infty$).

Given these assumptions, from (1.25) we can write $\mathbb{E}(v_t v_t')$ as follows:

$$\mathbb{E}(v_t v_t') = (\sigma_c^2 + \sigma_u^2)(\mathbf{I}_N - \rho W)^{-1}(\mathbf{I}_N - \rho W)^{-1} \tag{1.27}$$

Or using the stacked version of (1.26) instead, then we can write $\mathbb{E}(\varepsilon\varepsilon') = \Omega_\varepsilon$ in the following way:

$$\Omega_\varepsilon = \sigma_c^2(J_T \otimes \mathbf{I}_N) + \sigma_u^2\mathbf{I}_{NT} \tag{1.28}$$

where $J_T = e_t e_t'$. Therefore it follows that,

$$\mathbb{E}(vv') = \left[\mathbf{I}_T \otimes (\mathbf{I}_N - \rho W)^{-1}\right] \left[\sigma_c^2 (J_T \otimes \mathbf{I}_N) + \sigma_u^2 \mathbf{I}_{NT}\right] \left[\mathbf{I}_T \otimes (\mathbf{I}_N - \rho W)^{-1}\right] \tag{1.29}$$

Note that the middle of this matrix has a classic random effects structure. In order to compute this covariance matrix, it is assumed that the matrix $(\mathbf{I}_N - \rho W)$ is invertible and that $|\rho| < 1$ just as in the previous sections. Following the time series case and to facilitate the computation of the middle of (1.29), note that

$$\Omega_\varepsilon = \sigma_u^2 Q_0 + \sigma_1^2 Q_1 \tag{1.30}$$

where $Q_0 = \left(\mathbf{I}_T - \frac{J_T}{T}\right) \otimes \mathbf{I}_N$, $Q_1 = \frac{J_T}{T} \otimes \mathbf{I}_N$ and $\sigma_1^2 = \sigma_u^2 + T\sigma_c^2$. Noting that Q_0 and Q_1 are idempotent, symmetric, $Q_0 + Q_1 = \mathbf{I}_{NT}$ and that $Q_0 Q_1 = \mathbf{0}_{NT}$, it follows that $\Omega_\varepsilon^{-1} = \sigma_u^{-2} Q_0 + \sigma_1^{-2} Q_1$ and $\Omega_\varepsilon^{-\frac{1}{2}} = \sigma_u^{-1} Q_0 + \sigma_1^{-1} Q_1$. In short, if the researcher is willing to impose that the covariates are orthogonal to the individual heterogeneity vector c and the error term in (1.26) follows a spatial AR(1) process, then the matrix $\mathbb{E}(vv')$ will have a particular form that depends only on three parameters.

Knowing this, one can obtain an estimator that is potentially more efficient than the FE estimator. More specifically, the researcher can exploit the structure of the error term in (1.26) to remove the spatial correlation by performing a spatial Cochrane-Orcutt type transformation. Let

$$y^* = y - (\mathbf{I}_T \otimes \rho W) y$$

$$X^* = X - (\mathbf{I}_T \otimes \rho W) X$$

$$v^* = v - (\mathbf{I}_T \otimes \rho W) v$$

Therefore, the transformed model is

$$y^* = X^* \beta + v^* \tag{1.31}$$

Note that $v^* = \varepsilon$ so that (1.31) contains a classical composite error term. Given the structure of ε, we can perform a second transformation by multiplying (1.31) by $\Omega_\varepsilon^{-\frac{1}{2}}$ to obtain

$$\check{y} = \check{X} \beta + \check{\varepsilon} \tag{1.32}$$

where $\check{y} = \Omega_\varepsilon^{-\frac{1}{2}} y^*$ and similarly for the rest of the terms. Note that

$$\begin{aligned}
\mathbb{E}\left(\check{\varepsilon}\check{\varepsilon}'\right) &= \Omega_\varepsilon^{-\frac{1}{2}}\mathbb{E}(\varepsilon\varepsilon')\Omega_\varepsilon^{-\frac{1}{2}} \\
&= (\sigma_u^{-1}Q_0 + \sigma_1^{-1}Q_1)(\sigma_u^2 Q_0 + \sigma_1^2 Q_1)(\sigma_u^{-1}Q_0 + \sigma_1^{-1}Q_1) \\
&= Q_0 + Q_1 \\
&= \mathbf{I}_{NT} \qquad (1.33)
\end{aligned}$$

Hence (1.32) can be estimated by Pooled OLS to obtain a GLS-type estimator to obtain efficiency gains, denoted by β_{GLS}. If all the relevant matrices are well behaved as $N \longrightarrow \infty$ and non-singular, Kapoor et al. (2007) showed that

$$(NT)^{\frac{1}{2}}\left(\hat{\beta}_{GLS} - \beta\right) \xrightarrow{d} N(0, \Psi) \text{ as } N \longrightarrow \infty \qquad (1.34)$$

where $\Psi = \left(\sigma_u^2 M_{XX}^0 + \sigma_1^2 M_{XX}^1\right)^{-1}$ and $M_{XX}^j = \lim_{N\to\infty} \frac{1}{NT} X^{*\prime} Q_j X^*$ for $j = 0, 1$. The previous analysis requires knowledge of σ_c^2, σ_u^2 and ρ and therefore it is not feasible. Kapoor et al. (2007) proposed generalized moments estimators of these parameters and they showed that if $\hat{\beta}_{FGLS}$ is the Pooled OLS estimator of (1.32) using *any* consistent estimators $\hat{\sigma}_c^2, \hat{\sigma}_u^2$ and $\hat{\rho}$ instead of σ_c^2, σ_u^2 and ρ, then

$$(NT)^{\frac{1}{2}}\left(\hat{\beta}_{GLS} - \hat{\beta}_{FGLS}\right) \xrightarrow{p} 0 \text{ and } \hat{\Psi} - \Psi \xrightarrow{p} 0 \qquad (1.35)$$

where $\hat{\Psi} = \left(\frac{1}{NT}\hat{X}^{*\prime}\hat{\Omega}_\varepsilon^{-1}\hat{X}^*\right)^{-1}$, provided that the working assumptions used to derive (1.34) hold. Note that the hats over the components of $\hat{\Psi}$ denote the dependence of the terms on $\hat{\sigma}_c^2, \hat{\sigma}_u^2$ and $\hat{\rho}$.

The validity of the previous covariance matrix Ψ rests on the working assumptions that the error term v follows a spatial AR(1) and the conditions imposed on each element of c and u hold. However, from an empirical perspective it is always possible that the structure of Ω_ε does not have the RE form due to the presence of heteroskedasticity or serial correlation on u_i for example. It is important to stress out that even if Ω_ε does not have the same structure as in (1.28), $\hat{\beta}_{FGLS}$ remains consistent, provided that the strict exogeneity condition (more formally this would mean that $\mathbb{E}[X \otimes c] = 0$ and $\mathbb{E}[X \otimes u] = 0$) and the corresponding rank condition continue to hold.

Nevertheless, if the researcher is unsure about the assumptions related to the vectors of individual heterogeneity c or the idiosyncratic errors u made in this section, it is wise to make robust inference.

In these instances, the HACSC estimator presented in this paper can be useful to achieve this purpose. More specifically, consider the residuals

$$\check{\ddot{\varepsilon}}_t = \check{y}_t - \check{X}_t\hat{\beta}_{FGLS}, \quad t = 1\ldots T.$$

where $\hat{\beta}_{FGLS}$ is obtained by estimating (1.32). In this context, the (r, s)-th element of the middle of the robust covariance matrix is

$$\hat{\Sigma}_{rs} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{l=1}^{T}\check{X}_{it,r}\check{X}_{jl,s}\check{\ddot{\varepsilon}}_{it}\check{\ddot{\varepsilon}}_{jl}K\left[\frac{\rho^*(i,j)}{\rho_b}\right] \tag{1.36}$$

And the fully robust covariance matrix is:

$$\check{\Psi} = \left(\check{X}'\check{X}\right)^{-1}\left\{\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{l=1}^{T}\check{\ddot{\varepsilon}}_{it}\check{\ddot{\varepsilon}}_{jl}\check{X}'_{it}\check{X}_{jl}K\left[\frac{\rho^*(i,j)}{\rho_b}\right]\right\}\left(\check{X}'\check{X}\right)^{-1} \tag{1.37}$$

where $\check{X}_{it}$ is the $1\times k$ vector of covariates at time t for observation i. Note that the computation of $\check{\Psi}$ requires the use of the transformed variables and not the original ones, which is consistent with the *estimating* equation (1.32). As in the previous sections, the kernel function $K(\cdot)$ will provide weights so that the (possible) spatial correlation decreases for observations that are far apart according to the distance measure $\rho(\cdot,\cdot)$. Naturally, $\check{\Psi}$ will be valid whether the RE structure of Ω_ε holds or not and will be robust to arbitrary serial and spatial correlation, as well as heteroskedasticity.

Throughout this section we have assumed that all the elements of the explanatory variables are uncorrelated with the error term u. If some elements of X are endogenous (i.e. $\mathbb{E}[x'_{it}u_{it}] \neq 0$) and the researcher has available instruments Z, then the extension to an IV procedure is straightforward as discussed in Mutl and Pfaffermayr (2010) and B. Baltagi and Liu (2011). The estimation approach would be to apply Pooled 2SLS to the estimating equation 1.32 using instruments $\check{Z}$, where the $\check{}$ denotes the same transformations made earlier in the section. In this instance, the computation of the covariance matrix using the HACSC estimator would look like (1.16), but the researcher would need to use the transformed variables as in this section instead.

1.6 Alternative estimation: a Control Function approach

It is well known that Instrumental Variables estimation procedures such as 2SLS deliver consistent estimates of the parameters at the expense of losing precision when compared to OLS as pointed out by Cameron and Trivedi (2005). In such instances, if the researcher is willing to impose additional assumptions, she can resort to the control function approach (Blundell & Powell, 2003), which can deliver estimates that are (potentially) more efficient as it will be shown in simulations. To this end, consider the following estimating equation:

$$y_{it} = f(X_{1t}, X_{2t}, W) + c_i + u_{it} \tag{1.38}$$

where $f(\cdot)$ is a known function, $\mathbb{E}(X'_{1t}u_{it}) = 0$ and $\mathbb{E}(X'_{2t}u_{it}) \neq 0$. In practice, $f(\cdot)$ will almost certainly contain linear functions of X_{1t}, X_{2t} as well as spatial spillovers from these variables, but it can also include nonlinear terms of the endogenous variables such as interactions with X_{1t}, squared functions and so on. Now, to analyze the CF approach, consider equation (1.39), which is a special case of (1.38) and is very similar to (1.1) but without the spatial lag of the dependent variable on the right hand side[5], which will allow us to interpret it as a conditional mean function and for simplicity we will assume that there's only one element in x_{2it}:

$$\begin{aligned} y_{it} &= x_{1it}\beta_1 + x_{2it}\beta_2 + W_iX_{1t}\gamma_1 + W_iX_{2t}\gamma_2 + c_i + u_{it} \\ &= x_{it}\beta + W_iX_t\gamma + c_i + u_{it}, \quad i = 1\dots N, \ t = 1\dots T \end{aligned} \tag{1.39}$$

where the definitions are the same as in Section 1. By applying the within transformation, we obtain the estimating equation:

$$\ddot{y}_{it} = \ddot{x}_{it}\beta + W_i\ddot{X}_t\gamma + \ddot{u}_{it} \tag{1.40}$$

As with the 2SLS case and using obvious notation, this approach also requires the availability of a set of instruments $\ddot{Z}_{it} = (\ddot{x}_{1it} \ \ \ddot{w}_{1it} \ \ \ddot{z}_{2it} \ \ \ddot{\mathfrak{z}}_{2it})$. The first two assumptions of the Control Function (CF) approach are the same as with 2SLS, namely: $\mathbb{E}(\ddot{Z}'_{it}\ddot{u}_{jt}) = 0$ for $i, j = 1\dots N$ and $t = 1\dots T$ and the identification condition $\text{rank}[\mathbb{E}(\ddot{Z}'\ddot{A})] = 2k - 1$. The first stage of the estimation involves the reduced form of the endogenous variable on the instruments and obtaining the disturbances $\ddot{v}_{2it}$, i.e.

$$\ddot{v}_{2it} = \ddot{x}_{2it} - \ddot{Z}_{it}\psi \tag{1.41}$$

[5] It is certainly possible to use the control function approach with the spatial lag of the dependent variable as a covariate.

where $\mathbb{E}(\ddot{Z}_{it}'\ddot{v}_{it}) = 0$. Given that $\mathbb{E}(\ddot{Z}_{it}'\ddot{u}_{it}) = 0$, note that $\ddot{x}_{2it}$ and $\ddot{w}_{2it}$ are endogenous if and only if $\ddot{u}_{it}$ is correlated with $\ddot{v}_{2it}$ and $W_i\ddot{v}_{2t}$. At this point we state the additional assumption required by the CF approach:

$$\mathbb{E}(\ddot{u}_{it}|Z, X_2, W) = \mathbb{E}(\ddot{u}_{it}|Z, \ddot{v}_2, W) = \mathbb{E}(\ddot{u}_{it}|\ddot{v}_2, W) = \mu_1\ddot{v}_{2it} + \mu_2 W_i\ddot{v}_{2t} \tag{1.42}$$

This equation has two strong implicit restrictions. First, the second equality would hold under independence of Z and $(\ddot{u}, \ddot{v}_2, W)$ and second, we are assuming a linear conditional expectation of $\ddot{u}_{it}$ on the parameters. Given this, we can write

$$\ddot{u}_{it} = \mu_1\ddot{v}_{2it} + \mu_2 W_i\ddot{v}_{2t} + \ddot{e}_{it} \tag{1.43}$$

Replacing (1.43) in (1.40) yields:

$$\ddot{y}_{it} = \ddot{x}_{it}\beta + W_i\ddot{X}_t\gamma + \mu_1\ddot{v}_{2it} + \mu_2 W_i\ddot{v}_{2t} + \ddot{e}_{it} \tag{1.44}$$

Stacking again all the explanatory variables into a matrix A and the coefficients into a vector θ yields:

$$\ddot{y}_{it} = \ddot{a}_{it}\theta + \ddot{e}_{it} \tag{1.45}$$

The error term in (1.45) is uncorrelated with the rest of variables in the equation (including $\ddot{x}_{2it}$ and $\ddot{w}_{2it}$), so the parameters can be consistently estimated using Pooled OLS by replacing the disturbances with the computed residuals from the first stage. Therefore, the estimating equation for the main model becomes:

$$\ddot{y}_{it} = \hat{\ddot{a}}_{it}\theta + \ddot{e}_{it} \tag{1.46}$$

where the ^ denotes that we are using generated regressors. Two important observations from equation (1.44) is that by including both $\ddot{v}_{2it}$ and $W_i\ddot{v}_{2t}$, the parameters obtained from this estimation will be numerically the same as 2SLS.[6] Second, *if* $\mu_2 = 0$, then it would be enough to include only $\ddot{v}_{2it}$ in the estimating equation to get consistent estimates of θ and in this scenario, they would be different than 2SLS. Furthermore, it is precisely by excluding $W_i\ddot{v}_{2t}$ from the estimation that the

[6] In this sense, we do not get any efficiency gains compared to 2SLS by including both terms.

CF would probably be more efficient than 2SLS in this case, as it would be using information from this restriction.

The CF has some additional advantages over 2SLS. One, the inclusion of the generated regressors in (1.44) allows the researcher to perform a Hausman-type test to determine if the suspected variables are endogenous that can be made robust to heteroskedasticity, spatial and serial correlation using the estimator proposed below. Second, the CF can handle nonlinear functions of the endogenous variables in a parsimonious way: for example in model (1.39), x_{2it} could contain interactions with other exogenous variables or even squared terms, in which case the CF only requires to include only $\hat{\ddot{v}}_{2it}$ in the final estimating equation, whereas 2SLS would need a reduced form equation for each additional function of the endogenous variable. If such nonlinear functions of the endogenous variable are indeed present in the main model, the CF can be made more flexible by including terms such as $\hat{\ddot{v}}_{2it}^2$ (but again, this is not necessary as the inclusion of $\hat{\ddot{v}}_{2it}$ already "controls" for this endogeneity), at the cost of having to adapt the standard errors to account for these new generated regressors.

From this point onward, one has to decide how to deal with the error term. One option is to impose some structure to it and apply a Feasible GLS procedure in order to obtain further efficiency gains. Note that this is possible because in (1.42) we have conditioned on the whole set of exogenous variables and the weighting matrix. However, this would not be possible if we slightly modify the model. So far we have assumed that the model also contains spatial spillovers of the endogenous variable $\ddot{x}_{2it}$, but suppose that for some theoretical reason, the model does not include $W_i\ddot{X}_{2t}$. In this case we could relax (1.42) to

$$\mathbb{E}(\ddot{u}_{it}|Z_{it}, x_{2it}) = \mathbb{E}(\ddot{u}_{it}|Z_{it}, \ddot{v}_{2it}) = \mathbb{E}(\ddot{u}_{it}|\ddot{v}_{2it}) = \mu_1\ddot{v}_{2it} \tag{1.47}$$

Note that we are now conditioning only on the own control function. In this instance one could still estimate the transformed model by Pooled OLS, however it would preclude to apply a Feasible GLS procedure because the strict spatial exogeneity assumption would be violated since it will involve the weighting matrix W and the error terms of other observations.

Alternatively, the researcher can treat the error term non-parametrically and apply the HACSC estimator proposed in this paper to obtain robust standard errors. Nevertheless, in this case there's an additional layer of complication on top of the spatio-temporal correlation and the heteroskedasticity: by including $\hat{\ddot{v}}_{2it}$ in the estimating equation, we now have a generated regressor and therefore, the covariance matrix of the parameters needs to be adjusted to take into account the sampling error induced by the first stage estimation (i.e. we are getting *estimates* of ψ). Although Basile et al. (2014) recommends to perform a bootstrap to obtain the standard errors in a CF setup, sampling with spatially dependent data is not a trivial matter so having a formula is useful in practice.

In this setup, the fully robust covariance matrix is

$$B^{-1}MB^{-1} \tag{1.48}$$

where

$B = \mathbb{E}\left(\sum_i^N \sum_t^T \ddot{a}'_{it}\ddot{a}_{it}\right)$.

$M = \text{Var}\left[\sum_i^N \sum_t^T (\ddot{z}_{it}\psi)'(\ddot{e}_{it} + \ddot{v}_{it}\theta) - G \cdot r_{it}(\psi)\theta\right] = \text{Var}\left[\sum_i^N \sum_t^T m_{it}\right]$.

$G = \mathbb{E}\left[\sum_i^N \sum_t^T (\ddot{z}_{it}\psi)'\ddot{z}_{it}\right]$

$r_{it}(\delta) = \left(\frac{1}{NT}\sum_i^N \sum_t^T \ddot{z}'_{it}\ddot{z}_{it}\right)^{-1}\left[(NT)^{-\frac{1}{2}}\sum_i^N \sum_t^T \ddot{z}'_{it}\ddot{v}_{it}\right]$.

The derivation of (1.48) can be found in the Appendix. To estimate it, we can replace the population quantities by their sample analogues so that

$\hat{B} = \frac{1}{NT}\sum_i^N \sum_t^T \hat{\ddot{a}}'_{it}\hat{\ddot{a}}_{it}$.

$\hat{m}_{it} = (\ddot{z}_{it}\hat{\psi})'(\hat{\ddot{e}}_{it} + \hat{\ddot{v}}_{it}\hat{\theta}) - \hat{G} \cdot \hat{r}_{it}(\hat{\psi})\hat{\theta}$.

With these quantities calculated, the (r, s)-th element of M can be estimated as

$$\hat{M}_{rs} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{l=1}^{T}\hat{\ddot{m}}_{it,r}\hat{\ddot{m}}_{jl,s}K\left[\frac{\rho^*(i,j)}{\rho_b}\right]$$

Note that (1.48) also has a sandwich type form, very similar to the HACSC estimator presented earlier. Similarly, the kernel function is also used to operationalize the weak spatial dependence assumption, however in this case the terms it multiplies (m_{it} instead of $\ddot{z}'_{it}\ddot{u}_{it}$) have a different structure to take into account the first stage sampling error.

1.7 Simulations

1.7.1 Design

To test the performance of the HACSC estimator and the CF version of it, I performed a Monte Carlo study. In this experiment, the units of observation live in a squared regular grid of 20×20 and the distance between two adjacent individuals is normalized to one. To evaluate the performance of the estimator, consider the following data generating process:

$$
\begin{aligned}
y_{it} &= \beta_0 + x_{1it}\beta_1 + x_{2it}\beta_2 + x_{1it}x_{2it}\beta_3 + c_i + u_{it} \\
x_{1it} &= \delta_0 + \delta_1 z_{1it} + v_{it} \\
c_i &= (I - \rho W)_i^{-1} C \\
u_{it} &= \alpha v_{it} + e_{it} \\
e_t &= (I - \rho W)^{-1} a_t \\
a_{it} &= \psi a_{i,t-1} + \varepsilon_{it} \\
&\mathbb{E}(x_{1it}u_{it}) \neq 0, \mathbb{E}(x_{2it}c_i) \neq 0, \mathbb{E}(z_{1it}c_i) \neq 0
\end{aligned}
$$

where $[\beta_0 \quad \beta_1 \quad \beta_2 \quad \beta_3]' = [2 \quad 0.7 \quad 0.6 \quad 0.3]'$ and ε_{it}, and C are independent and identically distributed random variables following normal distributions and are independent from each other. z_{1it} is an instrument for x_{1it} and x_{2it} is exogenous with respect to the error term u_{it} and they follow a normal and gamma distributions respectively. Note that there is an interaction term between the endogenous and exogenous variable, for which we have a readily available instrument, $z_{1it}x_{1it}$.

In this setup, the error term u_{it} satisfy the CF assumption given that it depends linearly on the error term from the reduced-form equation, v_{it}. The error terms e and a follow a spatial and temporal AR(1) process respectively. The strength of the spatial correlation is governed by the parameter ρ, while the persistence of the serial correlation is moderated by ψ. Note also that the individual heterogeneity also follows a Spatial AR(1) model, however, since I am going to apply the within transformation for the estimation, its DGP does will not affect the results.

For the weighting matrix W, I used a rook-type weighting scheme so that each observation will have between two and four pseudo-neighbors and each of those will have an equal weight. W is row-normalized to ensure that $(I-\rho W)$ is invertible. I estimated the model using both FE 2SLS and the CF approach with $N = 400$ and $T = 5$ using 1,000 replications. I am interested in comparing the estimates of the coefficients by the two methods to see if there are some efficiency gains by using the CF approach. Furthermore, I also want to evaluate the performance of four different estimators of the covariance matrix: the HACSC proposed in this paper, a SHAC assuming no serial correlation, the cluster robust and the "regular" ones without any adjustment. In the case of the CF approach, I will compare the standard errors presented in Section 1.6 that account for the first stage and a HACSC that ignores the two-step procedure.

I conducted a simulation for every combination of $\rho = [0,\ 0.3,\ 0.7]$ and $\psi = [0,\ 0.3,\ 0.7]$. I used the Bartlett Kernel to perform the analysis, contrary to Kelejian and Prucha (2007), who used the Parzen Kernel. An important parameter in this experiment is the threshold distance ρ_b at which the Kernel will assign a zero weight for units that are apart by more than ρ_p. Following the recommendation of the authors mentioned above, I set $\rho_b = N^{\frac{1}{4}}$, i.e. the integer part of $N^{\frac{1}{4}}$. At each iteration, I draw a new set of covariates and keep it fixed across the iterations of the ρ and ψ parameters.

1.7.2 Results

This section describes the results of the simulations using two metrics for the estimated coefficients: the mean and the corresponding standard deviation across the 1,000 replications for different values of ρ and ψ. Table 1.1 presents the outcomes of this experiment and it shows that both estimators provided unbiased estimates of the parameters in the sense that the average of the estimated coefficients is centered around the true values for any combination of ρ and ψ. This is expected since in this exercise the CF assumption is true.

However, when analyzing the standard deviations, the CF consistently shows a lower value than 2SLS (e.g. 0.049 against 0.084 for β_3 when $\rho = \psi = 0.3$). Figure 1.3 exemplifies this finding: note that the distribution of the estimated parameters is tighter around the true value for the CF estimates compared to 2SLS'. Therefore, whenever the CF assumption holds, this estimator seems to be more efficient, which can be explained by the fact that we are using additional information when performing the estimation. Interestingly, these efficiency gains are more evident for β_1 and β_3, the coefficients associated with the endogenous variables, whereas for the coefficient of the exogenous covariate β_2, the differences between the standard deviations of both estimators are more modest across all pairs of ρ and ψ.

Table 1.1 Average estimated coefficients and standard deviation across the 1000 replications using a rook type weighting matrix, N=400 and T=5.

		β_1		β_2		β_3	
		CF	2SLS	CF	2SLS	CF	2SLS
ρ	ψ						
0.0	0.0	0.704 (0.196)	0.698 (0.294)	0.606 (0.269)	0.604 (0.283)	0.298 (0.049)	0.300 (0.089)
	0.3	0.696 (0.188)	0.692 (0.269)	0.595 (0.254)	0.593 (0.272)	0.300 (0.047)	0.301 (0.080)
	0.7	0.703 (0.18)	0.705 (0.266)	0.600 (0.250)	0.601 (0.265)	0.300 (0.045)	0.299 (0.079)
0.3	0.0	0.690 (0.207)	0.683 (0.301)	0.589 (0.275)	0.584 (0.299)	0.303 (0.052)	0.305 (0.090)
	0.3	0.706 (0.191)	0.718 (0.281)	0.603 (0.276)	0.608 (0.294)	0.299 (0.049)	0.294 (0.084)
	0.7	0.704 (0.189)	0.697 (0.288)	0.599 (0.254)	0.595 (0.282)	0.299 (0.048)	0.301 (0.085)
0.7	0.0	0.695 (0.236)	0.698 (0.330)	0.573 (0.352)	0.575 (0.371)	0.302 (0.057)	0.301 (0.095)
	0.3	0.691 (0.231)	0.693 (0.334)	0.580 (0.335)	0.581 (0.360)	0.303 (0.054)	0.301 (0.095)
	0.7	0.704 (0.221)	0.693 (0.309)	0.603 (0.317)	0.600 (0.336)	0.300 (0.054)	0.303 (0.089)

To analyze the performance of the HACSC estimator, I use two metrics: first the average of the variance[7] estimated for each coefficient for each pair of ρ and ψ across the 1,000 replications and I compare it with the "true value", which is computed as the variance of the set of estimated coefficients for each pair of ρ and ψ across the 1,000 replications. Tables D.1-D.3 present this comparison and the first thing to note in the case of the CF is that both estimated variances, with

[7] I used the estimated variances instead of the standard errors because the nonlinearity of the square root function could affect the results.

and without the first stage correction, are very close to the true value so at first glance, using this metric the correction does not seem to make an impact.

For the 2SLS estimator, the differences are more substantial. The HACSC estimator is consistently closer to the true value across all pairs of rho and psi compared to the SHAC that imposes no serial correlation and the non-robust one. In general, the variance estimated with the HACSC is on average larger compared to the one computed with these two alternatives. Admittedly, in this case the cluster-robust variances are also very close to the true value. Overall these results suggest making the standard errors robust to spatial correlation at the expense of imposing no serial correlation can result in unreliable inference. Furthermore, as shown in Figure E.1, using the HACSC estimator will provided standard errors that are, on average, properly centered around the true value.

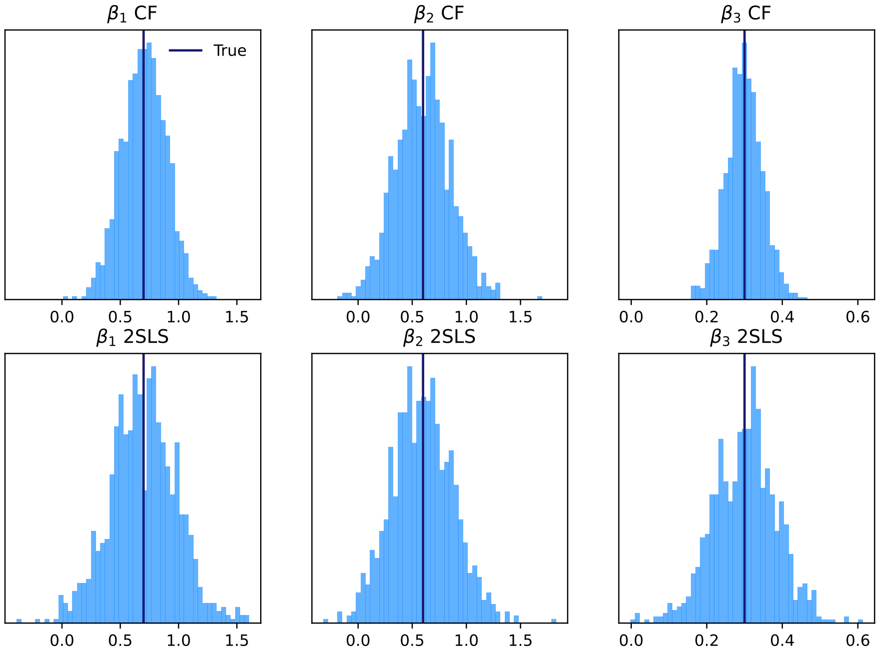


Figure 1.3 Distribution of coefficients estimated by 2SLS and the Control Function approach for $\rho = 0.3$ and $\psi = 0.7$ using a rook type weighting matrix.

As a second method to analyze the HACSC in this setup, I tested the null hypothesis H_0 : $\beta_3 = 0.3$ at a 5% of significance using a t-test over the 1,000 replications using the standard errors computed with the different estimators and I obtained the rejection probabilities. Using this metric, an estimator is performs better if the rejection probability is closer to 5%. Table 1.2 presents the results of this exercise.[8]

For the case of the CF approach, the rejection probabilities using the adjustment are slightly closer to 5% compared to the estimator that ignores the first stage so in this sense, the adjustment seems important to obtain more reliable inference if the researcher uses the CF approach. On the other hand, if we use 2SLS to estimate the coefficients, the HACSC estimator rejection probabilities are closer to the 5% compared to the SHAC and non-robust standard errors, which are over rejecting the null hypothesis. Using this metric, the cluster-robust standard errors seem to perform just as well as the HACSC estimator. Overall, the results suggest that the HACSC estimator, both in the case of 2SLS and the CF approach with the correction, provide more reliable inference compared to the existing SHAC.

Table 1.2 Rejection probabilities for the null hypothesis $H_0 : \beta_3 = 0.3$ at a 5% of significance using a t-test over the 1,000 replications with a rook type weighting matrix, N = 400, T=5.

		CF	CF_no1	HACSC	SHAC	Cluster	Non-Robust
ρ	ψ						
0.0	0.0	0.050	0.060	0.067	0.088	0.058	0.082
	0.3	0.046	0.060	0.054	0.072	0.046	0.068
	0.7	0.045	0.061	0.050	0.075	0.043	0.072
0.3	0.0	0.045	0.061	0.068	0.096	0.058	0.091
	0.3	0.050	0.064	0.047	0.074	0.040	0.067
	0.7	0.051	0.062	0.068	0.085	0.058	0.080
0.7	0.0	0.050	0.072	0.057	0.077	0.048	0.066
	0.3	0.041	0.057	0.066	0.095	0.065	0.090
	0.7	0.044	0.060	0.056	0.076	0.041	0.073

CF is the HACSC estimator using the first stage correction and CF_no1 refers to the HACSC estimator ignoring the first stage estimation using a CF approach.

1.8 Empirical application

To test the performance of the HACSC estimator with real world data, I revisit the problem of analyzing the effect of spending on the educational outcome of fourth graders in Michigan studied by Papke and Wooldridge (2008) using district level data from 1993 to 2001[9]. In short,

[8] Tables D.4 and D.5 show the results for $H_0 : \beta_1 = 0.7$ and $H_0 : \beta_2 = 0.6$ respectively.

[9] I want to thank Dr. Papke and Dr. Wooldridge for kindly sharing their data set.

Michigan changed the way schools were funded in 1994, going from a property-tax based system to a statewide system, which was possible trough an increase in the sales tax and lottery profits .

To measure the effect of spending on the academic achievement of students, the authors used as the dependent variable the fraction of fourth-graders that passed the math test (math4_{it}) of the Michigan Education Assessment Program (MEAP) given that the definition of this subject and the way it is evaluated has remained relatively constant over time. On the other hand, in addition to the current level of spending on a student, the authors also allow for the possibility that the level spending on the previous three years to play a role in the test scores. This is indeed a sensible choice given that one could argue that the previous years of education lay the foundations in the learning process of students.

The model also includes the proportion of students eligible for the free and reduced-price lunch program (lunch_{it}), the district enrollment (enroll_{it}) and time dummies. More details about the full model can be found in Papke (2005). Borrowing their notation, the estimated model is:

$$\text{math4}_{it} = \theta_t + \beta_1 \log(\text{avgrexp}_{it}) + \beta_2 \text{lunch}_{it} + \beta_3 \log(\text{enroll}_{it}) + c_{i1} + u_{it} \tag{1.49}$$

where avgrexp_{it} denotes the simple average of real spending from the current and previous three years. It is important to note that in addition to the linear probability model (LPM), Papke and Wooldridge (2008) also estimate the model with other nonlinear estimators but because they find that the LPM is a good approximation to the nonlinear estimates and since this paper focuses on linear models, we will compare the results only with their LPM results.

In order to replicate their results and use the HACSC estimator, we need a distance measure between the school districts. As mentioned in previous sections, this is not a trivial matter when we are working with geographical units but in this case, we will work with the geographic distance between the centroids of each district.[10] However, there have been changes in the school districts since 2001, which is why I could only use 98.6% of the original sample used by Papke and Wooldridge (2008). The main reason for this is that some districts have merged with others and in these cases, I used the data point of the district that absorbed the one disappearing. Table 1.3

[10] Roughly speaking, a centroid can be interpreted as the center of mass of a geometry.

compares the summary statistics from the original and new data sets and the t-tests show that there are no statistically significant differences between them.

Table 1.3 Sample means (standard deviations) of the original and new data sets and corresponding t-tests (p-values).

	1995			2001		
	Original	New	t-test	Original	New	t-test
Pass rate on fourth-grade	0.62	0.62	-0.30	0.76	0.76	-0.43
math test	(0.13)	(0.13)	(0.76)	(0.13)	(0.12)	(0.67)
Real expenditure	6329	6317	0.20	7161	7147	0.25
per pupil (2001$)	(986)	(978)	(0.85)	(933)	(916)	(0.80)
Real foundation	5962	5959	0.05	6348	6347	0.03
grant (2001$)	(1031)	(1035)	(0.96)	(689)	(692)	(0.98)
Fraction of eligible for	0.28	0.28	0.27	0.31	0.30	0.34
free and reduced lunch	(0.15)	(0.15)	(0.79)	(0.17)	(0.17)	(0.73)
Enrollment	3076	3099	-0.04	3078	3103	-0.05
	(8156)	(8210)	(0.97)	(7293)	(7341)	(0.96)
Number of observations	501	494	-	501	494	-

As a first step, I assume that all the explanatory variables are exogenous with respect to the error term u_{it} and apply the fixed effects estimator, sometimes referred to as "two-way fixed effects" because of the inclusion of the year dummies. Table 1.4 shows the estimates using the new data set and the ones reported by Papke and Wooldridge (2008). The coefficient associated with the average real expenditure is virtually the same, whereas the ones of lunch and the enrollment are negative with the new estimates. Nevertheless, the magnitudes of the latter are small and none of them are statistically significant in the original estimation either.

Table 1.4 Estimates assuming that all the explanatory variables are exogenous.

	Original results	New results	Standard errors for new results with different bandwidth values						
	Coefficient	Coefficient	ρ_b=1	ρ_b=100	ρ_b=200	ρ_b=300	ρ_b=400	ρ_b=500	ρ_b=600
log(avgrexp)	0.377	0.372	0.070	0.072	0.067	0.066	0.066	0.063	0.058
	(0.071)	(0.071)							
lunch	-0.042	0.029	0.064	0.077	0.079	0.072	0.061	0.060	0.061
	(0.073)	(0.064)							
log(enroll)	0.002	-0.02	0.048	0.045	0.033	0.028	0.026	0.023	0.022
	(0.049)	(0.048)							
Observations	501	493	-	-	-	-	-	-	-

Table 1.4 also shows the standard errors computed with the HACSC estimator using different bandwidth values. As expected and because the minimum distance between any two school districts in the data set is 1.05 kilometers, when the bandwidth is 1 kilometer the HACSC estimator is effectively treating the observations as if they have no effect on their neighbors (i.e. no spatial correlation) and consequently the standard errors are very similar to the ones computed using an estimator that is robust to heteroskedasticity and serial correlation. Interestingly, as the bandwidth increases, the standard error for each coefficients behaves differently: for the average spending, it first increases and then decreases, for enrollment it decreases monotonically whereas for lunch, there is not an evident pattern. Note that this exercise shows that even if the covariance matrix is robust to heteroskedasticity, spatial and serial correlation, this does *not* mean that the standard errors will be necessarily larger.

One of the issues with the estimates previously discussed is that the spending from a school district might be endogenous, mainly due to the fact that a school district might adjust its current spending if they suspect that the (bad) performance of a cohort throughout the year will be reflected on the pass rates of the MEAP test (Papke & Wooldridge, 2008). Fortunately, the change in the way that school districts brought with it a natural instrument: in the 1993/1994 school year, each district started to receive a per-student "foundation grant" based on the initial funding in 1994 that sought to increase the spending per student to a baseline level and had the effect of reducing the differences in spending between the districts across the state of Michigan by the year of 2001 (see Figure 1.4). The details of why this is a suitable instrument are discussed in Papke and Wooldridge (2008), but in broad terms, the identification assumption is that the idiosyncratic error term has a smooth relationship with both the dependent variable and the initial funding. On the other hand, the foundation grant depended on the initial funding in a non-smooth way [see Table 1 in Papke and Wooldridge (2008) to see this].

As a result of this concern, Papke and Wooldridge (2008) augmented the model by also including the real spending from 1994 with interactions with the time dummies, along with the time averages of lunch and enrollment, using as instruments the foundation grant interacted with the year binary

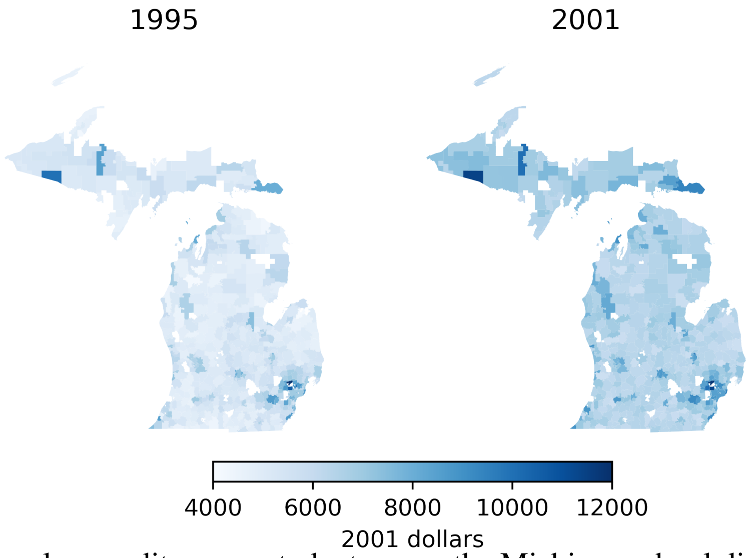


Figure 1.4 Average real expenditure per student across the Michigan school districts in 1995 and 2001.

variables. The new estimated model using instrumental variables is then

$$\begin{aligned} \text{math4}_{it} &= \theta_t + \beta_1 \log(\text{avgrexp}_{it}) + \beta_2 \text{lunch}_{it} + \beta_3 \log(\text{enroll}_{it}) \\ &+ \beta_{4t} \log(\text{rexppp}_{i,1994}) + \xi_1 \overline{\text{lunch}_i} + \xi_2 \overline{\log(\text{enroll}_i)} + v_{it1}. \end{aligned} \tag{1.50}$$

Note that because we have a single endogenous variable, in this case using Two Stage Least Squares (2SLS) would be numerically the same as estimating the model with the control function approach, and because of this, I used the latter. Table 1.5 shows the estimates from this model and once again, the coefficients obtained using the new sample are very similar to the ones computed using the original data set. In particular, the coefficient of the spending is considerable larger than the OLS estimate, which can be explained in the context of the local average treatment effect literature or by the fact that district authorities can decide to increase spending whenever they think the cohort might underperform Papke and Wooldridge (2008).

Table 1.5 Estimates assuming that the spending variable is endogenous.

	Original results	New results	Standard errors for new results with different bandwidth values						
	Coefficient	Coefficient	ρ_b=1	ρ_b=100	ρ_b=200	ρ_b=300	ρ_b=400	ρ_b=500	ρ_b=600
log(avgrexp)	0.555 (0.208)	0.546 (0.211)	0.221	0.265	0.292	0.253	0.221	0.202	0.187
lunch	-0.062 (0.075)	0.008 (0.067)	0.066	0.077	0.083	0.079	0.07	0.068	0.067
log(enroll)	0.046 (0.067)	0.023 (0.066)	0.069	0.075	0.079	0.071	0.065	0.058	0.054
v	-0.421 (0.232)	-0.476 (0.236)	0.250	0.349	0.411	0.383	0.365	0.357	0.353
Observations	501	493	-	-	-	-	-	-	-

Contrary to the case where all the independent variables were treated as exogenous, the standard errors computed using the HACSC estimator when the bandwidth parameter is set to 1 kilometer are somewhat different to the ones computed using an estimator that is only robust to serial correlation and heteroskedasticity, which is expected because the latter does not take into account the first stage estimation. Once again this results show that the standard errors can be larger or smaller depending on the value selected for the bandwidth.

So far I have assumed that there is only spatial correlation in the error term. However in this scenario there could be spatial spillovers from neighboring units that could be affecting the student performance on the math test. Figure 1.4 not only shows that the average real expenditure per student increased between 1995 and 2001 in all the school districts, but it also shows the spatial distribution of it. Note that there are districts where the surrounding neighbors have a very similar level of spending, for example, in 1995 the Detroit region shows multiple school districts with higher levels of expenditure compared to the rest of the state. Similarly, in Figure 1.5 the Upper Peninsula shows several neighboring school districts with higher passing rates than the rest of the region.

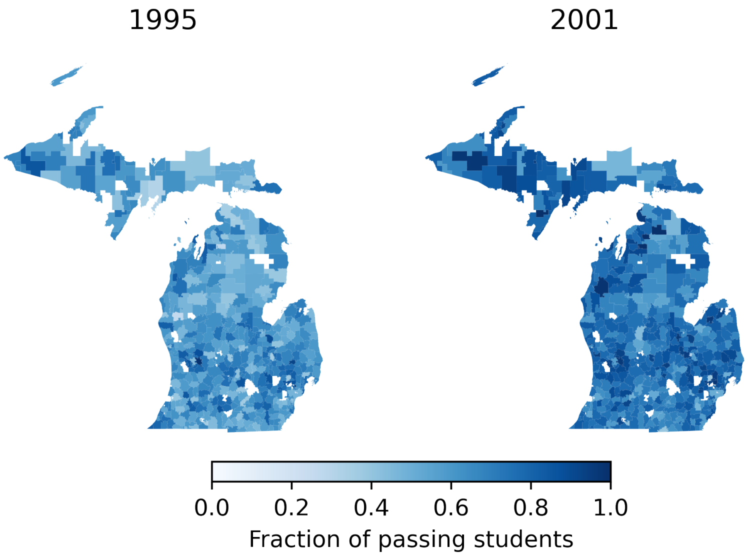


Figure 1.5 Average real expenditure per student across the Michigan school districts in 1995 and 2001.

Multiple reasons could be behind this pattern. For instance, it could be the case that parents with students that are underperforming identify school districts that are increasing spending and throughout the year, move to one of these districts in order to increase help their children to improve their grades. From the labor side, school districts might need to increase the expenditure in teachers' salaries to avoid losing them to other school districts within a reasonable commuting distance. All in all, it seems important to control for spillover effects of expenditure from neighbors, so I augment the models previously estimated with this additional variable[11] and Table 1.6 shows the estimates of this regression assuming that all the independent variables are exogenous with respect to the error term. Note that the coefficient on the average expenditure has decreased significantly so that an increase of approximately 10% in spending will now lead to an increase in the pass rate of about 2.8%. On the other hand, if neighboring school districts of unit i increase their expenditure around 10%, the pass rate in i is expected to improve around 3.2%, a larger effect than the own spending. To address the endogeneity issue, I also augmented the model 1.50 with the spending spillover variable using the control function approach[12] and the results are shown in Table 1.7.

[11] For this estimation, I used a rook type weighting matrix

[12] I used $W \cdot \log(\text{found})$ to instrument for $W \cdot \log(\text{avgrexp})$

Table 1.6 OLS with extension

	Coefficient	Standard errors with different bandwidth values						
	(st. error)	ρ_b=1	ρ_b=100	ρ_b=200	ρ_b=300	ρ_b=400	ρ_b=500	ρ_b=600
log(avgrexp)	0.281 (0.076)	0.076	0.077	0.071	0.067	0.065	0.061	0.056
lunch	0.030 (0.063)	0.063	0.077	0.082	0.076	0.066	0.064	0.064
log(enroll)	-0.008 (0.047)	0.047	0.044	0.035	0.03	0.028	0.025	0.024
W· log(avgrexp)	0.324 (0.090)	0.088	0.076	0.071	0.057	0.049	0.047	0.047
Number of districts	493	-	-	-	-	-	-	-

Once again, in this case the effect of the own expenditure is larger than in the exogenous case, but it is smaller compared to the original estimate. The spillover effect is significantly reduced to a marginal increase of around 0.7% in the pass rates due to an increase in the spending in surrounding school districts and moreover, the coefficient is not statistically significant.

Overall, the difference in the magnitude of the coefficients obtained for the spending in neighboring units make it difficult to interpret the effect of this variable. However in both cases it was positive, which supports the hypothesis that parents may move to school districts where the spending per student is higher. Of course, one cannot rule out the possibility that larger spending by neighboring school districts can attract better teachers to the area that are willing to commute, however, more detailed data may be needed to separate these effects.

Regarding the standard errors, most of the results show a pattern: if the bandwidth parameter is too small, they seem to be smaller relative to the ones computed with larger values, but at some point they become smaller again. This phenomenon has been documented in the time series literature: for example, Müller (2014) argues that when the bandwidth is too small, the estimate of the covariance matrix is downward biased. In the same line, Kiefer and Vogelsang (2005) show that for an AR(1) process, if the bandwidth is too small the estimator of is biased, whereas if every observation is given a weight of one in the estimation of the covariance matrix, then the estimates are going to tend to zero because the in-sample residuals have an average of zero, which is precisely what is being observed in this example as the bandwidth increases beyond some point.

Table 1.7 IV extension

	Coefficient	Standard errors for new results with different bandwidth values						
	(st. error)	ρ_b=1	ρ_b=100	ρ_b=200	ρ_b=300	ρ_b=400	ρ_b=500	ρ_b=600
log(avgrexp)	0.408 (0.231)	0.234	0.317	0.361	0.310	0.262	0.234	0.219
lunch	0.016 (0.067)	0.066	0.078	0.087	0.082	0.074	0.072	0.071
log(enroll)	-0.001 (0.067)	0.068	0.08	0.088	0.079	0.069	0.062	0.058
W · log(avgrexp)	0.071 (0.057)	0.056	0.076	0.083	0.077	0.07	0.067	0.065
v	-0.249 (0.254)	0.260	0.379	0.435	0.385	0.346	0.328	0.318
Number of districts	493	-	-	-	-	-	-	-

1.9 Conclusion

In this paper, I present a simple way to obtain standard errors that are robust to heteroskedasticity and both serial and spatial correlation in short panels with fixed effects and endogenous covariates. This is important because to the best of my knowledge, the current SHAC estimators do not explicitly allow for serial correlation in this context (admittedly the literature does not ignore this issue when $T \to \infty$). The estimator relies on averaging the moment conditions for a single individual across time, which allows to treat the estimation like a cross sectional problem without imposing any restrictions on the serial correlation of the residuals. This will help empirical researchers to obtain more reliable standard errors in different fields such as urban economics or international trade.

The proposed HACSC estimator can be directly applied in a Correlated Random Effects framework to obtain a fully robust Hausman-type test, which can help empirical researchers to choose between Fixed Effects and Random Effects specifications. In this paper I also showed that the Mundlak equivalence also holds in a particular spatial setting, which will allows to obtain the Fixed Effects coefficients of the time varying covariates in a Random Effects context. Similarly, the HACSC estimator can be used in a RE estimation procedure, whenever the researcher suspects that the structure imposed of the spatial error term might be misspecified.

I also presented a control function approach and the required assumptions to estimate the parameters of the model. Although even in the i.i.d. case it is a standard practice to use bootstrap to obtain the standard errors with this approach, in a spatial setting this is not a trivial procedure given the dependence between observations. For this reason, I also extended the HACSC estimator to this setup, which requires an adjustment of the covariance matrix to take into account the sampling error of the first stage estimation.

The Monte-Carlo experiment performed showed that the HACSC estimator works well in the presence of strong or moderate serial and spatial correlation compared to other methods used by the literature in terms of obtaining unbiased standard errors. As expected, the estimator also shows higher variance than such estimators, especially in settings with low spatial and/or serial correlation. The simulations also showed that if the CF assumptions hold, we can obtain efficiency gains compared to 2SLS.

An avenue for future research is to extend the Monte Carlo experiments in different directions. First, it would be interesting to use different weighting schemes for the weighting matrix W based on distance or a k-neighbor scheme in an irregular lattice, as well as different kernel functions. Analogous to the time series literature, the threshold for the distance bandwidth most certainly plays an important role on the finite sample behavior of the estimator, so implementing a data driven procedure to choose it is also a possibility to explore, particularly when the spatial correlation is strong.

CHAPTER 2

ESTIMATION OF MODELS WITH SPATIAL PANELS AND MISSING OBSERVATIONS IN THE COVARIATES

2.1 Introduction

Over the last years, the amount and type of data available for economic research has experienced an important increase. Many fields in economics have benefited from this, including areas that focus on spatial related issues such as development, trade, geography and urban economics. Unfortunately, a common issue that empirical researchers have to deal with is missing data, a problem that can arise in multiple ways and which often leads to the need of different methods, one of which is the use of the "complete cases" only, in other words, observations where either the response variable or one of the covariates is missing are dropped from the analysis.

The consequences of this will depend on the assumptions and the process that generates the missing data, but regardless of these, discarding observations results in a loss of information. This problem is more serious in a spatial context, where it is common to include spillover effects from "neighboring" units (i.e. spatial lags) in the model. For example, if we are working with county level data and the nature of the dependence between the units is a function of the geographical distance between them, the researcher might include the effects of surrounding counties as an additional explanatory variable using a weighting matrix W. However, in this setup if a unit i is a "neighbor" of l counties and i has a missing data point and the researcher is using the complete cases only, she might need to drop not only observation i, but all of its l neighbors as well, therefore, the loss of information in the spatial case is potentially more severe.

Furthermore, if we have a panel data set, the problem will be aggravated because the missing data could affect both dimensions. This is in fact a common problem in empirical work because the reason of the missing data could be that the units of observation (e.g. countries) have different lengths of their time series (i.e. unbalanced panel). Given this, a method to impute data in this case would be useful for empirical work so that the efficiency loss induced by the missing data is mitigated with respect to using only the complete cases. This work tries to fill this necessity by proposing a new GMM estimation procedure.

The problem of missing data has been a known issue in economic research for a long time. One of the approaches that empirical researchers use to deal with it, is dropping the incomplete cases, which induces an efficiency loss as mentioned previously. In this respect, Kelejian and Prucha (2010) present conditions in a spatial setting under which the missing data can be ignored asymptotically based on the proportion of the sample sizes related to the complete and incomplete observations. They also describe the case where the missing data cannot be ignored and will make inference more difficult.

In practice, there are alternatives other than just using the complete cases: for example, one might try to complete the sample first and then estimate the model using this "complete" data set. One of the methods documented in the spatial literature was introduced by Lesage and Pace (2004), who used the Expectation-Maximization algorithm to predict the value of the dependent variable that are missing in the context of real estate housing prices.

In the spatial context, one could also generate the spatial lags using the available data only, in which case the researcher has two options. First, a common practice is to replace the missing data with zeros (Kelejian & Prucha, 2010), nevertheless this technique does not seem sensible as having missing data is very a different problem and replacing these data points with zeroes will almost certainly lead to biased estimates. The second approach involves constructing the spatial lags using only the available "neighbors", but doing this could generate a misspecification of the weighting matrix and thus probably yield inconsistent estimates, as pointed out by Wang and Lee (2013). More concretely, if a unit i has four "neighbors", each of which has the same weight and the weighting matrix W is row-normalized then in theory each pseudo-neighbor should have a weight of $\frac{1}{4}$. However, if the data for one of the pseudo-neighbors is missing, then one would assign a weight of $\frac{1}{3}$ to each available unit, thus mispecifying W.

In non spatial settings, the literature has proposed multiple ways to deal with missing data. For instance, Dagenais (1973) proposed a generalized least squares estimator in which the missing variables are approximated using observed covariates. In a similar spirit and in the context of linear models, Gourieroux and Monfort (1981) present a maximum likelihood procedure in which the missing variables are explained by the observed ones. More recently, Dardanoni et al. (2011) suggest a framework with an augmented model to reduce the bias induced by replacing the missing observations with imputed values.

Abrevaya and Donald (2017) introduced a GMM framework in linear models in which they exploit moment conditions on the missing observations on the regressors to obtain an estimator that the claim to be more efficient than other estimators previously mentioned such as Dagenais'. Rai (2021) considered the panel data case of their estimator and find that it is more efficient than the fixed effects and correlated random effects using the Mundlak device that use only the complete cases. Rai (2023) extended their approach to the case of missing dependent variables and endogenous explanatory covariates, which is useful in cases where the researcher needs to combine data sets from different sources.

Going back to a spatial context, Wang and Lee (2013) also suggest three estimation procedures in the context a missing dependent variable only in the cross sectional case. They propose a GMM estimator based on linear moment conditions, a nonlinear least squares and a two stage least squares with imputation and compare the asymptotic properties of the three estimators. Wang and Lee (2013) extend the previous estimators to the case of spatial autoregressive panels using a random effects framework as a baseline and then generalize it by presenting the spatial Mundlak approach. Note that their work also focuses on the cases of a missing dependent variable only.

It is important to note that in the non spatial case, the Mundlak approach falls within the correlated random effects context, a middle ground between the random effects (RE) and fixed effects (FE) estimators. In the first case, the researcher must assume that there is no correlation between the explanatory variables in the model and the individual heterogeneity, whereas in the second case, this assumption is relaxed and allows these terms to be correlated. In this respect, Mundlak (1978) argues that the RE version is a misspecification of the FE model as it does not take into account the correlation between the heterogeneity and the regressors. To solve this problem, he proposed an auxiliary equation where the heterogeneity is modeled as a function of the time averages of the independent variables. By doing this, he shows that if we add these time averages to the main equation and estimate the model by RE, we will obtain the same numerical coefficients as if we estimate the model by FE. This equivalence carries over to the unbalanced panel case if we only use the complete cases, as shown by Wooldridge (2019) and byJoshi and Wooldridge (2019) for the case with endogenous covariates.

Debarsy (2012) was the first to extend the Mundlak approach to a spatial setting if the researcher working with a Spatial Durbin Model[1] (SDM). Nevertheless, this work does not show the aforementioned equivalence between the RE and FE specifications. In 2020, Li and Yang demonstrated that when the error term is modeled structurally[2], a very common practice in the spatial literature, then the equivalence holds *conditional* on the value of the parameter(s) associated with the error term, otherwise the FE and RE will yield different estimates generally. In addition, Wu-Chaves (2024) shows that when the error term is not modeled structurally, the equivalence holds if the model is estimated by ordinary least squares (OLS) or two-stage least squares (2SLS). One of the limitations of the work just described is that they focus on the case where the data is complete. In this paper, I will show that in the case of an unbalanced spatial panel, the the CRE equivalence also holds if the researcher uses the complete cases only to estimate the model.

In this chapter, I extend the work of Abrevaya and Donald (2017) and Rai (2021) to the case of spatial panels with spillover effects. The rest of the paper is organized as follows. Section 2.2 presents the model. Sections 2.3 and 2.4 state the assumptions and show the construction of GMM estimator, respectively. Section 2.5 shows the equivalence between FE and RE. Section 2.6 provides Monte-Carlo evidence related to the performance of the GMM estimator. Section 2.7 illustrates an empirical application of the estimator and Section 2.8 concludes.

[1] A SDM includes both a spatial lag of the dependent and independent variables on the right hand side of the equation.

[2] Note that by modeling the error term usually involves MLE or a GMM estimation.

2.2 Model

Consider the following model:

$$\begin{aligned} y_{it} &= x_{1it}\beta_1 + x_{2it}\beta_2 + W_iX_{1t}\gamma_1 + W_iX_{2t}\gamma_2 + c_i + u_{it} \\ &= x_{it}\beta + w_{it}\gamma + c_i + u_{it},\ i = 1\ldots N, t = 1\ldots T. \end{aligned} \tag{2.1}$$

where y_{it} is the response variable, x_{1it} is a $1 \times (k_1 + 1)$ set of exogenous variables that includes an intercept, x_{2it} is a $1 \times k_2$ vector of endogenous covariates (with $k_1 + k_2 = k$), $w_{it} = (w_{1it} \;\; w_{2it}) = (W_iX_{1t} \;\; W_iX_{2t})$ with W_i being the i-th row of an exogenous, non random, time invariant $N \times N$ weighting matrix, X_{1t} and X_{2t} are the $N \times k_1$ and $N \times k_2$ matrices of exogenous and endogenous covariates, respectively, for all observations at time t, c_i is the individual heterogeneity and u_{it} is the idiosyncratic error term. In this type of model, the terms W_iX_{1t} and W_iX_{2t} are known as spatial lags and they capture the effect of neighboring units on unit i's outcome[3]. $(\beta \;\; \gamma)$ are the parameters of interest and they are of dimension $(k + 1) \times 1$ and $k \times 1$ respectively[4].

In this paper, I will treat the error term in a non parametric way so that it might be serially and spatially correlated, but I do not impose any particular structure on it. The sense in which x_{1it} is exogenous is that it is uncorrelated with the error term u_{it} (i.e. $\mathbb{E}(x'_{1it}u_{it}) = 0$). Analogously the endogeneity of x_{2it} arises from the fact that $\mathbb{E}(x'_{2it}u_{it}) \neq 0$. I also assume that the asymptotics refer to the case where $N \longrightarrow \infty$ while T remains fixed.

Since x_{2it} is endogenous, we need a set of external instruments z_{2it} of dimension $l \times 1$ ($l \geq k_2$) that satisfy the usual requirements of relevance and exogeneity with respect to the error term u_{it}, that is $\mathbb{E}(z'_{2it}u_{it}) = 0$. Naturally, $\mathfrak{z}_{2it} = W_iZ_{2t}$ can be used as the instrument for W_iX_{2t}. For ease of notation, let $a_{it} = (x_{1it} \;\; x_{2it} \;\; w_{1it} \;\; w_{2it})$, and $z_{it} = (x_{1it} \;\; z_{2it} \;\; w_{1it} \;\; \mathfrak{z}_{2it})$. Under these assumptions,

[3] It is common to also include a spatial lag of the outcome variable, however by doing so the interpretation of the model as a conditional mean function is lost. For this reason, I am omitting this term in the paper.

[4] From a modeling perspective, it is not necessary to include all k variables in the spatial lag so the dimension of γ could me smaller

the set of first stage equations is:

$$
\begin{aligned}
x_{1it} &= x_{1it}\pi_{11} + z_{2it}\pi_{12} + w_{1it}\pi_{13} + \mathfrak{z}_{2it}\pi_{14} + r_{1it} \\
x_{2it} &= x_{1it}\pi_{21} + z_{2it}\pi_{22} + w_{1it}\pi_{23} + \mathfrak{z}_{2it}\pi_{24} + r_{2it} \\
w_{1it} &= x_{1it}\pi_{31} + z_{2it}\pi_{32} + w_{1it}\pi_{33} + \mathfrak{z}_{2it}\pi_{34} + r_{3it} \\
w_{2it} &= x_{1it}\pi_{41} + z_{2it}\pi_{42} + w_{1it}\pi_{43} + \mathfrak{z}_{2it}\pi_{44} + r_{4it} \qquad (2.2)
\end{aligned}
$$

where $\pi_{j1}, \pi_{j2}, \pi_{j3}$ and π_{j4} are vectors of dimensions $(k_1 + 1) \times (k_1 + 1)$, $l \times k_1$, $k_1 \times k_1$ and $l \times k_1$ respectively for $j = 1, 2, 3, 4$. Of course, the relevant equations of (2.2) are the second and fourth lines as $(x_{1it} \quad w_{1it})$ act as their own instruments. Given this, (2.1) and (2.2) can be written more compactly as:

$$y_{it} = a_{it}\theta + c_i + u_{it} \qquad (2.3)$$

$$a_{it} = z_{it}\pi + r_{it} \qquad (2.4)$$

where $\theta = (\beta \quad \gamma)$. By definition, $\mathbb{E}(z'_{it}r_{it}) = 0$ and because the instruments are relevant, it follows that $\pi^0 \neq 0$. Note that other than the exogeneity with respect to z_{it}, no other assumptions have been imposed on the error terms in (2.3) and (2.4). Furthermore, y_{it} can be expressed in terms of z_{it} as follows:

$$y_{it} = z_{it}\theta\beta + c_i + u_{it} + r_{it}\beta = z_{it}\theta\beta + c_i + v_{it} \qquad (2.5)$$

The parameters of this model can be consistently estimated by applying Fixed Effects Two Stage Least Squares (FE2SLS) or equivalently by applying Pooled 2SLS (P2SLS) to

$$\ddot{y}_{it} = \ddot{a}_{it}\theta + \ddot{u}_{it} \qquad (2.6)$$

using the instruments $\ddot{z}_{it}$ and where $\ddot{y}_{it} = y_{it} - \bar{y}_i$, $\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}$ and similar definitions apply to the other variables, provided that the corresponding rank conditions of the relevant matrices hold and

$$\mathbb{E}(\ddot{z}'_{it}\ddot{u}_{it}) = 0 \qquad (2.7)$$

The latter is implied by the following condition:

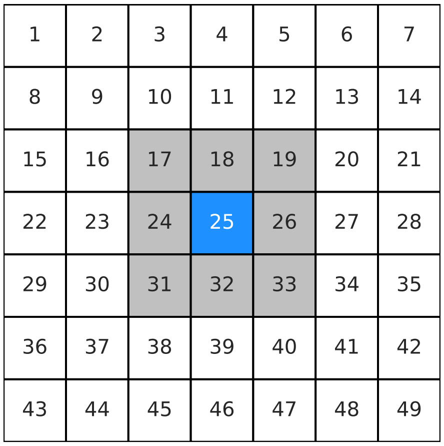
$$\mathbb{E}(u_{it}|Z, C) = 0 \tag{2.8}$$

which is a strict exogeneity assumption and where Z is the entire matrix of exogenous variables and C is the whole vector of individual heterogeneities. Note that (2.8) is a stronger condition than the classical strict exogeneity assumption because in this case the idiosyncratic error term at time t is not only uncorrelated with the exogenous variables at any time period, but it is also uncorrelated with the covariates of other units due to the nature of the spatial panel data set and in particular to the presence of the spatial lags. It is also important to emphasize that in this setup, the individual heterogeneity c_i is allowed to be arbitrarily correlated with the elements of z_{it} or the endogenous covariates.

2.3 Missing data mechanism

Before formalizing the missing data scheme, consider the consequences of missing observations in a model with spatial spillovers compared to a situation without such effects. As previously mentioned, a typical strategy when empirical researchers have missing data is to estimate the parameters use only the observations that have a full set of observed variables and discard the units that are incomplete. If we have a sample of 49 individuals living in a regular grid as shown in Figure 2.1 and there are no spillover effects and the researcher has no data on unit 25, then the loss of information is relatively small (around 2% of the sample). On the other hand, if the model contains spillover effects from neighboring units and we are using a queen type weighting scheme[5], then if unit 25 is missing and the researcher decides to use only the complete observations, then she would have to disregard unit 25's neighbors too (shown in gray in Figure 2.1) and end up losing almost 20% of the original sample.

[5] Under this weighting mechanism, a neighbor is an unit that shares an edge or a vertex.



1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31	32	33	34	35
36	37	38	39	40	41	42
43	44	45	46	47	48	49

Figure 2.1 Regular grid with unit 25 missing and its neighbors shown in gray.

The previous example shows that missing observations can result in a severe decrease in efficiency when estimating the parameters in a model with spillover effects. To formalize the missing mechanism, let

$$s_{it} = \begin{cases} 1 \text{ if } x_{1it} \text{ is observed for unit } i \text{ at time } t \\ 0 \text{ otherwise} \end{cases} \tag{2.9}$$

and let S_t be the $N \times N$ diagonal matrix with diagonal elements s_{it}. Note that s_{it} is indicating that the researcher observes either the full set of endogenous variables or none at all. Furthermore, I am also assuming that the response variable and the exogenous variables z_{it} are always fully observed for all individuals in all time periods. A common practice in empirical work is to ignore the missing data from neighbors in the spatial lag, so implicitly the missing neighbors are being assigned a weight of 0 (Kelejian & Prucha, 2010). This being the case, WS_tX_t would be enough to select units with available self-information but with possibly incomplete data on their the spatial lag. However, to select only the complete cases in the spatial lag, a new variable needs to be defined. To this end, for each i and its J neighbors, let

$$\tilde{s}_{it} = s_{it} \cdot \prod_{j=1}^{J} s_{jt} \tag{2.10}$$

so that $\tilde{s}_{it} = 1$ only when the full set of endogenous variables are observed for unit i and its corresponding neighbors. Then define $\tilde{S}_t$ as the diagonal matrix with diagonal elements $\tilde{s}_{it}$ so that $W\tilde{S}_t X_t$ will select only the fully complete cases. As previously mentioned, the researcher can consistently estimate the parameters using FE2SLS with the complete cases if (2.8) holds, at the expense of losing efficiency. More concretely, the estimator can be defined as follows:

$$\hat{\theta}_{CFE2SLS} = \left[\left(\sum_{i=1}\sum_{t=1}\tilde{s}_{it}\ddot{a}'_{it}\ddot{z}_{it}\right)\left(\sum_{i=1}\sum_{t=1}\tilde{s}_{it}\ddot{z}'_{it}\ddot{z}_{it}\right)^{-1}\left(\sum_{i=1}\sum_{t=1}\tilde{s}_{it}\ddot{z}'_{it}\ddot{a}_{it}\right)\right]^{-1} \cdot$$

$$\left(\sum_{i=1}\sum_{t=1}\tilde{s}_{it}\ddot{a}'_{it}\ddot{z}_{it}\right)\left(\sum_{i=1}\sum_{t=1}\tilde{s}_{it}\ddot{z}'_{it}\ddot{z}_{it}\right)^{-1}\left(\sum_{i=1}\sum_{t=1}\tilde{s}_{it}\ddot{z}'_{it}\ddot{y}_{it}\right) \tag{2.11}$$

where $\ddot{y}_{it} = y_{it} - \bar{y}_i$ and where $\bar{y}_i = \frac{1}{T_i}\sum_q \tilde{s}_{iq}y_{iq}$ and $T_i = \sum_q^T \tilde{s}_{iq}$. The rest of the variables are similarly defined. Note that T_i is a random variable as it is a function of the selection[6]. In words, for each unit i the time average is computed using only the periods where the observation as a full set of observed variables.

To develop an alternative estimator, consider again the within transformation of the variables similar to the one in (2.11), that is, the averages are computed using only the complete cases. Furthermore, define:

$$\dot{x}_{2it} = x_{2it} - \frac{1}{T - T_i}\sum_{t=1}^{T}(1 - \tilde{s}_{it})x_{2it} \tag{2.12}$$

that is, $\dot{x}_{2it}$ is a within transformation where the average is computed using the incomplete cases only and similar definitions apply to other variables. Regarding this transformation, it is important to point out that it may be possible that only one time period for a particular unit i is missing, in which case the within transformation (2.12) will remove that observation as the "average" is taken over a single time period. In such cases, these units are uninformative and are essentially removed from the estimation and therefore will not help to provide efficiency gains. Note that by applying the within transformation from (2.11) to the main model, the term c_i disappears. The resulting estimating equations are

$$\ddot{y}_{it} = \ddot{a}_{it}\theta + \ddot{u}_{it} \tag{2.13}$$

$$\ddot{a}_{it} = \ddot{z}_{it}\pi + \ddot{r}_{it} \tag{2.14}$$

[6] Note that I am implicitly assuming that $\Pr(T_i = 0) = 0$ for all i so that $\theta_{CFE2SLS}$ is well defined.

And by replacing (2.14) in (2.13), we obtain an expression of y_{it} in terms of the always observed variables:

$$\begin{aligned} \ddot{y}_{it} &= (\ddot{z}_{it}\pi + r_{it})\theta + \ddot{u}_{it} \\ &= \ddot{z}_{it}\pi\theta + \ddot{v}_{it} \end{aligned} \tag{2.15}$$

where $\ddot{v}_{it} = \ddot{u}_{it} + \ddot{r}_{it}\theta$. In order to obtain efficiency gains and still get a consistent estimator using fixed effects, consider the following assumption:

Assumption 1

$i)$ $\mathbb{E}(\tilde{s}_{it}\ddot{z}'_{it}\ddot{u}_{it}) = 0$ $\qquad ii)$ $\mathbb{E}(\tilde{s}_{it}\ddot{z}'_{it}\ddot{r}_{it}) = 0$ $\qquad iii)$ $\mathbb{E}[(1 - \tilde{s}_{it})\dot{z}'_{it}\dot{v}_{it}] = 0$

Part $i)$ of Assumption 1 imposes that θ is the same in both the complete and incomplete cases and it is also necessary for the complete cases estimator. The second and third points of Assumption 1 are the basis of the potential efficiency gains that can be achieved with the proposed estimator. Specifically, $ii)$ states that π is the same in both the observed and unobserved samples, whereas $iii)$ (along with $i)$) amounts to say that the model and the imputation method are the same for the complete and incomplete observations. Similarly to the non-missing units case, the conditions in 1 are implied by the following zero conditional mean assumptions:

Assumption 2

$i)$ $\mathbb{E}(u_{it}|Z, S, C) = 0$ and $ii)$ $\mathbb{E}(r_{it}|Z, S, C) = 0$

These are strict exogeneity conditions analogous to the non spatial case. Note that these are weaker than a missing at random (MAR) mechanism, in which the missingness is allowed to depend on the always observed data (Little & Rubin, 2019). More formally and borrowing notation from Rai (2023), in this context the data would be considered MAR if $s_{it} \perp (y_{it}, x_{1it}, w_{1it})|Z$ or equivalently $s_{it} \perp (u_{it}, r_{1it}, r_{3it})|Z$, where r_{1it} and r_{3it} are the errors related to x_{1it} and w_{1it} respectively in the first stage. A sense in which Assumption 2 is weaker than MAR is that in the former, the condition would still hold if the selection is a function of Z, provided that $\mathbb{E}(r_{it}|Z) = 0$. Both of these assumptions are weaker than the missing completely at random (MCAR) mechanism, where the probability of missing is independent of the rest of the variables, i.e. $s_{it} \perp (y_{it}, x_{1it}, w_{1it}, z_{it})$.

2.4 GMM estimation

Using equations (2.13), (2.14), (2.15) and Assumption 1, we can create a vector of moment conditions to perform GMM estimation. Let[7]

$$g_{it}(\theta,\pi) = \begin{bmatrix} \tilde{s}_{it}\ddot{z}'_{it}\ddot{u}_{it} \\ \tilde{s}_{it}\ddot{z}'_{it} \otimes \ddot{r}'_{it} \\ (1-\tilde{s}_{it})\dot{z}'_{it}\dot{v}_{it} \end{bmatrix} = \begin{bmatrix} \tilde{s}_{it}\ddot{z}'_{it}(\ddot{y}_{it} - \ddot{a}_{it}\theta) \\ \tilde{s}_{it}\ddot{z}'_{it} \otimes (\ddot{a}_{it} - \ddot{z}_{it}\pi)' \\ (1-\tilde{s}_{it})\dot{z}'_{it}(\dot{y}_{it} - \dot{z}_{it}\pi\theta) \end{bmatrix} = \begin{bmatrix} g_{1it}(\theta,\pi) \\ g_{2it}(\theta,\pi) \\ g_{3it}(\theta,\pi) \end{bmatrix} \tag{2.16}$$

Since I am assuming that Assumption 2 holds, it follows that $\mathbb{E}[g(\theta^0,\pi^0)] = 0$, where (θ^0,π^0) is the vector of true population parameters. Note that $g_{1it}(\cdot)$ and $g_{2it}(\cdot)$ use the complete cases, while the $g_{3it}(\cdot)$ moment condition uses the incomplete cases. Furthermore $g_{it}(\cdot)$ provides $[2(k_2+l)+1][2k+3]$ moment conditions, while there are $2(2k+1)(k_2+l+1)$ parameters to estimate, which leaves $2(2l+k_2-k_1)+1$ overidentifying restrictions. Once again it is important to note that the potential efficiency gains from the proposed estimator come from imposing that the π is the same among the observed and unobserved units and that the model and imputation method are the same among those same groups.

In order to obtain an efficient GMM estimator, we need to construct an optimal weighting matrix. let:

$$V \equiv \mathbb{E}\left[g(\theta,\pi)g(\theta,\pi)'\right] = \mathbb{E}\begin{bmatrix} V_{11} & V_{12} & 0 \\ V'_{12} & V_{22} & 0 \\ 0 & 0 & V_{33} \end{bmatrix} \tag{2.17}$$

where,

$$\begin{aligned} V_{11} &= \tilde{s}_{it}\ddot{u}^2_{it}\ddot{z}'_{it}\ddot{z}_{it} & V_{12} &= \tilde{s}_{it}\ddot{z}'_{it}\ddot{u}_{it}\ddot{z}'_{it} \otimes \ddot{r}_{it} \\ V_{22} &= \tilde{s}_{it}\ddot{z}'_{it} \otimes \ddot{r}'_{it}\ddot{z}_{it} \otimes \ddot{r}_{it} & V_{33} &= (1-\tilde{s}_{it})\dot{z}'_{it} \otimes \dot{v}^2_{it}\dot{z}'_{it}\dot{z}_{it} \end{aligned} \tag{2.18}$$

In this setup, the sample GMM objective function is:

$$\bar{g}(\theta,\pi)'\hat{\Omega}\bar{g}(\theta,\pi) \tag{2.19}$$

[7] Formally $g_{it}(\cdot)$ is also a function of the $(\ddot{y}, \ddot{a}, \ddot{z}, \tilde{S})$, but for notation simplicity I suppress these.

where $\bar{g}(\theta, \pi) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} g_{it}(\theta, \pi)$, Ω is a square, non-random, symmetric and positive semi-definite matrix of order $[2(k_1 + l) + 1][2k + 3]$ and $\hat{\Omega}$ is a consistent estimator of Ω. To obtain $\hat{\Omega}$, we can replace the expectations with sample averages in (2.21) and (2.17) and we can get consistent estimates of $\ddot{u}_{it}$, $\ddot{r}_{it}$ and $\dot{v}_{it}$ by applying GMM to to $g_{1it}(\cdot)$, only $g_{2it}(\cdot)$ and $g_{3it}(\cdot)$ only, respectively. It is noteworthy to point out that restricting π to be different across $g_{2it}(\cdot)$ and $g_{3it}(\cdot)$ make these moment functions redundant in the estimation of θ, as pointed out by Ahn and Schmidt (1995) and Rai (2023).

The proposed estimator in this paper minimizes (2.19) with respect to (θ, π) using $\hat{\Omega} = \hat{V}^{-1}$ and will be denoted as $(\hat{\theta}, \hat{\pi})$. Before stating the asymptotic normality result, we need to present the other component of the covariance matrix for the GMM estimator. First, let

$$D \equiv \mathbb{E}\left[\nabla g(\theta^0, \pi^0)\right] = \mathbb{E}\begin{bmatrix} D_{11} & 0 \\ 0 & D_{22} \\ D_{31} & D_{32} \end{bmatrix} \tag{2.20}$$

where $\nabla g(\theta^0, \pi^0)$ denotes the matrix of derivatives of $g(\theta, \pi)$ with respect to $[\theta', \mathrm{vec}(\pi)']'$ evaluated at the true population parameters and where

$$\begin{aligned} D_{11} &= -(\tilde{s}_{it}\dddot{z}_{it}'\ddot{x}_{it}) & D_{22} &= -\tilde{s}_{it}(\dddot{z}_{it}'\ddot{z}_{it} \otimes e_1, \ldots, \dddot{z}_{it}'\ddot{z}_{it} \otimes e_{2k+1}) \\ D_{31} &= -(1 - \tilde{s}_{it})\dddot{z}_{it}'\ddot{z}_{it}\pi^0 & D_{32} &= -(1 - \tilde{s}_{it})\beta^{0'} \otimes \dot{z}_{it}'\dot{z}_{it} \end{aligned} \tag{2.21}$$

In this case e_j denotes a row vector of zeros of dimension $[2k + 1]$ with the j-th element being equal to one. In order to get identification, I assume that $\mathrm{rank}(D) = 2(2k + 1)(k_2 + l + 1)$ and is of dimension $[2(k_2 + l) + 1](2k + 1) \times 2(2k + 1)(k_2 + l + 1)$ so that is has full column rank.

Now given the spatial panel structure being considered in this paper, we need to impose some regularity conditions and assumptions on the variables of the model. In particular I will assume that the conditions specified in Nazgul and Prucha (2009) for non-stationary random fields are satisfied, however, in this paper I am going only to focus only on those that are more relevant for the empirical researcher.

Assumption 3

The lattice where the units are located is infinitely countable and there is a distance measure $\rho(\cdot,\cdot)$ and a distance $\rho_0 > 0$ available to the researcher such that $\rho(i, j) \geq \rho_0$ for any two pair of observations i and j.

Assumption 4

The random field is α-mixing satisfying the properties outlined by Nazgul and Prucha (2009).

In practical terms, this means that the degree of dependence between the observations decays as the distance between them increases[8]. From a modeling perspective, this implies that the weights specified in W, the weighting matrix capturing the spillover effects, for any two observations i and j have to decrease as $\rho(i, j) \to \infty$. Note that this assumption also applies to the selection random variables s_{it} so that the missingness of one unit will not affect the availability of observations that are at a large distance from it. The following assumption is related to the error terms.

Assumption 5

At each time period, the $N \times 1$ vectors of errors are generated as:

$$u = F_t \varepsilon \qquad r = M_t \eta \tag{2.22}$$

where the ε and η are a $N \times 1$ vectors of i.i.d. random variables with mean 0, variance of 1, independent of each other and $\mathbb{E}(|\varepsilon|^q) < \infty$ and $\mathbb{E}(|\eta|^q) < \infty$ for $q \geq 4$ and the F_t and M_t are $N \times N$ non-singular unknown matrices whose row and column sums are uniformly bounded.

This assumption allows for many structures of spatial correlation between the error terms without imposing any restrictions on the time dimension. Assumption 6 states that all the relevant matrices are well behaved.

Assumption 6

The matrix of exogenous variables, $\ddot{z}$, has full column rank and its elements are uniformly bounded in absolute value by the finite constant $0 < c_Z < 0$. For a fixed and finite T, the matrices:

[8] Recall that we are working with $N \to \infty$ and fixed T asymptotics so there is not need to impose a weak dependence restriction on the time dimension.

1. $\lim\limits_{N\to\infty} (NT)^{-1}\ddot{z}'\ddot{z} = Q_{zz}$.

2. $\lim\limits_{N\to\infty} (NT)^{-1}\ddot{z}'RR'\ddot{z} = Q_{zRRz}$.

3. $\operatorname*{plim}\limits_{N\to\infty} (NT)^{-1}\ddot{z}'\ddot{z} = Q_{zz}$.

are finite and non-singular[9]. Furthermore, the matrix $\operatorname*{plim}\limits_{N\to\infty} (NT)^{-1}\ddot{z}'\ddot{a} = Q_{za}$ has full column rank $2k+1$. Similarly, the diagonal elements of W are zero and all of its elements are uniformly bounded by a finite constant $0 < c_W < \infty$.

Having state these conditions, the asymptotic normality is summarized in the following proposition.

Proposition 1. *Under Assumptions 2-6,*

$$\sqrt{NT}\left[\left(\hat{\theta}', vec(\hat{\pi})'\right)' - \left(\theta^{0\prime}, vec(\pi^0)'\right)'\right] \xrightarrow{d} N\left[0, \left(D'V^{-1}D\right)^{-1}\right]$$

Furthermore,

$$NT\bar{g}(\hat{\theta},\hat{\pi})'\hat{C}^{-1}\bar{g}(\hat{\theta},\hat{\pi}) \xrightarrow{d} \chi^2_{2(2l+k_2-k_1)+1}$$

This result follows directly from the Uniform Law of Large Numbers and the Central Limit Theorem derived by Nazgul and Prucha (2009) and therefore I omit the proof. This chi-square statistic is useful to determine if the overidentification restrictions (i.e. the moment conditions in Assumption 1 evaluated at the true population parameters) hold. More specifically, this test can help to determine if the mechanism that generated the missing observations is responsible for the violation of Assumption 1, however, it might not be useful in determining if the model is misspeficied (Rai, 2023).

2.5 Correlated Random Effects

2.5.1 The Mundlak Device

When working panel data, researchers usually have to decide between two main estimators, Random Effects (RE) and Fixed Effects (FE). The former provides efficiency gains over the second,

[9] Formally these conditions should also hold for the variables with incomplete observations.

while the later is more robust to violations of one of the main assumptions of the RE estimator, namely that the exogenous variables are uncorrelated with the individual heterogeneity, since the FE approach leaves this relationship unrestricted. Mundlak (1978) proposed a middle ground between these by restricting the relationship with a particular functional form, which falls under the Correlated Random Effects (CRE) approach. Consider the following standard linear model without spatial effects:

$$y_{it} = x_{it}\beta + c_i + u_{it} \tag{2.23}$$

Mundlak's approach is to to model the individual effects c_i as a linear function of the time averages of the covariates:

$$c_i = \bar{x}_i\delta + h_i \tag{2.24}$$

where h_i is uncorrelated with $\bar{x_i}$. By replacing (2.24) in (2.23) we obtain:

$$y_{it} = x_{it}\beta + \bar{x}_i\delta + h_i + u_{it} = x_{it}\beta + \bar{x}_i\delta + r_{it} \tag{2.25}$$

It turns out that if (2.25) is estimated either by POLS or RE, the estimated coefficient of β will be numerically the *same* as if the FE estimator is used in (2.23), a result attributed to Mundlak (1978). This equivalence has been extended to other contexts: Joshi and Wooldridge (2019) proved it for the case of unbalanced panels, Wooldridge (2019) showed it for models with unbalanced panels and endogenous variables. In the spatial context, Debarsy (2012) was the first to introduce the Mundlak device, while Li and Yang (2020) discuss some conditions under which the equivalence holds. Wang and Lee (2013) discuss how to implement the Mundlak device on spatial panels with missing data on the dependent variable, however they do not show the equivalence. In this paper, I show the equivalence holds for models with missing observations on the endogenous covariates in a spatial panel. To this end, consider again the model (2.1):

$$\begin{aligned} y_{it} &= x_{1it}\beta_1 + x_{2it}\beta_2 + W_i X_{1t}\gamma_1 + W_i X_{2t}\gamma_2 + c_i + u_{it} \qquad (2.1) \\ &= a_{it}\theta + c_i + u_{it}, \; i = 1 \dots N, t = 1 \dots T. \end{aligned}$$

where the same definitions and conditions described earlier still apply, including the availability of a set of instruments $(z_{2it} \quad \mathfrak{z}_{2it})$. Consider also the same selection variables and in particular, the complete cases selection variable $\tilde{s}_{it}$ as defined in (2.10). In this context, the Mundlak approach involves modeling the heterogeneity as a function of all the time averages of *all* the exogenous variables $z_{it} = (x_{1it}\ z_{2it}\ w_{1it}\ \mathfrak{z}_{2it})$ in (2.1):

$$c_i = \bar{z}_i\delta + \eta_i \tag{2.26}$$

and multiply the equation by the complete cases selection variable to obtain:

$$\tilde{s}_{it}y_{it} = \tilde{s}_{it}a_{it}\theta + \tilde{s}_{it}\bar{z}_i\delta + \tilde{s}_{it}\tilde{u}_{it} \tag{2.27}$$

where $\tilde{u}_{it} = u_{it} + \eta_i$. Then we can recover the FE estimates of θ by applying Pooled 2SLS to (2.27) using the instruments $\tilde{s}_{it}(z_{2it}\ \ \mathfrak{z}_{2it})$.This result is summarized in the following proposition:

Proposition 2. *Suppose $\tilde{\theta}$ is the estimated coefficient of θ by applying Pooled 2SLS to equation* (2.27). *Then $\tilde{\theta} = \hat{\theta}_{CFE2SLS}$, the coefficient defined in* (2.11).

The proof of Proposition 2 can be found in the Appendix. One of the advantages of the Mundlak device over the FE estimator is that it allows to estimate the effects of variables that do not show variation over time. Note however that as with the FE estimator, this approach is using only using the complete cases so the researcher can obtain efficiency gains with a GMM estimator that uses the information contained in the incomplete cases. The following subsection describes this procedure.

2.5.2 A GMM approach to CRE with missing data

Instead of applying the within transformation to recover the FE estimates of θ, in this section we construct moment conditions using the Mundlak approach, for which I will use equations (2.3), (2.4) and (2.5). As a first step, we model c_i in (2.3) as:

$$\begin{aligned} c_i &= \bar{x}_{2i}\tilde{\theta}_1 + \bar{z}_{2i}\tilde{\theta}_2 + \bar{w}_{2i}\tilde{\theta}_3 + \bar{\mathfrak{z}}_{2i}\tilde{\theta}_4 + \eta_i \\ &= \bar{z}_i\tilde{\theta}_i + \eta_i \end{aligned} \tag{2.28}$$

where the bar over the variables denotes the time average taken over the observations only where $\tilde{s}_{it} = 1$. Here we impose the following condition:

Assumption 7

$\mathbb{E}(\eta_i|Z,S) = 0$

Plugging (2.28) into (2.3) yields:

$$y_{it} = a_{it}\theta + \bar{z}_i\tilde{\theta} + \tilde{u}_{it} \tag{2.29}$$

where $\tilde{u}_{it} = u_{it} + \eta_i$. Since the main model has been augmented with these additional set of variables, the first stage equations need to be adjusted to include these exogenous variables. In particular, letting $\acute{z} = (z_{it} \;\; \bar{z}_i)$ we now have:

$$\begin{aligned} a_{it} &= z_{it}\tilde{\pi}_1^0 + \bar{z}_i\tilde{\pi}_2^0 + \tilde{r}_{it} \\ &= \acute{z}_{it}\tilde{\pi}^0 + \tilde{r}_{it} \end{aligned} \tag{2.30}$$

where $\mathbb{E}(\acute{z}'\tilde{r}_{it}) = 0$ holds by definition. Finally we replace (2.30) in (2.29) to obtain a reduced form of y_{it} on the always observed variables $\acute{z}$:

$$\begin{aligned} y_{it} &= z_{it}\tilde{\pi}_1^0\theta + \bar{z}_i(\tilde{\pi}_2^0\theta + \tilde{\theta}) + \tilde{v}_{it} \\ &= z_{it}\mu_1 + \bar{z}_i\mu_2 + \tilde{v}_{it} \\ &= \acute{z}_{it}\mu + \tilde{v}_{it} \end{aligned} \tag{2.31}$$

where $\tilde{v}_{it} = \tilde{u}_{it} + \tilde{r}_{it}\theta$. Note that as a consequence of Assumption 7 and $\mathbb{E}(\acute{z}'\tilde{r}_{it}) = 0$, $\mathbb{E}(\acute{z}_{it}'\tilde{v}_{it}) = 0$. If we let $\check{\theta} = (\theta' \;\; \tilde{\theta}')'$, from here we can construct the vector of moment conditions as follows:

$$\tilde{g}_{it}(\check{\theta},\tilde{\pi}) = \begin{bmatrix} \tilde{s}_{it}\acute{z}_{it}'\tilde{u}_{it} \\ \tilde{s}_{it}\acute{z}_{it}' \otimes \tilde{r}_{it}' \\ (1-\tilde{s}_{it})\acute{z}_{it}'\tilde{v}_{it} \end{bmatrix} = \begin{bmatrix} \tilde{s}_{it}\acute{z}_{it}'(y_{it} - a_{it}\theta - \bar{z}_i\tilde{\theta}) \\ \tilde{s}_{it}\acute{z}_{it}' \otimes (a_{it} - \acute{z}_{it}\tilde{\pi})' \\ (1-\tilde{s}_{it})\acute{z}_{it}'(y_{it} - \acute{z}_{it}\mu) \end{bmatrix} = \begin{bmatrix} g_{1it}(\check{\theta},\tilde{\pi}) \\ g_{2it}(\check{\theta},\tilde{\pi}) \\ g_{3it}(\check{\theta},\tilde{\pi}) \end{bmatrix} \tag{2.32}$$

From this point the estimation proceeds as in the previous section, but now we have

$$\tilde{V} \equiv \mathbb{E}\left[\tilde{g}(\check{\theta},\tilde{\pi})\tilde{g}(\check{\theta},\tilde{\pi})'\right] \quad \text{and} \quad \tilde{D} \equiv \mathbb{E}\left[\nabla\tilde{g}(\check{\theta}^0,\tilde{\pi}^0)\right] \tag{2.33}$$

Once again, the efficiency gains from this estimator come from the second and third moment conditions of (2.32) by imposing the same coefficients on both the complete and incomplete sub-populations.

2.6 Simulations

2.6.1 Data generating process

To analyze the performance of the proposed GMM estimator in this paper, I ran a Monte-Carlo study where I compared it to the complete cases (CC) estimator, the dummy variable method (DVM) and the estimator that used the data set without any missing observations. Note that although the DVM has been shown to deliver biased results (Jones, 1996), in some simulation studies its performance has been somewhat acceptable, like in Rai (2023). To this end, the benchmark data generating process is as follows:

$$y_{it} = \beta_0 + x_{1it}\beta_1 + W_i X_{1t}\beta_2 + x_{2it}\beta_3 + W_i X_{2t}\beta_4 + c_i + u_{it} \tag{2.34}$$

where $(x_{1it} \quad x_{2it})$ are scalars and the latter is potentially missing and it is endogenous so that it is correlated with the idiosyncratic error term u_{it}. I also generate the variable z_{2it} that will serve to instrument for x_{2it}. Naturally $W_i Z_{2t}$ will serve as an instrument for $W_i X_{2t}$. The individual heterogeneity is correlated with $(x_{1it} \quad z_{2it})$. The observations live in a regular square grid and the weighting matrix that captures the spillover effects follows a rook type scheme. The variables x_{1it}, z_{2it}, u_{it} follow a standard normal distribution and are independent of each other. The population parameter values are $\beta_0 = 2, \beta_1 = 1.5, \beta_2 = 0.7, \beta_3 = 1.2$ and $\beta_4 = 0.4$. The sample size was $N = 900$, $T = 5$ and the number of Monte-Carlo repetitions was 1000 for each scenario described below.

To incorporate the missing data, I used three different mechanisms. In the first one, the data is missing completely at random (MCAR) for which the selection variable followed a binomial distribution with parameter $p = 0.85$. Under this scheme, the average proportion of observations across the 1000 simulated data sets with a complete "own" set of data was 85% as expected. However, the percentage of units with a complete information set (i.e. both the "own" and neighbors information is non-missing) dropped down to 53%. In the second design, the data is missing at random (MAR) so that the selection variable is allowed to depend on the always observed variables. In this instance, I allowed the missingness to depend on x_{1it} and designed it so that around 85% of the observations had their own x_{2it} available. In this case the average proportion across the 1000 repetitions of complete cases was around 51%. As an extension of the first design, the data is again MCAR but the error term follows a spatial autoregressive process of order one (SAR) with a parameter $\rho = 0.4$ and I repeated this design with a smaller sample size where $N = 400$ and $T = 5$ for a total of 2000 observations. Finally, in the third experiment I allow the data to be MAR again but in this case the missingness also depends on the individual heterogeneity.

2.6.2 Results

The simulations showed that the proposed GMM behaves well in finite samples and consistently across the different designs. For example, Table 2.1 shows the average bias, standard deviation and root mean squared error for β_3 and β_4, the coefficients associated with the endogenous and potentially missing variables x_{2it} and $W_i X_{2t}$ for the case when the data is MCAR across the 1000 repetitions. The proposed GMM has an average bias just as small as the estimator that uses the complete data, showing that it is indeed a consistent estimator.

Table 2.1 Average bias, standard deviation and root mean squared error for β_3 and β_4 across the 1000 repetitions when the data is MCAR.

		β_3			β_4	
	Bias	S.D.	RMSE	Bias	S.D.	RMSE
Whole data	0.0004	0.0247	0.0247	0.0010	0.0480	0.0480
Complete cases	-0.0016	0.0364	0.0364	0.0013	0.0713	0.0713
Proposed GMM	0.0004	0.0308	0.0308	0.0008	0.0604	0.0603
Dummy variable	0.9800	0.1304	0.9886	0.3211	0.1865	0.3713

More importantly, the standard deviation of the estimated coefficients for the proposed GMM is smaller than the estimator that uses the complete cases only, showing that it provides some efficiency gains relative to it. As expected, it is not as efficient as the estimator that uses the whole data set as this one uses the full set of available information, whereas the proposed estimator might lose some information (e.g. cases where there is only one incomplete time period and therefore the unit becomes uninformative). This is illustrated in Figure 2.2, where the estimator that uses the whole data set has tighter distributions around the true population values, followed by the proposed GMM estimator and finally, the complete cases estimator, which has more disperse distributions.

The simulations also show that the DVM estimator is inconsistent as the average bias for the parameters associated with the endogenous variables is substantial, although this appears to be limited to these covariates as the β_1 and β_2 coefficients seem to be well behaved.

The simulations show a very small loss in efficiency when the data is MAR compared to the MCAR case. Similarly, this loss is also small when the data is MCAR but the error term follows a SAR(1) process, as in the latter the standard deviations are slightly larger relative to the the first two scenarios. Nevertheless, the proposed GMM estimator shows again to be more efficient than the complete cases estimator. Of course, if the researcher is confident that the error terms follows a SAR(1), she might be able to exploit efficiency gains using alternative estimators that use this information such as maximum likelihood, at the risk of misspecifying the structure of data generating process. As expected, when the sample size is smaller the distribution of the coefficients show a greater dispersion, but the proposed GMM estimator continues to show to be well behaved under this scenario with a small bias and a standard deviation that is smaller than the complete cases estimator. Finally, when the missingness is also allowed to depend on the individual heterogeneity, there are no substantial differences in the results: the proposed GMM seems to be consistent and the root mean squared error is between the estimator that uses the whole data set and the complete cases one. This result is somewhat expected as the within transformation removes the individual heterogeneity from the estimating equations.

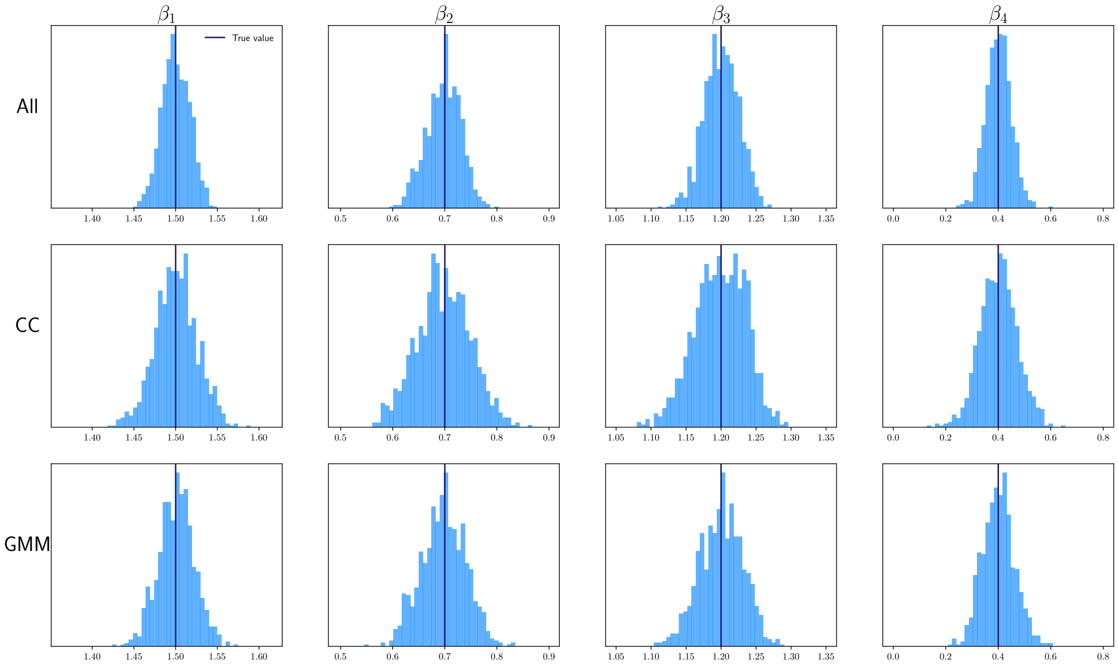


Figure 2.2 Distribution of estimated coefficients across the 1,000 Monte-Carlo repetitions when the data is MCAR.

2.7 Empirical Application

I this section, I revisit the problem of analyzing the impact of different variables on crime in the state of North Carolina at the county level between 1981 and 1987. This problem was studied by Cornwell and Trumbull (1994) and by B. Baltagi (2006), where they modeled the crime rate as a function of a set of covariates that included deterrent variables and returns to legal opportunities. However, as pointed out by B. Baltagi (2006), most of the fixed effects estimates presented by Cornwell and Trumbull (1994) turned out to be statistically insignificant, therefore, in this paper I present a simplified version of the model that focuses on the deterrent variables.

The original data set used in their estimation contained 90 counties[10] and seven time periods for a total of 630 observations. Note that their data has no missing observations and therefore, for the purpose of this illustration, the missing variables will be generated artificially so that around 5%

[10] North Carolina has a total of 100 counties, nevertheless their data only contained information for 90.

of the observations has one of their variables missing. To this end, consider the following model:

$$\begin{aligned}\text{crime}_{it} &= \beta_0 + \beta_1\text{arrest}_{it} + \beta_2\text{conviction}_{it} + \beta_3\ \text{prison}_{it} \\ &\quad + \beta_4\text{police}_{it} + \beta_5\text{avgsent}_{it} + \beta_6\text{dens}_{it} + c_i + u_{it} \qquad (2.35)\end{aligned}$$

where crime is the crime rate (crimes committed per person), arrest is the “probability” of arrest (number of arrests per crimes), conviction is the proportion of convictions to arrests, prison is the ratio of sentences that results in jail time to the total number of convictions, police is the police per capita, avgsent is the average sentence in days and dens is the the number of people living in the county per square mile. Cornwell and Trumbull (1994) argued that both the arrest and police variables are endogenous, for which they proposed two external instrumental variables (IV): the tax revenue per capita is correlated with the police covariate as we would expect that higher tax revenues are correlated with larger police forces. On the other hand, the ratio of crimes that involve face to face contact to those that do not (denoted by mix) is the IV for arrests, the rationale being that when a crime is committed in person, identification of the perpetrator is facilitated.

Columns 1 and 2 of Table 2.2[11] show the results of estimating the model 2.35 using the complete data set by FE2SLS and using the proposed GMM (PGMM) estimator respectively[12]. The first thing to note is that all the coefficients are similar in magnitude and most of them have the expected sign. Indeed, deterrent variables such as arrests, conviction and prison have a negative effect on the crime rate. On the other hand, police and density have a positive impact on the dependent variable, which is expected as the latter increases the likelihood of offenders finding victims. On the other hand, as pointed out by B. Baltagi (2006), there might be simultaneity involved in the relationship between the crime rate, arrests and police. Note however that none of the estimates is statistically significant.

From a spatial perspective, one could argue that criminal activity in some areas might affect the surrounding counties. For example, if the people living in the big cities of North Carolina are

[11] All the specifications include time dummies.

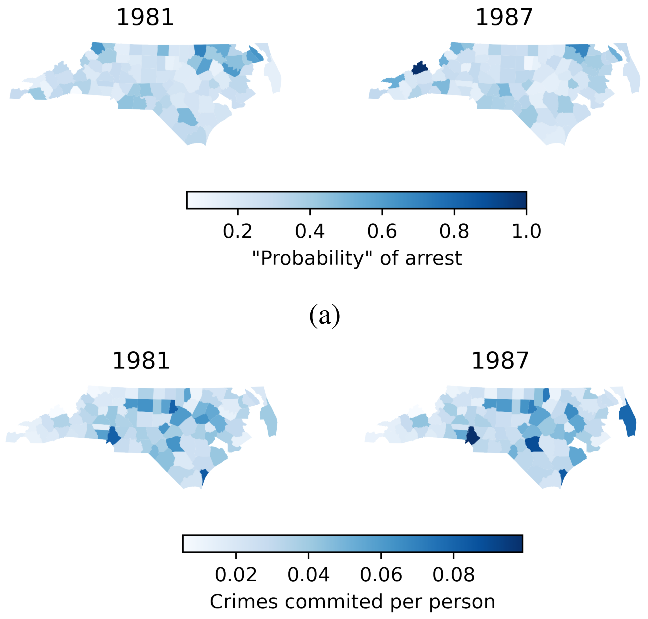
[12] Admittedly the results from the missing data case are going to depend on which observations are missing, therefore I estimated the model 200 times and at each iteration, a different set of observations was missing. The table shows the average estimated coefficients across the 200 repetitions and the “standard errors” presented for the PGMM columns are the sample standard deviation of the computed coefficients across the iterations.

more affluent and are more densely populated than those in rural areas, it could be expected that the former have larger crime rates or arrests. Figures 2.3a and 2.3b show these variables plotted over the counties at the beginning and end of the period of analysis. They reflect that indeed counties with high (low) proportion of arrests are neighbors to other counties with higher (lower) "probabilities" of arrest. To capture this, I augment model 2.35 by including the spatial lag of the variable arrest. The results of this are shown in columns 3 and 4 of Table 2.2.

Table 2.2 Results from the estimation (standard errors in parenthesis)

	FE2SLS	PGMM	FE2SLS	PGMM
Arrest	-0.0202 (0.0128)	-0.0182 (0.0169)	-0.0224 (0.0242)	-0.0175 (0.0162)
Police	3.7286 (1.7727)	4.1822 (2.145)	4.0688 (4.0708)	3.9161 (2.6127)
Conviction	-0.0019 (0.0009)	-0.0023 (0.0013)	-0.0206 (0.0169)	-0.0021 (0.0015)
Prison	-0.0012 (0.0045)	-0.0023 (0.0049)	-0.0020 (0.0021)	-0.0018 (0.0054)
Sentence	0.0002 (0.0002)	0.0004 (0.0002)	-0.0012 (0.0072)	0.0004 (0.0003)
Density	0.0039 (0.0049)	0.0011 (0.0027)	0.0002 (0.0004)	0.0006 (0.0028)
W × Arrest	-	-	0.0046 (0.0063)	-0.0151 (0.0533)

After adding the additional covariate, none of the estimates from the estimates for the other variables changes significantly and they remain statistically insignificant at the usual confidence levels. The sign of the coefficient for the spatial lag of arrests is positive for the case of the FE2SLS estimator but negative for the PGMM. One could argue that the expected sign of this variable would be positive because if the number of arrests in counties that are neighbors of *i* increases, the criminals might move their activities *i*. However the empirical evidence does not support this theory, as both estimators find a statistically insignificant coefficient, which coincides with the findings of Cornwell and Trumbull (1994).



(a)

(b)

Figure 2.3 Maps of the “probability” of arrest (a) and crime rates (b) at the beginning and end of the period of study.

2.8 Conclusion

Missing data is a more serious problem in spatial models with spillover effects because the loss of information is greater if the researcher decides to use only the complete cases. This paper presented a simple way to exploit the information of incomplete observations in spatial panel data models with potentially missing endogenous explanatory variables. The estimator is presented in a GMM framework that imposes restrictions on the coefficients in the complete and incomplete subsamples to obtain a more efficient estimator relative to the fixed effects estimator that only uses the complete cases.

An alternative to the FE estimator in panel data is the correlated random effects approach, which restricts the relationship between the unobserved heterogeneity and the explanatory variables. In particular, by using Mundlak’s device the researcher can recover the same numerical FE coefficients for the time varying variables and also estimate the effects of time invariant covariates. In this paper, I show that this equivalence carries over to the missing data with endogenous independent variables. In addition to this equivalence, I also present a potentially more efficient GMM estimator that exploits the incomplete cases information using the the additional restrictions of the Mundlak approach.

The simulations show that the proposed GMM estimator behaves well in finite samples with an average bias very close to the estimator that uses the whole set of non missing data but more importantly, it consistently had a smaller standard deviation across the Monte-Carlo study compared to the estimator that uses only the complete cases, which shows that the GMM indeed provides some efficiency gains.

CHAPTER 3

ESTIMATION OF MODELS WITH MULTIPLE FIXED EFFECTS AND ENDOGENOUS VARIABLES: A CORRELATED RANDOM EFFECTS APPROACH

3.1 Introduction

Gravity type models have been widely used in a variety of economic fields to analyze the flows of goods or services between multiple regions or entities. The international trade literature has had a long tradition of using this type of model to quantify the relationship between bilateral trade flows and other variables such as trade costs and economic integration agreements (Baier et al., 2014), although its use to estimate these relationships can be documented back to 1885 (Kabir et al., 2017). Studies in this area that use the gravity equation include Flach and Unger (2022), Anderson and Van Wincoop (2003) and B. H. Baltagi et al. (2003), but the list of papers is extensive. Furthermore, gravity type models have also been used to explain migration flows (Beine et al., 2015) and international financial assets outflows (Okawa & Van Wincoop, 2012). Kabir et al. (2017) provides an excellent overview of other areas where the gravity equation has been applied. However, for the remainder of the paper I will focus on the international trade case.

The main idea behind gravity models is that the bilateral economic relationship between two entities is proportional with their economic size (e.g. a country's GDP is often used in the trade literature (Matyas, 1997)) and negatively correlated with their economic or geographical distance. Intuitively this idea is appealing and is analogous to Newton's Universal Gravity Law, however, it was recognized that the inclusion of covariates such as policy variables (e.g. border taxes) lacked theoretical justification (Anderson, 1979). However, Anderson (1979) made a seminal contribution in this direction by presenting a commodities model that are differentiated by the country of origin and deriving a gravity equation from it. Other papers that also presented theoretical foundations for these models include Krugman (1980) Bergstrand (1985), Eaton and Kortum (2002) and Chaney (2018).

Gravity type models are at least double indexed: in the cross sectional case, one index corresponds to the originating country and the other to the destination country. If time series data is available, then one of the indices identifies the time dimension instead of the originating country and if the researcher is using panel data, then a third index can be added to the model to identify each of the components previously mentioned. More details about the formulation of a gravity model can be found in Matyas (1997). Although more details will be provided later in the paper, each of these dimensions will have a corresponding term (unobserved heterogeneities, latent variables or "fixed effects") in the model that captures their corresponding effect on the response variable. Depending on the assumptions imposed on these terms, the estimation approach can vary between a random effects (RE) procedure or a fixed effects (FE) estimator (Matyas, 1997). An excellent overview of both the RE and FE with multi-dimensional panels can be found in Matyas (2017, Chapters 1 and 2).

As previously mentioned, one of the main differences between the FE and RE estimators is the restriction related to the relationships between the explanatory variables and the unobserved heterogeneities that is imposed to achieve consistency. In particular, the FE allows for arbitrary correlation between the latent variables and the covariates, while the RE assumes zero correlation among the dependent variables and each of the fixed effects. However, the literature has proposed a middle ground between these approaches, the correlated random effects (CRE). For instance, in the one way panel case, Mundlak (1978) suggested to model the individual heterogeneity as a linear function of the time averages of the right hand side variables, use this auxiliary equation in the main model and estimate the parameters with Pooled Ordinary Least Squares (POLS). By following these steps, he showed that the researcher can obtain same numerical estimates of the FE estimator for the time varying covariates. It is important to note that this result is an algebraic equivalence that does *not* depend on the statistical properties of the estimators nor the conditions assumed to obtain a consistent estimator laid out earlier.

This equivalence between the Mundlak device and the FE estimator has been extended to other contexts. For example, Wooldridge (2021) showed it for the case of a two-way panel, Debarsy (2012) was the first to propose it for spatial panels, Joshi and Wooldridge (2019) demonstrated it for the case of unbalanced panels and Yang (2022) proved it for models with multiple fixed effects.

It is important to note that Yang (2022) does not allow for correlation between the covariates and the idiosyncratic error term. In this paper, I extend the result by relaxing this assumption and show that the FE estimates and be recovered using two different sets of variables to model the fixed effects. The rest of the paper is organized as follows. Section 3.2 presents the model and its assumptions. Section 3.3 shows how to consistently estimate the model, while Section 3.4 introduces the equivalence between the FE and the CRE approach and Section 3.5 concludes.

3.2 Model

To motivate the use of a FE or RE approach, consider the following linear model with additive heterogeneities, which is common to see in gravity-type models:

$$\begin{aligned} y_{ijt} &= x_{1ijt}\beta_1 + x_{2ijt}\beta_2 + \alpha_i + \phi_j + \gamma_t + u_{ijt} \\ &= x_{ijt}\beta + e_{ijt}, \quad i = 1 \dots N_1, j = 1, \dots N_2, t = 1 \dots T \end{aligned} \tag{3.1}$$

where y_{ijt} is the dependent variable, x_{ijt} is a vector of K explanatory variables, including a constant. I decompose the error term e_{ijt} into four components: α_i is the individual specific heterogeneity along one of the dimensions of the data (e.g. exporter "fixed effect"), ϕ_j is the heterogeneity along the other dimension (e.g. importer "fixed effect"), γ_t is the time specific effect and u_{ijt} is the idiosyncratic error term. I divide x_{ijt} in two subsets: x_{1ijt} are K_1 exogenous variables in the sense that $\mathbb{E}(x'_{1ijt}u_{ijt}) = 0$ and x_{2ijt} are K_2 endogenous variables so that $\mathbb{E}(x'_{2ijt}u_{ijt}) \neq 0$. In light of the endogenous x_{2ijt}, to obtain consistent estimates of β, we could construct Hausman-Taylor type instrumental variables, however I will assume that we have L (with $L \geq K_2$) external instrumental variables available and denoted by z_{2ijt} that satisfy the usual relevance $[\mathbb{E}(z'_{2ijt}x_{2ijt}) \neq 0]$ and exogeneity $[\mathbb{E}(z'_{2ijt}u_{ijt}) = 0]$ conditions and let the set of exogenous variables be $z_{ijt} = (x_{1ijt} \;\; z_{2ijt})$.

In this paper, I do not consider formal asymptotic analysis, nor do I focus on whether the individual heterogeneities and time effects are parameters to be estimated or should be treated as random variables since the equivalence derived below using the Mundlak approach is an algebraic result. However, at least one of the indices should go to infinity to obtain a consistent estimator of β, conditional on not treating the heterogeneities or time effects associated with that index as parameters to be estimated to avoid the incidental parameters problem. Matyas (2017) has a nice review of asymptotic properties of fixed effects and random effects estimators for the different cases that can arise in empirical work.

Throughout the paper I assume that the data is ordered such that the i index is the slowest to change, then j and t is the fastest. I also assume that all the relevant matrices have full column rank and are therefore invertible. I also maintain the following exogeneity assumption:

$$\mathbb{E}\left(u_{ijt}|z_{111}, z_{112}, \ldots z_{N_1N_2T}, \alpha_i, \phi_j, \gamma_t\right) = 0 \tag{3.2}$$

This is an extension to the three dimensional panel of the strict exogeneity assumption found in the one way panel data literature. Note that the equivalence that will be presented in Section 3.4 does not depend on any of the assumptions stated so far, it is an algebraic equivalence that is unrelated to the statistical properties of the estimators presented in the next section.

3.3 Estimation

The estimation approach of the parameters in equation (3.1) will depend on the variables of interest and the assumptions the researcher is willing to make. If we assume that the exogenous variables z_{2ijt} are uncorrelated to all the individual heterogeneities (α_i, ϕ_j and γ_t) and the following conditions are met:

1. The heterogeneities are pairwise uncorrelated.

2. $\mathbb{E}(\alpha_i) = \mathbb{E}(\phi_j) = \mathbb{E}(\gamma_t) = 0$.

Furthermore, if we assume

$$\mathbb{E}(\alpha_i\alpha_{i'}) = \begin{cases} \sigma_\alpha^2 \text{ if } i = i' \\ 0 \text{ otherwise} \end{cases}$$

$$\mathbb{E}(\phi_j\phi_{j'}) = \begin{cases} \sigma_\phi^2 \text{ if } j = j' \\ 0 \text{ otherwise} \end{cases}$$

$$\mathbb{E}(\gamma_t\gamma_{t'}) = \begin{cases} \sigma_\gamma^2 \text{ if } t = t' \\ 0 \text{ otherwise} \end{cases}$$

then the structure of the covariance matrix is given by

$$
\begin{aligned}
\mathbb{E}(e_{ijt}e'_{i'j't'}) &= \mathbb{E}\left[(\alpha_i + \phi_j + \gamma_t + u_{ijt})(\alpha_{i'} + \phi_{j'} + \gamma_{t'} + u_{i'j't'})'\right] \\
&= \sigma^2_\alpha && \text{if } i = i', j \neq j', t \neq t' \\
&= \sigma^2_\phi && \text{if } i \neq i', j = j', t \neq t' \\
&= \sigma^2_\gamma && \text{if } i \neq i', j \neq j', t = t' \\
&= \sigma^2_\alpha + \sigma^2_\phi && \text{if } i = i', j = j', t \neq t' \\
&= \sigma^2_\alpha + \sigma^2_\gamma && \text{if } i = i', j \neq j', t = t' \\
&= \sigma^2_\phi + \sigma^2_\gamma && \text{if } i \neq i', j = j', t = t' \\
&= \sigma^2_\alpha + \sigma^2_\phi + \sigma^2_\gamma + \sigma^2_u && \text{if } i = i', j = j', t = t'
\end{aligned}
$$

Which translates into the following matrix:

$$\Omega = \mathbb{E}(ee') = \sigma^2_\alpha(\mathbf{I}_{N_1} \otimes J_{N_2T}) + \sigma^2_\phi(J_{N_1} \otimes \mathbf{I}_{N_2} \otimes J_T) + \sigma^2_\gamma(J_{N_1N_2} \otimes \mathbf{I}_T) + \sigma^2_u\mathbf{I}_{N_1N_2T} \tag{3.3}$$

where ⊗ represents the kronecker product and $\mathbf{I}$ and J denote an identity matrix and a square matrix of ones, respectively, of size given by their subscript. We can transform the data to obtain an efficient estimator that exploits this information. Indeed, the RE estimator presented in Matyas (2017) can be obtained by applying Pooled Two Stage Least Squares (P2SLS) to the following equation:

$$\Omega^{-\frac{1}{2}}y = \Omega^{-\frac{1}{2}}X\beta + \Omega^{-\frac{1}{2}}e \tag{3.4}$$

using the instruments $\Omega^{-\frac{1}{2}}z_2$, where the absence of subscripts indicate that the data has been stacked. Denote the estimated coefficient from this estimation as $\hat{\beta}_{RE2SLS}$. A few observations are in order. First, the assumptions stated above related to the second moments of the individual heterogeneities are *not* necessary to get a consistent estimator of the parameters. These conditions only determine the specific structure of the matrix Ω which in turn is used to perform the GLS-type transformation of the data to get efficiency gains, but the consistency of the estimator hinges on other assumptions. One the other hand, the second moment conditions are important to get a particular structure of the covariance matrix, but if these do not hold, inference can be misleading. For this reason, researchers should use a robust covariance matrix to obtain the associated standard errors.

Sometimes imposing a zero correlation between the exogenous variables and the heterogeneities might be an unrealistic restriction. In these instances, a FE approach is also available and it has the advantage of leaving the relationship between the exogenous variables and the heterogeneities unrestricted. One way to obtain the FE2SLS estimator is to include dummy variables to account for the different heterogeneities (see Wooldridge (2021) for a description of the Two-Way Fixed Effects estimator). Alternatively, we can apply a transformation to the data to end up with an estimating equation that does not contain the "fixed effects". To this end, we define the following notation. Let

$$\bar{y}_{i\cdot\cdot} = \frac{1}{N_2 T} \sum_{j}^{N_2} \sum_{t}^{T} y_{ijt} \quad \text{and} \quad \bar{y}_{\cdot j \cdot} = \frac{1}{N_1 T} \sum_{i}^{N_1} \sum_{t}^{T} y_{ijt} \tag{3.5}$$

be the unit specific averages over the remaining dimensions for variable y. Also define

$$\bar{y}_{\cdot\cdot t} = \frac{1}{N_1 N_2} \sum_{i}^{N_1} \sum_{j}^{N_2} y_{ijt} \tag{3.6}$$

be the cross sectional average for each t. Let

$$\bar{y}_{\cdot\cdot\cdot} = \frac{1}{N_1 N_2 T} \sum_{i}^{N_1} \sum_{j}^{N_2} \sum_{t}^{T} y_{ijt} \tag{3.7}$$

be the overall average. Note that

$$\bar{y}_{\cdot\cdot\cdot} = \frac{1}{N_1} \sum_{i}^{N_1} \bar{y}_{i\cdot\cdot} = \frac{1}{N_2} \sum_{i}^{N_2} \bar{y}_{\cdot j \cdot} = \frac{1}{T} \sum_{t}^{T} \bar{y}_{\cdot\cdot t} \tag{3.8}$$

Finally, transform and denote the original data as follows:

$$\ddot{y}_{ijt} = y_{ijt} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j \cdot} - \bar{y}_{\cdot\cdot t} + 2\bar{y}_{\cdot\cdot\cdot} \tag{3.9}$$

and other variables can be constructed similarly. This transformation gives rise to the within estimator and it will remove the heterogeneity and time effects. It was first introduced by Matyas (1997) and its extension to a model with endogenous variables is straightforward. Indeed, the FE estimator of β, denoted as $\hat{\beta}_{FE2SLS}$ can be obtained by applying P2SLS to:

$$\ddot{y}_{ijt} = \ddot{x}_{ijt}\beta + \ddot{e}_{ijt} = \ddot{x}_{ijt}\beta + \ddot{u}_{ijt} \tag{3.10}$$

using the instruments $\ddot{z}_{2ijt}$. A few comments are in order related to the within estimator. First, this is not the only transformation that removes the individual heterogeneities from the estimating equation. As pointed out by Balazsi et al. (2018), the following operation would also remove the heterogeneities and time effects in equation (3.1):

$$\dot{y}_{ijt} = y_{ijt} - \bar{y}_{ij\cdot} - \bar{y}_{\cdot\cdot t} + \bar{y}_{\cdot\cdot\cdot} \tag{3.11}$$

The transformation needed to remove the "fixed effects" will vary depending on the structure of the heterogeneities.

A second and perhaps more important point related to this operation is that some of the coefficients might not be identifiable as the associated variables will also be removed after the operation. In particular, variables that are either time or individual invariant (in either dimension) will be also removed by the transformation. From an empirical point of view, this is not a trivial issue: for example, in the trade literature and gravity models it is common to include the GDP of the exporter or importer region or some policy variables as covariates, which will be invariant from at least one of the cross sectional dimensions and thus eliminated. This problem and the efficiency gains give more appeal to the RE estimator over the FE, at the cost of imposing additional assumptions. It is essential to stress out that the equivalence presented in the next section is an algebraic result and is not related to other statistical properties of the estimators such as consistency.

3.4 Correlated Random Effects

As noted in the previous section, the FE and RE rely on opposing assumptions related to the relationship between the exogenous variables and the individual and time effects. On the one hand, the RE estimator assumes that there is no correlation between the exogenous and these unobserved effects, while the FE places no restrictions in this sense. As a result, the usual bias-variance trade-off arises between both estimators from the imposition and plausibility of this condition. In one-way panels, the literature has proposed a middle ground in which the dependence between the unobserved heterogeneity and the covariates is not zero but is restricted in a specific way. In particular, Mundlak (1978) proposed to model the individual heterogeneity as linear function of the time average of the covariates. Chamberlain (1982) provided a more flexible approach in which the heteretogeneity is linearly projected into the space of the whole history of explanatory variables. One of the drawbacks of the latter is that the number of coefficients to be estimated grows linearly as the sample size grows, which can be a greater issue in higher dimensional panels.

An interesting fact about the Mundlak device is that he showed that by adding the time averages to the estimating equation, one can recover the same FE estimates if the equation is estimated by RE or POLS. This result has been extended to the two-way panel: Wooldridge (2021) proves that the two-way FE estimates can be recovered by applying POLS to the main equation and adding the time and cross sectional averages as regressors, while B. H. Baltagi (2023) demonstrates that the GLS-type transformation and POLS are equivalent in this sense. In addition, Yang (2022) extends this equivalence to three-way panels and presents conditions under which a weighted variable addition test is equivalent to the Hausman specification test. More concretely, the linear projection of the individual and time effects in the three-way panel using the Mundlak approach is given by:

$$\begin{aligned}
\mathbb{L}(\alpha_i | z_{111}, z_{112}, \dots z_{N_1 N_2 T}) &= \bar{z}_{i\cdot\cdot}\delta_1 \\
\mathbb{L}(\phi_j | z_{111}, z_{112}, \dots z_{N_1 N_2 T}) &= \bar{z}_{\cdot j \cdot}\delta_2 \\
\mathbb{L}(\gamma_t | z_{111}, z_{112}, \dots z_{N_1 N_2 T}) &= \bar{z}_{\cdot\cdot t}\delta_3 \qquad (3.12)
\end{aligned}$$

where $\mathbb{L}(\cdot)$ denotes the linear projection operator. One aspect that these papers have in common is that they show the result for the case in which the explanatory variables are exogenous with respect to the idiosyncratic error term. In this paper, I show that the equivalence carries over when there are endogenous variables on the right hand side of the equation, which can be useful as this situation often arises in empirical work. Once again it is important to stress out that this is an algebraic result that is unrelated to the consistency of the estimators. To fix ideas, it is useful to first re-write in scalar form the RE transformation from (3.4), which yields the following:

$$\sigma_u^2 \Omega^{-\frac{1}{2}} y_{ijt} = \tilde{y}_{ijt} = y_{ijt} - \tilde{\theta}_1 \bar{y}_{i\cdot\cdot} - \tilde{\theta}_2 \bar{y}_{\cdot j \cdot} - \tilde{\theta}_4 \bar{y}_{\cdot\cdot t} + \tilde{\theta}_4 \bar{y}_{\cdot\cdot\cdot} \qquad (3.13)$$

where,

$$
\begin{aligned}
\tilde{\theta}_1 &= \left(1 - \sqrt{\theta_1}\right) \\
\tilde{\theta}_2 &= \left(1 - \sqrt{\theta_2}\right) \\
\tilde{\theta}_3 &= \left(1 - \sqrt{\theta_3}\right) \\
\tilde{\theta}_4 &= \left(2 - \sqrt{\theta_1} - \sqrt{\theta_2} - \sqrt{\theta_3} - \sqrt{\theta_4}\right) \\
\theta_1 &= \frac{\sigma_u^2}{N_2 T \sigma_\alpha^2 + \sigma_u^2} \\
\theta_2 &= \frac{\sigma_u^2}{N_1 T \sigma_\phi^2 + \sigma_u^2} \\
\theta_3 &= \frac{\sigma_u^2}{N_2 T \sigma_\gamma^2 + \sigma_u^2} \\
\theta_4 &= \frac{\sigma_u^2}{N_2 T \sigma_\alpha^2 + N_1 T \sigma_\phi^2 + N_1 N_2 \sigma_\gamma^2 + \sigma_u^2}
\end{aligned}
$$

and where we can transform the rest of the variables in a similar way. Therefore, the RE2SLS estimator can be once again obtained by applying Pooled 2SLS to

$$\tilde{y}_{ijt} = \tilde{x}_{ijt}\beta + \tilde{e}_{ijt} \tag{3.14}$$

using instrumental variables $\tilde{z}_{2ijt}$. Note that the Pooled 2SLS estimator is a special case of (3.14) by setting $\tilde{\theta_s} = 0$ for $s = 1, 2, 3, 4$. To obtain the CRE 2SLS estimator using the Mundlak device, we can apply P2SLS to the following equation:

$$\tilde{y}_{ijt} = \tilde{x}_{ijt}\beta + \tilde{\bar{x}}_{1ijt}\pi + \tilde{\bar{z}}_{2ijt}\delta \tag{3.15}$$

using instruments $\tilde{z}_{2ijt}$, where $\tilde{\bar{x}}_{1ijt} = (\tilde{\bar{x}}_{1i\cdot\cdot} \quad \tilde{\bar{x}}_{1\cdot j\cdot} \quad \tilde{\bar{x}}_{1\cdot\cdot t})$ and $\tilde{\bar{z}}_{2ijt} = (\tilde{\bar{z}}_{2i\cdot\cdot} \quad \tilde{\bar{z}}_{2\cdot j\cdot} \quad \tilde{\bar{z}}_{2\cdot\cdot t})$. Two observations are in order related to (3.15). First note that $\tilde{\bar{x}}_{1i\cdot\cdot} = (1 - \tilde{\theta}_1)\bar{x}_{1i\cdot\cdot}$, $\tilde{\bar{x}}_{1\cdot j\cdot} = (1 - \tilde{\theta}_2)\bar{x}_{1\cdot j\cdot}$, $\tilde{\bar{x}}_{1\cdot\cdot t} = (1 - \tilde{\theta}_3)\bar{x}_{1\cdot\cdot t}$ and similarly for $\tilde{\bar{z}}_2$, so that the averages of the transformed variables do not depend on parameters that are associated with the other dimensions' averages. Second, note that we only need to include the averages of the exogenous variables $\tilde{\bar{x}}_{1ijt}$ and $\tilde{\bar{z}}_{2ijt}$ and not the ones from the endogenous variables $\tilde{\bar{x}}_{2ijt}$. By doing so, the β recovered from this estimation, denoted as $\hat{\beta}_{M_1}$ will be numerically the same as $\hat{\beta}_{FE2SLS}$. This result in summarized in Proposition 1.

Proposition 1. *Suppose that all the relevant matrices have full column rank. Let $\hat{\beta}_{FE2SLS}$ be the coefficient obtained by estimating equation 3.10 by Pooled 2SLS and $\hat{\beta}_{M_1}$ be the coefficient computed from applying Pooled 2SLS to equation* (3.15)*. Then $\hat{\beta}_{FE2SLS} = \hat{\beta}_{M_1}$.*

The proof of this proposition can be found in the Appendix. This result is useful because it allows the researcher to perform a Hausman-type test using a variable addition test. Specifically, the researcher can analyze the significance of the coefficients associated with the averages to decide between a FE or RE specifications. A discussion of this procedure can be found in Joshi and Wooldridge (2019). As Matyas (2017) notes, there can be many causes of endogeneity in three-way panels, which might require at least as many instruments for each of these sources. In order to obtain the CRE equivalence, the researcher has to include all their averages in the estimating equation, which can consume an important number of degrees of freedom and can be costly in finite samples when conducting inference. Fortunately, we can recover the FE estimates by adding a different set of variables. If we let $\hat{x}_{2ijt}$ denote first stage predicted values for the endogenous variables, then applying Pooled 2SLS to

$$\begin{aligned}\tilde{y}_{ijt} &= \tilde{x}_{1ijt}\beta_1 + \hat{\tilde{x}}_{2ijt}\beta_2 + \tilde{\bar{\bar{x}}}_{1ijt}\pi_1 + \hat{\tilde{\bar{\bar{x}}}}_{2ijt}\pi_2 \\ &= \hat{\tilde{x}}_{ijt}\beta + \hat{\tilde{\bar{\bar{x}}}}_{ijt}\pi \qquad (3.16)\end{aligned}$$

using the instruments $(\tilde{z}_{2ijt} \quad \tilde{\bar{\bar{z}}}_{2ijt})$ will also yield the same β as FE2SLS. Proposition 2 formally states the equivalence.

Proposition 2. *Suppose that all the relevant matrices have full column rank. Let $\hat{\beta}_{FE2SLS}$ be the coefficient obtained by estimating equation 3.10 by Pooled 2SLS and $\hat{\beta}_{M_2}$ be the coefficient computed from applying Pooled 2SLS to equation* (3.16)*. Then $\hat{\beta}_{FE2SLS} = \hat{\beta}_{M_2}$.*

As noted previously, the advantage of using this set of variables instead of the instruments is that it allows to preserve the degrees of freedom if we have more than one instrument for each endogenous covariate. An important feature of the CRE approach using the Mundlak device is that it allows to estimate the effect of variables that are constant across one of the dimensions of the panel, something that cannot be done using the within estimator as its transformation wipes out any variable of this nature. In fact, Wooldridge (2021) proves in the two-way panel that adding additional variables that only vary across one of the dimensions will not change the FE estimates, a result that most likely carries over to the three-way panel. This result makes intuitive sense as the within estimator is supposed to remove these variables but it also shows that adding the averages (either from the exogenous variables or from the predicted values as in Proposition 2) is enough to control for the the individual and time effects.

3.5 Conclusion

In this paper, I establish the algebraic equivalence between FE2SLS and RE2SLS in three-way panels with additive unobserved heterogeneities in the presence of endogenous variables using the Mundlak device. Namely, by including either the averages of the exogenous variables or the means of the predicted values explanatory covariates across all the different dimensions, is enough to control for the unobserved heterogeneities and to recover the FE2SLS estimates. The first approach has the disadvantage that if there are multiple instruments available, the degrees of freedom could be reduced considerably, an issue that is more severe in finite samples. The use of Mundlak's device also allows to relax the no correlation between the covariates and the unobserved heterogeneities, which allows the researcher to obtain more robust estimates of the coefficients.

Furthermore, this result also offers the researchers a flexible and easy to implement solution to choose between a FE and RE specification. In particular, Yang (2022) shows that a modified variable addition test associated with the averages is equivalent to the Hausman-type test, with the additional advantage that the former can be made robust to heteroskedasticity and serial correlation. One of the limitations of the result shown in this paper is that the algebraic equivalence is likely to break with other structures of heterogeneities. For example, Yang (2022) argues that if the cross sectional heterogeneities are time varying, then the result no longer holds in the case of exogeneous variables, a result that most like carries over in the presence of endogenous covariates. However, future research might could extend the result to more general models of unobserved heterogeneities.

BIBLIOGRAPHY

Abrevaya, J., & Donald, S. G. (2017). A gmm approach for dealing with missing data on regressors. *The Review of Economics and Statistics*, *99*(4), 657–662.

Ahn, S. C., & Schmidt, P. (1995). Efficient estimation of models for dynamic panel data. *Journal of Econometrics*, *68*(1), 5–27.

Amemiya, T. (1985). *Advanced econometrics*. Harvard University Press.

Anderson, J. (1979). A theoretical foundation for the gravity equation. *The American Economic Review*, *69*(1), 106–116.

Anderson, J., & Van Wincoop, E. (2003). Gravity with gravitas: A solution to the border puzzle. *The American Economic Review*, *93*(1), 170–192.

Arellano, M. (1987). Computing robust standard errors for within-groups estimators. *Oxford Bulletin of Economics and Statistics*, *49*(4), 431–434.

Baier, S. L., Bergstrand, J., & Feng, M. (2014). Economic integration agreements and the margins of international trade. *Journal of International Economics*, *93*(2), 339–350.

Balazsi, L., Matyas, L., & Wansbeek, T. (2018). The estimation of multidimensional fixed effects panel data models. *Econometric Reviews*, *3*(3), 212–227.

Baltagi, B. (2006). Estimating an economic model of crime using panel data from north carolina. *Journal of Applied Econometrics*, *21*, 543–547.

Baltagi, B., & Liu, L. (2011). Instrumental variable estimation of a spatial autoregresive panel model with random effects. *Economic Letters*, (111), 135–137.

Baltagi, B. H. (2023). *The two-way mundlak estimator* (tech. rep. No. Working Paper No. 256). Center for Policy Research.

Baltagi, B. H., Egger, P., & Pfaffermayr, M. (2003). A generalized design for bilateral trade flow models. *Economics Letters*, *80*(3), 391–397.

Basile, R. (2009). Productivity polarization across regions in europethe role of nonlinearities and spatial dependence. *International Regional Science Review*, 92–115.

Basile, R., Durban, M., Minguez, R., Montero, J. M., & Mur, J. (2014). Modeling regional economic dynamics: Spatial dependence, spatial heterogeneity and nonlinearities. *Journal of Economic Dynamics and Control*, 229–245.

Beine, M., Bertoli, S., & Fernandez-Huertas-Moraga, J. (2015). A practitioners' guide to gravity models of international migration. *World Econ*, *39*, 496–512.

Bergstrand, J. H. (1985). The gravity equation in international trade: Some microeconomic foundations and empirical evidence. *The Review of Economics and Statistics*, *67*(3), 474–481.

Bester, C. A., Conley, T., Hansen, C., & Vogelsang, T. (2016). Fixed-b asymptotics for spatially dependent robust nonparametric covariance matrix estimators. *Econometric Theory*, *32*(1), 154–186.

Bester, C. A., Conley, T. G., & Hansen, C. B. (2011). Inference with dependent data using cluster covariance estimators. *Journal of Econometrics*, *165*(2), 137–151.

Blundell, R., & Powell, J. (2003). Endogeneity in nonparametric and semiparametric regression models. *Econometric society monographs*, *36*, 312–357.

Cameron, C., & Trivedi, P. (2005). *Microeconometrics: Methods and applications*. Cambridge University Press.

Chamberlain, G. (1982). Multivariate regression models for panel data. *Journal of Econometrics*, *18*(1), 5–46.

Chaney, T. (2018). The gravity equation in international trade: An explanation. *Journal of Political Economy*, *126*(1), 150–177.

Conley, T. G. (1999). Gmm estimation with cross sectional dependence. *Journal of Econometrics*, *92*(1), 1–45.

Conley, T. G., & Molinari, F. (2007). Spatial correlation robust inference with errors in location or distance. *Journal of Econometrics*, *140*, 76–96.

Cornwell, C., & Trumbull, W. N. (1994). Estimating the economic model of crime with panel data. *The Review of Economics and Statistics*, *76*(2), 360–366.

Dagenais, M. (1973). The use of incomplete observations in multiple regression analysis. *Journal of Econometrics*, *1*(4), 317–328.

Dardanoni, Valentino, S., Modica, F., & Peracchi. (2011). Regression with imputed covariates: A generalized missing-indicator approach. *Journal of Econometrics*, *162*(2), 362–368.

Debarsy, N. (2012). The mundlak approach in the spatial durbin panel data model. *Spatial Economic Analysis*, *7*(1), 109–131.

Driscoll, J. C., & Kraay, A. C. (1998). Consistent covariance matrix estimation with spatially dependent panel data. *The Review of Economics and Statistics*, *80*(4), 549–560.

Eaton, J., & Kortum, S. (2002). Technology, geography, and trade. *Econometrica*, *70*(5), 1741–1779.

Flach, L., & Unger, F. (2022). Quality and gravity in international trade. *Journal of International Economics*, *137*, 103578.

Gourieroux, C., & Monfort, A. (1981). On the problem of missing data in linear models. *The Review of Economic Studies*, *48*(4), 579–586.

Greene, W. (2007). *Econometric analysis* (7th ed.). Prentice Hall.

Jones, M. (1996). Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American Statistical Association*, *91*(433), 222–230.

Joshi, R., & Wooldridge, J. M. (2019). Correlated random effects models with endogenous explanatory variables and unbalanced panels. *Annals of Economics and Statistics*, (134).

Kabir, M., Salim, R., & Al-Mawali, N. (2017). The gravity model and trade flows: Recent developments in econometric modeling and empirical evidence. *Economic Analysis and Policy*, *56*, 60–71.

Kapoor, M., Kelejian, H., & Prucha, I. R. (2007). Panel data models with spatially correlated error components. *Journal of Econometrics*, *140*(140), 97–130.

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

Kelejian, H., & Prucha, I. (2007). Hac estimation in a spatial framework. *Journal of Econometrics*, *140*(1), 131–154.

Kelejian, H., & Prucha, I. (2010). Spatial models with spatially lagged dependent variables and incomplete data. *Journal of Geographical Systems*, (12), 241–257.

Kelejian, H., Prucha, I. R., & Yuzefovich, Y. (2004). Instrumental variable estimation of a spatial autoregressive model with autoregressive disturbances: Large and small sample results. In *Spatial and spatiotemporal econometrics*. Emerald Group Publishing Limited.

Kiefer, N. M., & Vogelsang, T. J. (2005). A new asymptotic theory for heteroskedasticity-autocorrelation robust tests. *Econometric Theory*, *21*(6), 1130–1164.

Kim, M. S., & Sun, Y. (2011). Spatial heteroskedasticity and autocorrelation consistent estimation of covariance matrix. *Journal of Econometrics*, *160*, 346–371.

Kim, M. S., & Sun, Y. (2013). Heteroskedasticity and spatiotemporal dependence robust inference for linear panel models with fixed effects. *Journal of Econometrics*, *177*, 85–108.

Krugman, P. (1980). Scale economies, product differentiation, and the pattern of trade. *American Economic Review*, *70*(5), 950–959.

Kunsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *Annals of Statistics*, *17*(3), 1217–1241.

Lee, L.-f. (2003). Best spatial two-stage least squares estimators for a spatial autoregressive model with autoregressive disturbances. *Econometric Reviews*, *22*(4), 307–335.

Lesage, J., & Pace, K. (2004). Models for spatially dependent missing data. *Journal of Real Estate Finance and Economics*, *29*(2), 233–254.

LeSage, J., & Pace, R. K. (2009). *Introduction to spatial econometrics*. CRC Press.

Li, L., & Yang, Z. (2020). Spatial dynamic panel data models with correlated random effects. *Journal of Econometrics*.

Little, R., & Rubin, D. (2019). *Statistical analysis with missing data* (3rd edition). Wiley.

Matyas, L. (1997). Proper econometric specification of the gravity model. *The World Economy*, *20*, 363–368.

Matyas, L. (Ed.). (2017). *The econometrics of multi-dimensional panels: Theory and applications*. Springer.

McMillen, D. P. (1996). One hundred fifty years of land values in chicago: A nonparametric approach. *Journal of Urban Economics*, *40*, 100–124.

Müller, U. K. (2014). Hac corrections for strongly autocorrelated time series. *Journal of Business and Economic Statistics*, *32*(3), 311–321.

Müller, U. K., & Watson, M. W. (2022a). Spatial correlation robust inference. *Econometrica*, *90*(6), 2901–2935.

Müller, U. K., & Watson, M. W. (2022b). Spatial correlation robust inference in linear regression and panel models. *Journal of Business and Economics Statistics*, *00*(0), 1–15.

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica*, *46*(1), 69–85.

Mutl, J., & Pfaffermayr, M. (2010). The hausman test in a cliff and ord panel model. *Econometrics Journal*, *10*, 1–30.

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

Newey, W. K., & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticityand autocorrelation consistent covariance matrix. *Econometrica*, *55*, 703–798.

Okawa, Y., & Van Wincoop, E. (2012). Gravity in international finance. *Journal of International Economics*, *87*(2), 205–215.

Papke, L. E. (2005). The effects of spending on test pass rates: Evidence from Michigan. *Journal of Public Economics*, *89*(5-6), 821–839.

Papke, L. E., & Wooldridge, J. M. (2008). Panel data methods for fractional response variables with an application to test pass rates. *Journal of Econometrics*, *145*(1-2), 121–133.

Politis, D. N., & White, H. (2004). Automatic block-length selection for the dependent bootstrap. *Econometric Reviews*, *23*(1), 53–70.

Rai, B. (2021). *Efficient estimation with missing values in cross section and panel data.* [Doctoral dissertation, Michigan State University].

Rai, B. (2023). Eficient estimation with missing data and endogeneity. *Econometric Reviews*, *42*(2), 220–239.

Vogelsang, T. (2012). Heteroskedasticity, autocorrelation, and spatial correlation robust inference in linear panel models with fixed-effects. *Journal of Econometrics*, *166*, 303–319.

Wang, W., & Lee, L.-F. (2013). Estimation of spatial autoregressive models with randomly missing data in the dependent variable. *The Econometrics Journal*, *16*, 73–102.

Wang, W., & Lee, L.-f. (2013). Estimation of spatial panel data models with randomly missing data in the dependent variable. *Regional Science and Urban Economics*, (43), 521–538.

Wheeler, D., & Tiefelsdorf, M. (2005). Multicollinearity and correlation among local regression coefficients in geographically weighted regression. *Journal of Geographical Systems*, *7*(2), 161–187.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, *48*(4), 817–838.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data* (2nd ed.). MIT Press.

Wooldridge, J. M. (2019). Correlated random effects models with unbalanced panels. *Journal of Econometrics*, *211*, 137–150.

Wooldridge, J. M. (2021). Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators. *SSRN Electronic Journal*.

Wu-Chaves, S. (2024). *Essays in spatial panel data econometrics.* [Unpublished doctoral dissertation], Michigan State University.

Yang, Y. (2022). A correlated random effects approach to the estimation of models with multiple fixed effects. *Economics Letters*, *213*, 110408.

APPENDIX A

ADDITIONAL ASSUMPTIONS AND DEFINITIONS FOR CHAPTER 1

Assumption 1

The functions $g_i(\cdot, \theta)$ satisfy these conditions:

1. $g_i(\cdot, \theta)$ are Borel measurable on $\mathcal{Z}$, the σ-algebra generated by Z, for all $\theta \in \Theta$.

2. $\sup_N \sup_{i \in D_N} \mathbb{E}[|g_i(Z_{i,N}, \theta)|^{2+\eta}] < \infty \;\; \forall \theta \in \Theta$ for some $\eta > 0$.

Assumption 2

The $g(\cdot, \cdot)$ satisfy the following conditions:

1. For some $p \geq 1$:

$$\limsup_{n \to \infty} \frac{1}{|D_N|} \sum_{i \in D_N} \mathbb{E}\left[d_{i,N}^p \mathbb{1}(d_{i,N}^p > k)\right] \to 0 \text{ as } k \to \infty$$

   where $d_{i,N} = \sup\limits_{\theta \in \Theta} |g_{i,N}(Z_{i,N}, \theta)|$.

2. $g_i(Z_{i,N}, \theta)$ are L_0 stochastically equicontinuous.

Assumption 3

The true parameter θ_0 and the $g_i(\cdot, \cdot)$ satisfy these conditions:

1. $\theta_0 \in \text{int}(\Theta)$.

2. $g_i(Z_i, \cdot)$ is continuously differentiable on the interior of Θ.

3. $|\nabla_\theta g_i(Z_i, \theta)| < \infty$, where ∇_θ denotes the gradient of $g_i(Z_i, \theta)$ with respect to the parameter vector θ.

4. $\nabla_\theta g_i(Z_i, \theta)$ is Borel measurable, $\mathbb{E}[\nabla_\theta g_i(Z_i, \theta)]$ exists and rank $\{\mathbb{E}[\nabla_\theta g_i(A_i, \theta)]\} = P$, where $P = \dim(\theta_0)$.

5. $\mathbb{E}[|g_i(Z_i, \theta_0)|^{2+\epsilon}] < \infty$ for some $\epsilon > 0$.

Assumption 4

There exist finite dimensional vectors m_i and Δ such that $\hat{u_i} - u_i = m_i\Delta$ and

$$\frac{1}{N}\sum_{i=1}^{N} ||z_i||^2 = O_p(1) \quad \text{and} \quad N^{\frac{1}{2}}||\Delta|| = O_p(1)$$

Definitions

***α*-mixing for random fields**

Let D_N be a subset of D. For $U \subseteq D_N$ and $V \subseteq D_N$, let $\sigma_n(U) = \sigma(X_{i,N} : i \in U)$, $\alpha_N(U,V) = \alpha(\sigma_n(U), \sigma_n(V))$. Then the α-mixing coefficients for the random field $\{X_{i,N} : i \in D_N, N \in \mathbb{N}\}$ is defined as follows:

$$\alpha_{k,l,N}(r) = \sup(\alpha_n(U,V), |U| \leq k, |V| \leq l, \rho(U,V) \geq r)$$

for $k, l, r, n \in \mathbb{N}$. Define also

$$\bar{\alpha}_{k,l}(r) = \sup_N \alpha_{k,l,N}(r)$$

Upper tail quantile function

Let X be a random variable. Then the upper quantile function $Q_X : (0,1) \to [0,\infty)$ is defined as:

$$Q_X(u) = \inf\{t : P(X > t) \leq u\}$$

“Inverse" function of mixing coefficients

For the non-increasing sequence of the mixing coefficients $\{\bar{\alpha}_{1,1}\}_{m=1}^{\infty}$, set $\bar{\alpha}_{1,1}(0) = 1$ and define its “inverse” function $\alpha_{\text{inv}}(u) : (0,1) \to \mathbb{N} \cup \{0\}$ as:

$$\alpha_{\text{inv}}(u) = \max\{m \geq 0 : \bar{\alpha}_{1,1}(m) > u\}$$

Stochastic equicontinuity

The array of random functions $\{f_{i,N}(Z_{i,N}, \theta) : i \in D_N, n \geq 1\}$ is:

1. L_0 stochastically equicontinuous on Θ iff for every $\varepsilon > 0$,

$$\limsup_{N\to\infty} \frac{1}{|D_N|} \sum_{i \in D_N} P\left[\sup_{\theta' \in \Theta} \sup_{\theta \in B(\theta',\delta)} |f_{i,N}(Z_{i,N}, \theta) - f_{i,N}(Z_{i,N}, \theta')| > \varepsilon\right] \to 0 \text{ as } \delta \to 0.$$

2. L_p stochastically equicontinuous, $p > 0$, on Θ iff

$$\limsup_{N\to\infty} \frac{1}{|D_N|} \sum_{i\in D_N} \mathbb{E}\left[\sup_{\theta'\in\Theta} \sup_{\theta\in B(\theta',\delta)} |f_{i,N}(Z_{i,N},\theta) - f_{i,N}(Z_{i,N},\theta')|^p\right] \to 0 \text{ as } \delta \to 0.$$

3. a.s. stochastically equicontinuous on Θ iff

$$\limsup_{N\to\infty} \frac{1}{|D_N|} \sum_{i\in D_N} \sup_{\theta'\in\Theta} \sup_{\theta\in B(\theta',\delta)} |f_{i,N}(Z_{i,N},\theta) - f_{i,N}(Z_{i,N},\theta')| \to 0 \text{ a.s. as } \delta \to 0.$$

APPENDIX B

PROOFS FOR CHAPTER 1

Proof of Proposition 5

For notation simplicity, we will assume that $W_i y_t$ is included in x_{2it}, $x_{it} = [x_{1it} \;\; x_{2it}]$, where the x_2 are $k_2 + 1$ endogenous variables and $z_{it} = [x_{1it} \;\; z_{2it} \;\; w^2_{1it} \ldots w^s_{1it}]$, where z_2 is a vector of L_2 instruments for x_2, with $L_2 \geq k_2$, and similarly for the spatial variables (note however that $W_i y_t$ is not in $W_i X_t$). Therefore, the problem is to apply Pooled 2SLS to the following equation:

$$
\begin{aligned}
y_{it} - \eta_i \bar{y}_i &= (x_{it} - \eta_i \bar{x}_i)\beta + W_i(X_t - \eta_i \bar{X})\gamma + (1 - \eta_i)\bar{z}_i\delta + (1 - \eta_i)W_i\bar{Z}\lambda \\
&= (x_{it} - \eta_i \bar{x}_i)\beta + (w_{it} - \eta_i \bar{w}_i)\gamma + (1 - \eta_i)\bar{z}_i\delta + (1 - \eta_i)\bar{\mathfrak{z}}_i\lambda
\end{aligned}
$$

using IV's: $[(z_{it} - \eta\bar{z}_i) \quad (\mathfrak{z}_{it} - \eta\bar{\mathfrak{z}}_i) \quad (1-\eta)\bar{z}_{2i} \quad (1-\eta)\bar{\mathfrak{z}}_i]$.

We first orthogonalize the IV's, i.e., we run $z_{it} - \eta\bar{z}_i = (1-\eta)\bar{z}_i\epsilon_1 + (1-\theta_i)\bar{\mathfrak{z}}_i\epsilon_2$ and obtain the residuals r_{it} and $\mathfrak{z}_{it} - \eta\bar{\mathfrak{z}}_i = (1-\eta)\bar{z}_i\epsilon_3 + (1-\theta_i)\bar{\mathfrak{z}}_i\epsilon_4$ and get the residuals s_{it}. To do so, we use the Frish-Waugh-Lovell theorem sequentially.

1.a) $z_{it} - \eta\bar{z}_i$ on $(1-\eta)\bar{z}_i$. The coefficient will be:

$$
\begin{aligned}
\tilde{\epsilon}_1 &= \left[\sum_{i=1}^{N}\sum_{t=1}^{T}(1-\eta)^2\bar{z}'_{1i}\bar{z}_{1i}\right]^{-1}\left[\sum_{i=1}^{N}\sum_{t=1}^{T}(1-\eta)^2\bar{z}'_{1i}\bar{z}_i\right] \\
&= \left[\sum_{i=1}^{N}\sum_{t=1}^{T}(1-\eta)^2\bar{z}'_{1i}\bar{z}_{1i}\right]^{-1}\left[\sum_{i=1}(1-\eta)\bar{z}'_i\sum_{t=1} z_{it} - \sum_{i=1} T(1-\eta)\eta\bar{z}'_i\bar{z}_i\right] \\
&= \left[\sum_{i=1}^{N} T(1-\eta)^2\bar{z}'_i\bar{z}_i\right]^{-1}\left[\sum_{i=1}^{N} T(1-\eta)^2\bar{z}'_{1i}\bar{z}_{1i}\right] \\
&= \left[\sum_{i=1}^{N}(1-\eta)^2\bar{z}'_{1i}\bar{z}_{1i}\right]^{-1}\left[\sum_{i=1}^{N}(1-\eta)^2\bar{z}'_{1i}\bar{z}_{1i}\right] = \mathbf{I}_L
\end{aligned}
$$

Therefor the residuals will be $v_{it} = z_{it} - \bar{z}_i$.

1.b) Run $(1-\eta)\bar{\mathfrak{z}}_i$ on $(1-\eta)\bar{z}_i$. In this case the coefficient and the residuals will depend only on the i index, call the latter f_i.

1.c) Run v_{it} on f_i to get ϵ_2. The coefficient will be:

$$\begin{aligned}
\epsilon_2 &= \left[\sum_{i=1}\sum_{t=1} f_i' f_i\right]^{-1}\left[\sum_{i=1}\sum_{t=1} f_i' v_{it}\right] \\
&= \left[\sum_{i=1}\sum_{t=1} f_i' f_i\right]^{-1}\left[\sum_{i=1} f_i' \sum_{t=1} v_{it}\right] \\
&= \left[\sum_{i=1}\sum_{t=1} f_i' f_i\right]^{-1}\left[\sum_{i=1} f_i' \sum_{t=1} (z_{it} - \bar{z}_i)\right] = \mathbf{0}_L
\end{aligned}$$

where we used the fact that the sum of deviations from the mean add up to zero for all i in the second term. This implies that $\epsilon_1 = \mathbf{I}_L$ and therefore, $r_{it} = z_{it} - \bar{z}_i$.

Using very similar steps, it can be shown that if we run $\mathfrak{z}_{it} - \eta\bar{\mathfrak{z}}_i = (1-\eta)\bar{z}_i\epsilon_3 + (1-\theta_i)\bar{\mathfrak{z}}_i\epsilon_4$, then $\epsilon_3 = \mathbf{0}_L$ and $\epsilon_4 = \mathbf{I}_L$, and therefore the residuals of this regression will be $s_{it} = \mathfrak{z}_{it} - \bar{\mathfrak{z}}_i$. Since we have orthogonalized the instrumental variables with respect to $(1-\eta)\bar{z}_i$ and $(1-\eta)\bar{\mathfrak{z}}_i$, we now have to apply Pooled 2SLS to the following equation:

$$y_{it} - \eta\bar{y}_i = (x_{it} - \eta\bar{x}_i)\beta + (w_{it} - \eta\bar{w}_i)\gamma$$

using IV's $[(z_{it} - \bar{z}_i) \quad (\mathfrak{z}_{it} - \bar{\mathfrak{z}}_i)]$. We now define the following notation: $\ddot{z}_{it} = z_{it} - \bar{z}_i$, $\ddot{\mathfrak{z}}_{it} = \mathfrak{z}_{it} - \bar{\mathfrak{z}}_i$, $\hat{z}_{it} = [\ddot{z}_{it} \quad \ddot{\mathfrak{z}}_{it}]$, $\tilde{y}_{it} = y_{it} - \eta\bar{y}_i$, $\tilde{x}_{it} = [(x_{it} - \eta\bar{x}_i) \quad (w_{it} - \eta\bar{w}_i)]$, $\hat{y}_{it} = y_{it} - \bar{y}_i$ and $\hat{x}_{it} = [(x_{it} - \bar{x}_i) \quad (w_{it} - \bar{w}_i)]$. Then the $\Gamma = (\beta \;\; \gamma)$ from the previous problem can be obtained as:

$$\begin{aligned}
\hat{\Gamma}_{2SLS} = &\left[\left(\sum_{i=1}\sum_{t=1}\tilde{x}_{it}'\hat{z}_{it}\right)\left(\sum_{i=1}\sum_{t=1}\hat{z}_{it}'\hat{z}_{it}\right)^{-1}\left(\sum_{i=1}\sum_{t=1}\hat{z}_{it}'\tilde{x}_{it}\right)\right]^{-1}\cdot \\
&\left(\sum_{i=1}\sum_{t=1}\tilde{x}_{it}'\hat{z}_{it}\right)\left(\sum_{i=1}\sum_{t=1}\hat{z}_{it}'\hat{z}_{it}\right)^{-1}\left(\sum_{i=1}\sum_{t=1}\hat{z}_{it}'\tilde{y}_{it}\right) \qquad \text{(B.1)}
\end{aligned}$$

The first term of the square bracket term can be rewritten as follows (the third term of that inverse matrix can also be written in a similar way):

$$\sum_{i=1}\sum_{t=1}\tilde{x}'_{it}\hat{z}_{it} = \sum_{i=1}\sum_{t=1}\begin{bmatrix}(x_{it}-\eta\bar{x}_i)' \\ (w_{it}-\eta\bar{w}_i)'\end{bmatrix}\begin{bmatrix}\ddot{z}_{it} & \ddot{\mathfrak{z}}_{it}\end{bmatrix}$$

$$= \sum_{i=1}\sum_{t=1}\begin{bmatrix}(x_{it}-\eta\bar{x}_i)'\ddot{z}_{it} & (x_{it}-\eta\bar{x}_i)'\ddot{\mathfrak{z}}_{it} \\ (w_{it}-\eta\bar{w}_i)'\ddot{z}_{it} & (w_{it}-\eta\bar{w}_i)'\ddot{\mathfrak{z}}'_{it}\end{bmatrix} \tag{B.2}$$

We focus on the (1,1) term, but the following algebraic manipulation holds for the rest of the terms in the matrix and for the second term in (B.1):

$$\begin{aligned}
\sum_{i=1}\sum_{t=1}(x_{it}-\eta\bar{x}_i)'\ddot{z}_{it} &= \sum_{i=1}\sum_{t=1}x'_{it}\ddot{z}_{it} - \sum_{i=1}\eta\bar{x}'_i\sum_{t=1}\ddot{z}_{it} \\
&= \sum_{i=1}\sum_{t=1}x'_{it}\ddot{z}_{it} - \sum_{i=1}\eta\bar{x}'_i\sum_{t=1}(z_{it}-\bar{z}_i) \\
&= \sum_{i=1}\sum_{t=1}x'_{it}\ddot{z}_{it} \\
&= \sum_{i=1}\sum_{t=1}x'_{it}\ddot{z}_{it} - \sum_{i=1}\bar{x}'_i\sum_{t=1}(z_{it}-\bar{z}_i)' \\
&= \sum_{i=1}\sum_{t=1}(x_{it}-\bar{x}_i)'\ddot{z}_{it}
\end{aligned}$$

where in the second and fourth lines we used the fact that the sum of deviations from the mean over t add up to zero for all observations. Therefore, (B.2) can be rewritten as:

$$\begin{aligned}
\sum_{i=1}\sum_{t=1}\begin{bmatrix}(x_{it}-\eta\bar{x}_i)'\ddot{z}_{it} & (x_{it}-\eta\bar{x}_i)'\ddot{\mathfrak{z}}_{it} \\ (w_{it}-\eta\bar{w}_i)'\ddot{z}_{it} & (w_{it}-\eta\bar{w}_i)'\ddot{\mathfrak{z}}'_{it}\end{bmatrix} &= \sum_{i=1}\sum_{t=1}\begin{bmatrix}(x_{it}-\bar{x}_i)'\ddot{z}_{it} & (x_{it}-\bar{x}_i)'\ddot{\mathfrak{z}}_{it} \\ (w_{it}-\bar{w}_i)'\ddot{z}_{it} & (w_{it}-\bar{w}_i)'\ddot{\mathfrak{z}}'_{it}\end{bmatrix} \\
&= \sum_{i=1}\sum_{t=1}\hat{x}'_{it}\hat{z}_{it}
\end{aligned}$$

Similarly,

$$\sum_{i=1}\sum_{t=1}\hat{z}'_{it}\tilde{y}_{it} = \sum_{i=1}\sum_{t=1}\hat{z}'_{it}\hat{y}_{it}$$

Therefore,

$$\begin{aligned}
\hat{\Gamma}_{2SLS} =& \left[\left(\sum_{i=1} \sum_{t=1} \tilde{x}'_{it} \hat{z}_{it} \right) \left(\sum_{i=1} \sum_{t=1} \hat{z}'_{it} \hat{z}_{it} \right)^{-1} \left(\sum_{i=1} \sum_{t=1} \hat{z}'_{it} \tilde{x}_{it} \right) \right]^{-1} \cdot \\
& \left(\sum_{i=1} \sum_{t=1} \tilde{x}'_{it} \hat{z}_{it} \right) \left(\sum_{i=1} \sum_{t=1} \hat{z}'_{it} \hat{z}_{it} \right)^{-1} \left(\sum_{i=1} \sum_{t=1} \hat{z}'_{it} \tilde{y}_{it} \right) \\
=& \left[\left(\sum_{i=1} \sum_{t=1} \hat{x}'_{it} \hat{z}_{it} \right) \left(\sum_{i=1} \sum_{t=1} \hat{z}'_{it} \hat{z}_{it} \right)^{-1} \left(\sum_{i=1} \sum_{t=1} \hat{z}'_{it} x_{it} \right) \right]^{-1} \cdot \\
& \left(\sum_{i=1} \sum_{t=1} \hat{x}'_{it} \hat{z}_{it} \right) \left(\sum_{i=1} \sum_{t=1} \hat{z}'_{it} \hat{z}_{it} \right)^{-1} \left(\sum_{i=1} \sum_{t=1} \hat{z}'_{it} \hat{y}_{it} \right) \\
=& \hat{\Gamma}_{FE2SLS}
\end{aligned}$$

□

Proof of Proposition 6

For notation simplicity and without loss of generality, I will omit $W_i y_t$ in the proof. This term can be treated as an additional endogenous variable included in x_{2it} with its respective instruments $[w^2_{1it} \dots w^s_{1it}]$. Let $x_{it} = (x_{1it} \quad x_{2it})$, where x_{1it} is a $1 \times \mathrm{k}_1$ vector of exogenous variables and x_{2it} is a $1 \times \mathrm{k}_2$ vector of endogenous covariates.

Similarly, $X_t = (X_{1t} \quad X_{2t})$, $z_{it} = (x_{1it} \quad z_{2it})$, $\bar{z}_i = (\bar{x}_{1i} \quad \bar{z}_{2i})$, $Z_t = (X_{1t} \quad Z_{2t})$ and $\bar{Z}_t = (\bar{X}_1 \quad \bar{Z}_2)$, $\mathfrak{z}_{2it} = W_i Z_{2it}$, $\bar{\mathfrak{z}}_{2i} = W_i \bar{Z}_2$.

Finally denote $\hat{x}_{it} = (x_{1it} \quad \hat{x}_{2it})$, $\hat{\bar{x}}_i = (\bar{x}_{1i} \quad \hat{\bar{x}}_{2i})$, $\hat{\bar{X}} = (\bar{X}_1 \quad \hat{\bar{X}}_2)$, where the hats denote the linear projections of x_2 on $(x_1 \quad z_2)$ and their spatial lags.

In a spatial setting, $(\beta \quad \gamma)_{FE2SLS}$ can be obtained by applying Pooled 2SLS to

$$y_{it} - \bar{y}_i = (x_{1it} - \bar{x}_{1i})\beta_1 + (x_{2it} - \bar{x}_{2i})\beta_2 + W_i(X_{1t} - \bar{X}_1)\gamma_1 + W_i(X_{2t} - \bar{X}_2)\gamma_2 + (u_{it} - u_i)$$

using IV's: $[(z_{2it} - \bar{z}_{2i}) \quad W_i(Z_{2t} - \bar{Z}_2)]$

We want to show that applying Pooled 2SLS to:

$$\begin{aligned}
y_{it} - \theta\bar{y}_i =& (x_{1it} - \theta\bar{x}_{1i})\beta_1 + (x_{2it} - \theta\bar{x}_{2i})\beta_2 + W_i(X_{1t} - \theta\bar{X}_1)\gamma_1 + W_i(X_{2t} - \theta\bar{X}_2)\gamma_2 \\
&+ (1-\theta)\bar{x}_{1i}\delta_1 + (1-\theta)\bar{x}_{2i}\delta_2 + (1-\theta)\bar{W}_i\bar{X}_1\lambda_1 + (1-\theta)\bar{W}_i\bar{X}_2\lambda_2 + u_{it}
\end{aligned}$$

using IV's: $[(z_{2it} - \theta\bar{z}_{2i}) \quad W_i(Z_{2t} - \theta\bar{Z}_2) \quad (1-\theta)\bar{z}_{2i} \quad (1-\theta)W_i\bar{Z}_2]$ yields the same $(\beta \quad \gamma)$.

In order to proof the result, I will follow these steps:

1. Orthogonalize with respect to $[(1-\theta)\bar{x}_{1i} \quad (1-\theta)\bar{w}_{1i}]$ the instrumental variables and $[(x_{1it} - \theta\bar{x}_{1i}) \quad (w_{1it} - \theta\bar{w}_{1i})]$
2. Orthogonalize with respect to $[(1-\theta)\bar{z}_{2i} \quad (1-\theta)\bar{\mathfrak{z}}_{2i}]$ in the first stage equation.
3. Show that we get the same predicted values using the orthogonalized variables and the original ones.
4. Use the Frisch-Waugh-Lovell (WFL) theorem to show the equivalence.

So the model is:

$$\begin{aligned} y_{it} - \theta\bar{y}_i &= (x_{1it} - \theta\bar{x}_{1i})\beta_1 + (x_{2it} - \theta\bar{x}_{2i})\beta_2 + (w_{1it} - \theta\bar{w}_{i1})\gamma_1 + (w_{2it} - \theta\bar{w}_{i2})\gamma_2 \\ &+ (1-\theta)\bar{x}_{1i}\delta_1 + (1-\theta)\bar{x}_{2i}\delta_2 + (1-\theta)\bar{w}_{i1}\lambda_1 + (1-\theta)\bar{w}_{i2}\lambda_2 + u_{it} \end{aligned}$$

using IV's: $[(z_{2it} - \theta\bar{z}_{2i}) \quad (\mathfrak{z}_{2it} - \theta\bar{\mathfrak{z}}_{2i}) \quad (1-\theta)\bar{z}_{2i} \quad (1-\theta)\bar{\mathfrak{z}}_{2i}]$.

Step 1

a. $z_{2it} - \theta\bar{z}_{2i}$ on $(1-\theta)\bar{x}_{1i}$, $(1-\theta)\bar{w}_{1i}$

The residuals will be: $z_{2it} - \theta\bar{z}_{2i} - (1-\theta)\bar{x}_{1i}\hat{\eta}_1 - (1-\theta)\bar{w}_{1i}\hat{\eta}_2 = l_{it}$

Applying the FWL theorem: for $(1-\theta)\bar{x}_{1i}$ on $(1-\theta)\bar{w}_{1i}$, the coefficient will be:

$$\begin{aligned} \hat{\mu}_1 &= \left[\sum_{i=1}^{N}\sum_{t=1}^{T}(1-\theta)^2\bar{w}'_{1i}\bar{w}_{1i}\right]^{-1}\left[\sum_{i=1}^{N}\sum_{t=1}^{T}(1-\theta)^2\bar{w}'_{1i}\bar{x}_{1i}\right] \\ &= \left[\sum_{i=1}^{N}T(1-\theta)^2\bar{w}'_{1i}\bar{w}_{1i}\right]^{-1}\left[\sum_{i=1}^{N}T(1-\theta)^2\bar{w}'_{1i}\bar{x}_{1i}\right] \\ &= \left[\sum_{i=1}^{N}(1-\theta)^2\bar{w}'_{1i}\bar{w}_{1i}\right]^{-1}\left[\sum_{i=1}^{N}(1-\theta)^2\bar{w}'_{1i}\bar{x}_{1i}\right] \end{aligned}$$

The residuals will be $(1-\theta)\bar{x}_{1i} - (1-\theta)\bar{w}_{1i}\hat{\mu}_1 = s_i$.

Now we regress $z_{2it} - \theta\bar{z}_{2i}$ on $(1-\theta)\bar{w}_{1i}$. The coefficient will be:

$$\begin{aligned}
\hat{\mu}_2 &= \left[\sum_{i=1}^{N}\sum_{t=1}^{T}(1-\theta)^2\bar{w}_{1i}'\bar{w}_{1i}\right]^{-1}\left[\sum_{i=1}^{N}\sum_{t=1}^{T}(1-\theta)^2\bar{w}_{1i}'(z_{2it}-\theta\bar{z}_{2i})\right] \\
&= \left[\sum_{i=1}^{N}T(1-\theta)^2\bar{w}_{1i}'\bar{w}_{1i}\right]^{-1}\left[\sum_{i=1}^{N}(1-\theta)^2\bar{w}_{1i}'\sum_{t=1}^{T}(z_{2it}-\theta\bar{z}_{2i})\right] \\
&= \left[\sum_{i=1}^{N}T(1-\theta)^2\bar{w}_{1i}'\bar{w}_{1i}\right]^{-1}\left[\sum_{i=1}^{N}(1-\theta)^2\bar{w}_{1i}'\{T\times(\bar{z}_{2i}-\theta\bar{z}_{2i})\}\right] \\
&= \left[\sum_{i=1}^{N}(1-\theta)^2\bar{w}_{1i}'\bar{w}_{1i}\right]^{-1}\left[\sum_{i=1}^{N}(1-\theta)^2\bar{w}_{1i}'\bar{z}_{2i}\right]
\end{aligned}$$

The residuals will be $z_{2it} - \theta\bar{z}_{2i} - (1-\theta)\bar{w}_{1i}\hat{\mu}_2 = g_{it}$.

Finally, we run g_{it} on s_i. The coefficient will be:

$$\begin{aligned}
\hat{\eta}_1 &= \left[\sum_{i=1}^{N}\sum_{t=1}^{T}s_i's_i\right]^{-1}\left[\sum_{i=1}^{N}\sum_{t=1}^{T}s_i'g_{it}\right] \\
&= \left[\sum_{i=1}^{N}T\times s_i's_i\right]^{-1}\left[\sum_{i=1}^{N}s_i'\sum_{t=1}^{T}g_{it}\right] \\
&= \left[\sum_{i=1}^{N}T\times s_i's_i\right]^{-1}\left[\sum_{i=1}^{N}T\times s_i'\bar{g}_i\right] \\
&= \left[\sum_{i=1}^{N}s_i's_i\right]^{-1}\left[\sum_{i=1}^{N}(1-\theta)s_i'(\bar{z}_{2i}-\bar{w}_{1i}\hat{\mu}_2)\right]
\end{aligned}$$

Using similar steps, $\hat{\eta}_2$ will be:

$$\hat{\eta}_2 = \left[\sum_{i=1}^{N}s_i^{*\prime}s_i^*\right]^{-1}\left[\sum_{i=1}^{N}(1-\theta)s_i^{*\prime}(\bar{z}_{2i}-\bar{x}_{1i}\hat{\mu}_2^*)\right]$$

where $\hat{\mu}_2^*$ is the coefficient of regressing $z_{2it} - \theta\bar{z}_{2i}$ on $(1-\theta)\bar{x}_{1i}$ and s_i^* are the residuals of regressing $(1-\theta)\bar{w}_{1i}$ on $(1-\theta)\bar{x}_{1i}$

b. $(\mathfrak{z}_{2it} - \theta\bar{\mathfrak{z}}_{2i})$ on $(1-\theta)\bar{x}_{1i}$, $(1-\theta)\bar{w}_{1i}$.

The residuals will be $(\mathfrak{z}_{2it} - \theta\bar{\mathfrak{z}}_{2i}) - (1-\theta)\bar{x}_{1i}\hat{\eta}_3 - (1-\theta)\bar{w}_{1i}\hat{\eta}_4 = m_{it}$.

c. $(1-\theta)\bar{z}_{2i}$ on $(1-\theta)\bar{x}_{1i}$, $(1-\theta)\bar{w}_i$.

The residuals are: $(1-\theta)\bar{z}_i - (1-\theta)\hat{\eta}_5 - (1-\theta)\bar{w}_{1i}\hat{\eta}_6 = v_i$, which only depend on the i subscript.

Applying the FWL theorem, regressing $(1-\theta)\bar{x}_{1i}$ on $(1-\theta)\bar{w}_{1i}$ yields $\hat{\mu}_1$, the same as in step 1a. The residuals will be only a function of the i subscript, say f_i.

Finally, run f_i on s_i and the coefficient will be:

$$\left[\sum_{i=1} T s_i' s_i\right]^{-1}\left[\sum_{i=1} T s_i' f_i\right] = \left[\sum_{i=1} T s_i' s_i\right]^{-1}\left[\sum_{i=1} s_i'(1-\theta)(\bar{z}_{2i} - \bar{w}_{1i}\hat{\mu}_2)\right] = \hat{\eta}_5 = \hat{\eta}_1$$

The same coefficient as above. Following similar steps, it can be shown that

$$\hat{\eta}_6 = \hat{\eta}_2 = \left[\sum_{i=1} T s_i^{*\prime} s_i^*\right]\left[\sum_{i=1} s_i^{*\prime}(1-\theta)(\bar{z}_{2i} - \bar{x}_{1i}\hat{\mu}_2^*)\right]$$

where $\hat{\mu}_2^*$ is defined in step 1a.

$\therefore v_i = (1-\theta)\bar{z}_{2i} - (1-\theta)\bar{x}_{1i}\hat{\eta}_5 - (1-\theta)\bar{w}_{1i}\hat{\eta}_6 = (1-\theta)\bar{z}_{2i} - (1-\theta)\bar{x}_{1i}\hat{\eta}_1 - (1-\theta)\bar{w}_{1i}\hat{\eta}_2$

d. $(1-\theta)\bar{z}_{2i}$ on $(1-\theta)\bar{x}_{1i}, (1-\theta)\bar{w}_{1i}$

The coefficients will only depend in i, denote them by r_i. If $(1-\theta)\bar{z}_{2i} = (1-\theta)\bar{x}_{1i}\hat{\eta}_7 + (1-\theta)\bar{w}_{1i}\hat{\eta}_8$, it can be shown using similar arguments than in the previous step that $\hat{\eta}_7 = \hat{\eta}_3$ and $\hat{\eta}_8 = \hat{\eta}_4$.

e. $x_{1it} - \theta\bar{x}_{1i}$ on $(1-\theta)\bar{x}_{1i}, (1-\theta)\bar{w}_{1i}$

We can apply the FWL theorem to get the coefficients:

i. First if we regress $x_{1it} - \theta\bar{x}_{1i}$ on $(1-\theta)\bar{x}_{1i}$. The coefficient is:

$$\begin{aligned}
&\left[\sum_i \sum_t (1-\theta)^2 \bar{x}'_{1i}\bar{x}_{1i}\right]^{-1} \left[\sum_i \sum_t (1-\theta)\bar{x}'_{1i}(x_{1it} - \theta\bar{x}_{1i})\right] \\
=&\left[\sum_i T(1-\theta)^2 \bar{x}'_{1i}\bar{x}_{1i}\right]^{-1} \left[\sum_i (1-\theta)\bar{x}'_{1i} \sum_t (x_{1it} - \theta\bar{x}_{1i})\right] \\
=&\left[\sum_i T(1-\theta)^2 \bar{x}'_{1i}\bar{x}_{1i}\right]^{-1} \left[\sum_i (1-\theta)\bar{x}'_{1i} T(\bar{x}_{it} - \theta\bar{x}_{1i})\right] \\
=&\left[\sum_i T(1-\theta)^2 \bar{x}'_{1i}\bar{x}_{1i}\right]^{-1} \left[\sum_i T(1-\theta)^2 \bar{x}'_{1i}\bar{x}_{1i}\right] \\
=&\mathbf{I}_{k_1}
\end{aligned}$$

where $\mathbf{I}_{k_1}$ denotes an identity matrix of size k_1. Therefore, the residuals will be $x_{1it} - \bar{x}_{1i}$.

ii. Now regress $(1-\theta)\bar{w}_{1i}$ on $(1-\theta)\bar{x}_{1i}$.

The coefficients and residuals will only depend on i. Denote the later by d_i.

iii. Finally regress $x_{1it} - \bar{x}_{1i}$ on d_i. The coefficient will be:

$$\begin{aligned}
&\left[\sum_i \sum_t d'_i d_i\right]^{-1} \left[\sum_i \sum_t d'_i (x_{1it} - \bar{x}_{1i})\right] \\
=&\left[\sum_i \sum_t d'_i d_i\right]^{-1} \left[\sum_i d'_i \sum_t (x_{1it} - \bar{x}_{1i})\right] = \mathbf{0}_{k_1}
\end{aligned}$$

where we used the fact that $\sum_t (x_{1it} - \bar{x}_{1i}) = 0$. Therefore $x_{1it} - \theta\bar{x}_{1i} = (1-\theta)\bar{x}_{1i}\mathbf{I}_{k_1} + (1-\theta)\bar{w}_{1i}\mathbf{0}_{k_1}$ and the residuals will be $x_{1it} - \bar{x}_{1i}$.

f. $w_{1it} - \theta\bar{w}_{1i}$ on $(1-\theta)\bar{x}_{1i}$ and $(1-\theta)\bar{w}_{1i}$.

Applying the FWL theorem in a similar way than the previous step, we get the following relationship:

$w_{1it} - \theta\bar{w}_{1i} = (1-\theta)\bar{x}_{1i}\mathbf{0}_{k_1} + (1-\theta)\bar{w}_{1i}\mathbf{I}_{k_1}$ and the residuals will be $w_{1it} - \bar{w}_{1i}$.

Therefore, after orthogonalizing, we can apply Pooled 2SLS to:

$$
\begin{aligned}
y_{it} - \theta\bar{y}_i = (x_{1it} - \bar{x}_{1i})\beta_1 + (x_{2it} - \theta\bar{x}_{2i})\beta_2 + (w_{1it} - \bar{w}_{i1})\gamma_1 + (w_{2it} - \theta\bar{w}_{i2})\gamma_2 \\
+ (1-\theta)\bar{x}_{2i}\delta_2 + (1-\theta)\bar{w}_{i2}\lambda_2 + u_{it}
\end{aligned}
$$

using IV's: $[l_{it} \quad m_{it} \quad v_i \quad r_i]$.

Step 2

In this step we orthogonalize with respect to v_i and r_i in the first stage equation. Note that these are the residuals from the previous step associated with $(1-\theta)\bar{z}_{2i}$ and $(1-\theta)\bar{\mathfrak{z}}_{2i}$ respectively, the instrumental variables.

a. $l_{it} = v_i\zeta_1 + r_i\zeta_2 + \varepsilon_1$.

i. l_{it} on v_i. The coefficient will be:

$$
\begin{aligned}
\tilde{\eta}_1 &= \left[\sum_i \sum_t v_i' v_i\right]^{-1} \left[\sum_i \sum_t v_i' l_{it}\right] \\
&= \left[\sum_i T v_i' v_i\right]^{-1} \left[\sum_i v_i' \sum_t l_{it}\right] \\
&= \left[\sum_i v_i' v_i\right]^{-1} \left[\sum_i v_i' \bar{l}_i\right]
\end{aligned}
$$

Note that $l_{it} = z_{2it} - \theta\bar{z}_{2i} - (1-\theta)\bar{x}_{1i}\hat{\eta}_1 - (1-\theta)\bar{w}_{1i}\hat{\eta}_2$, therefore

$$
\begin{aligned}
\bar{l}_i &= \frac{1}{T}\sum_t \left[z_{2it} - \theta\bar{z}_{2i} - (1-\theta)\bar{x}_{1i}\hat{\eta}_1 - (1-\theta)\bar{w}_{1i}\hat{\eta}_2\right] \\
&= (1-\theta)\bar{z}_{2i} - (1-\theta)\bar{x}_{1i}\hat{\eta}_1 - (1-\theta)\bar{w}_{1i}\hat{\eta}_2 \\
&= (1-\theta)(\bar{z}_{2i} - \bar{x}_{1i}\hat{\eta}_1 - \bar{w}_{1i}\hat{\eta}_2) = v_i
\end{aligned}
$$

Therefore, $\tilde{\eta}_1 = \mathbf{I}_l$ since $\hat{\eta}_1 = \hat{\eta}_5$ and $\hat{\eta}_2 = \hat{\eta}_6$. The residuals are $z_{2it} - \bar{z}_{2i}$.

ii. r_i on v_i. In this case, both the coefficient and the residuals are going to depend only on i, call them h_i.

iii. Regress $z_{2it} - \bar{z}_{2i}$ on h_i. The coefficient is:

$$\begin{aligned}\hat{\eta}_3 &= \left[\sum_i \sum_t h_i' h_i\right]^{-1} \left[\sum_i \sum_t h_i'(z_{2it} - \bar{z}_{2i})\right] \\ &= \left[\sum_i \sum_t h_i' h_i\right]^{-1} \left[\sum_i h_i' \sum_t (z_{2it} - \bar{z}_{2i})\right] = \mathbf{0}_l\end{aligned}$$

Because the sum of deviations from the mean add up to zero. Therefore $l_{it} = v_i \mathbf{I}_l + r_i \mathbf{0}_l + \varepsilon$ and the residuals will be $z_{2it} - \bar{z}_{2i}$.

b. $m_{it} = v_i \pi_1 + r_i \pi_2 + \varepsilon_2$

i. m_{it} on r_i

The coefficient will be, after some algebra, $\tilde{\pi}_2 = \left[\sum_i r_i' r_i\right]^{-1} \left[\sum_i r_i' \bar{m}_i\right]$. Noting that

$$\begin{aligned}\bar{m}_i &= \frac{1}{T} \sum_t \left[(\mathfrak{z}_{2it} - \theta\bar{\mathfrak{z}}_{2i}) - (1-\theta)\bar{x}_{1i}\hat{\eta}_3 - (1-\theta)\bar{w}_{1i}\hat{\eta}_4\right] \\ &= (1-\theta)(\bar{\mathfrak{z}}_{2i} - \bar{x}_{1i}\hat{\eta}_3 - \bar{w}_{1i}\hat{\eta}_4) \\ &= (1-\theta)(\bar{\mathfrak{z}}_{2i} - \bar{x}_{1i}\hat{\eta}_7 - \bar{w}_{1i}\hat{\eta}_8) = r_i\end{aligned}$$

We conclude that $\tilde{\pi}_2 = \mathbf{I}_l$ and the residuals are $\mathfrak{z}_{2it} - \bar{\mathfrak{z}}_{2i}$.

ii. v_i on r_i. The coefficient will be denoted by $\tilde{\pi}_1 = \left[\sum_i r_i' r_i\right]^{-1} \left[\sum_i r_i' v_i\right]$, and the residuals will depend on i, call them $\tilde{h}_i$.

iii. $\mathfrak{z}_{2it} - \bar{\mathfrak{z}}_{2i}$ on $\tilde{h}_i$.

Using again the fact that $\sum_t \mathfrak{z}_{2it} - \bar{\mathfrak{z}}_{2i} = 0$, we conclude that $\pi_1 = \mathbf{0}_l$, which implies that $\tilde{\pi}_2 = \pi_2 = \mathbf{I}_l$ and therefore, the residuals will be $\mathfrak{z}_{2it} - \bar{\mathfrak{z}}_{2i}$.

In the original first stage we have:

$$\begin{aligned}x_{2it} - \theta\bar{x}_{2it} = (x_{1it} - \theta\bar{x}_{1i})\phi_1 + (w_{1it} - \theta\bar{w}_{1i})\phi_2 + (\mathfrak{z}_{2it} - \theta\bar{\mathfrak{z}}_{2i})\phi_3 + (w_{2it} - \theta\bar{w}_{i2})\phi_4 \\ + (1-\theta)\bar{x}_{1i}\rho_1 + (1-\theta)\bar{w}_{1i}\rho_2 + (1-\theta)\bar{z}_{2i}\rho_3 + (1-\theta)\bar{\mathfrak{z}}_{2i}\rho_4 + \varepsilon_{FS}\end{aligned}$$

After orthogonalizing with respect to $[(1-\theta)\bar{x}_{1i} \quad (1-\theta)\bar{w}_{1i} \quad (1-\theta)\bar{z}_{2i} \quad (1-\theta)\bar{\mathfrak{z}}_{2i}]$, to get $\Phi = (\phi_1\ \phi_2\ \phi_3\ \phi_4)$, we have to regress

$x_{2it} - \theta\bar{x}_{2it}$ on $[(x_{2it} - \bar{x}_{2it}) \quad (x_{1it} - \bar{x}_{1i}) \quad (w_{1it} - \bar{w}_{1i}) \quad (\mathfrak{z}_{2it} - \bar{\mathfrak{z}}_{2i})]$.

We note that if $z_{it} = [x_{1it}\ \ w_{1it}\ \ z_{2it}\ \ \mathfrak{z}_{2it}]$, then the coefficient of $x_{2it} - \theta\bar{x}_{2it}$ on $z_{it} - \bar{z}_i$ is

$$\begin{aligned}
\check{\Phi} &= \left[\sum_i \sum_t (z_{it} - \bar{z}_i)'(z_{it} - \bar{z}_i)\right]^{-1} \left[\sum_i \sum_t (z_{it} - \bar{z}_i)'(x_{2it} - \theta\bar{x}_{2i})\right] \\
&= \left[\sum_i \sum_t (z_{it} - \bar{z}_i)'(z_{it} - \bar{z}_i)\right]^{-1} \left[\sum_i \sum_t (z_{it} - \bar{z}_i)'x_{2it} - \sum_i \left\{\sum_t (z_{it} - \bar{z}_i)'\right\} \theta\bar{x}_{2i}\right] \\
&= \left[\sum_i \sum_t (z_{it} - \bar{z}_i)'(z_{it} - \bar{z}_i)\right]^{-1} \left[\sum_i \sum_t (z_{it} - \bar{z}_i)'x_{2it} - \sum_i \left\{\sum_t (z_{it} - \bar{z}_i)'\right\} \bar{x}_{2i}\right] \\
&= \left[\sum_i \sum_t (z_{it} - \bar{z}_i)'(z_{it} - \bar{z}_i)\right]^{-1} \left[\sum_i \sum_t (z_{it} - \bar{z}_i)'(x_{2it} - \bar{x}_{2i})\right]
\end{aligned}$$

Where we used the fact that the terms in curly brackets are zero. Therefore, Φ can also be obtained by regressing $(x_{2it} - \bar{x}_{2it})$ on $[(x_{2it} - \bar{x}_{2it})\ \ (x_{1it} - \bar{x}_{1i})\ \ (w_{1it} - \bar{w}_{1i})\ \ (\mathfrak{z}_{2it} - \bar{\mathfrak{z}}_{2i})]$.

Step 3

In this step we show that $\widehat{x_{2it} - \theta\bar{x}_{2i}} = \widetilde{x_{2it} - \theta\bar{x}_{2i}}$, where

$$\begin{aligned}
\widehat{x_{2it} - \theta\bar{x}_{2i}} = {} & (x_{1it} - \theta\bar{x}_{1i})\hat{\phi}_1 + (w_{1it} - \theta\bar{w}_{1i})\hat{\phi}_2 + (z_{2it} - \theta\bar{z}_{2i})\hat{\phi}_3 + (\mathfrak{z}_{2it} - \theta\bar{\mathfrak{z}}_{2i})\hat{\phi}_4 \\
& + (1-\theta)\bar{x}_{1i}\hat{\rho}_1 + (1-\theta)\bar{w}_{1i}\hat{\rho}_2 + (1-\theta)\bar{z}_{2i}\hat{\rho}_3 + (1-\theta)\bar{\mathfrak{z}}_{2i}\hat{\rho}_4
\end{aligned}$$

$$\begin{aligned}
\widetilde{x_{2it} - \theta\bar{x}_{2i}} = {} & (x_{1it} - \bar{x}_{1i})\tilde{\phi}_1 + (w_{1it} - \bar{w}_{1i})\tilde{\phi}_2 + (z_{2it} - \bar{z}_{2i})\tilde{\phi}_3 + (\mathfrak{z}_{2it} - \bar{\mathfrak{z}}_{2i})\tilde{\phi}_4 \\
& + (1-\theta)\bar{x}_{1i}\tilde{\rho}_1 + (1-\theta)\bar{w}_{1i}\tilde{\rho}_2 + (1-\theta)\bar{z}_{2i}\tilde{\rho}_3 + (1-\theta)\bar{\mathfrak{z}}_{2i}\tilde{\rho}_4
\end{aligned}$$

First we note that $\hat{\phi}_j = \tilde{\phi}_j$ for $j = 1, 2, 3, 4$ because in the second equation the respective explanatory variables are orthogonalized with respect to the terms related to the time averages of the independent variables. Given this fact and after some algebra, we have that $\widehat{x_{2it} - \theta\bar{x}_{2i}} = \widetilde{x_{2it} - \theta\bar{x}_{2i}}$ if $\hat{\phi}_j + \hat{\rho}_j = \tilde{\rho}_j$ for $j = 1, 2, 3, 4$.

To show that the previous equality holds, we start with $\widetilde{x_{2it} - \theta\bar{x}_{2i}}$. Since $z_{it} = [x_{1it}\ w_{1it}\ z_{2it}\ \mathfrak{z}_{2it}]$ as above, we have

$z_{it} - \bar{z}_i = \left[(x_{1it} - \bar{x}_{1i})\ \ (w_{1it} - \bar{w}_{1i})\ \ (z_{2it} - \bar{z}_{2i})\ \ (\mathfrak{z}_{2it} - \bar{\mathfrak{z}}_{2i})\right]$, $\tilde{\rho} = (\tilde{\rho}_1'\ \ \tilde{\rho}_2'\ \ \tilde{\rho}_3'\ \ \tilde{\rho}_4')'$ and $\tilde{\phi} = (\tilde{\phi}_1'\ \ \tilde{\phi}_2'\ \ \tilde{\phi}_3'\ \ \tilde{\phi}_4')'$, therefore, $\widetilde{x_{2it} - \theta\bar{x}_{2i}} = (z_{it} - \bar{z}_i)\hat{\phi} + (1-\theta)\bar{z}_i\tilde{\rho}$. Greene (2007) shows that given

$\hat{\phi}$, one can get $\tilde{\rho}$ as:

$$\begin{aligned}
\tilde{\rho} &= \left[\sum_i \sum_t (1-\theta)^2 \bar{z}_i' \bar{z}_i\right]^{-1} \left[\sum_i \sum_t (1-\theta) \bar{z}_i' \left\{x_{2it} - \theta \bar{x}_{2i} - (z_{it} - \bar{z}_i)\hat{\phi}\right\}\right] \\
&= \left[\sum_i T(1-\theta)^2 \bar{z}_i' \bar{z}_i\right]^{-1} \left[\sum_i \left((1-\theta)\bar{z}_i' \sum_t (x_{2it} - \theta \bar{x}_{2i}) - (1-\theta)\bar{z}_i' \left\{\sum_t (z_{it} - \bar{z}_i)\right\} \hat{\phi}\right)\right] \\
&= \left[\sum_i (1-\theta)^2 \bar{z}_i' \bar{z}_i\right]^{-1} \left[\sum_i (1-\theta)^2 \bar{z}_i' \bar{x}_{2i}\right]
\end{aligned}$$

where we used the fact that $\sum_t (z_{it} - \bar{z}_i) = 0$ on the second line.

We turn now to $\widehat{x_{2it} - \theta \bar{x}_{2i}}$. With similar definitions as above, given $\hat{\phi}$, we get $\hat{\rho}$ as:

$$\begin{aligned}
\hat{\rho} &= \left[\sum_i \sum_t (1-\theta)^2 \bar{z}_i' \bar{z}_i\right]^{-1} \left[\sum_i \sum_t (1-\theta) \bar{z}_i' \left\{(x_{2it} - \theta \bar{x}_{2i}) - (z_{it} - \theta \bar{z}_i)\hat{\phi}\right\}\right] \\
&= \left[\sum_i \sum_t (1-\theta)^2 \bar{z}_i' \bar{z}_i\right]^{-1} \left[\sum_i (1-\theta) \bar{z}_i' \left\{\sum_t (x_{2it} - \theta \bar{x}_{2i}) - \sum_t (z_{it} - \theta \bar{z}_i)\hat{\phi}\right\}\right] \\
&= \left[\sum_i T(1-\theta)^2 \bar{z}_i' \bar{z}_i\right]^{-1} \left[\sum_i T(1-\theta)^2 \bar{z}_i' \bar{x}_{2i}\right] \\
&- \left[\sum_i T(1-\theta)^2 \bar{z}_i' \bar{z}_i\right]^{-1} \left[\sum_i (1-\theta) \bar{z}_i (T\bar{z}_i - T\theta \bar{z}_i)\hat{\phi}\right] \\
&= \tilde{\rho} - \left[\sum_i T(1-\theta)^2 \bar{z}_i' \bar{z}_i\right]^{-1} \left[\sum_i T(1-\theta)^2 \bar{z}_i' \bar{z}_i\right] \hat{\phi} \\
&= \tilde{\rho} - \mathbf{I}_{2(k_1+l)+1} \hat{\phi} = \tilde{\rho} - \hat{\phi}
\end{aligned}$$

Therefore, $\tilde{\rho} = \hat{\rho} + \hat{\phi}$ and hence $\widehat{x_{2it} - \theta \bar{x}_{2i}} = \widetilde{x_{2it} - \theta \bar{x}_{2i}}$.

In a similar way and using obvious notation, it can be shown that

$\widehat{w_{2it} - \theta \bar{w}_{2i}} = \widetilde{w_{2it} - \theta \bar{w}_{2i}}$.

Step 4

Given the previous step, the problem becomes:

$$\begin{aligned}
y_{it} - \theta \bar{y}_i = (x_{1it} - \bar{x}_{1i})\beta_1 + (x_{2it} - \theta \bar{x}_{2i})\beta_2 + (w_{1it} - \bar{w}_{i1})\gamma_1 + (w_{2it} - \theta \bar{w}_{i2})\gamma_2 \\
+ (1-\theta)\bar{x}_{2i}\delta_2 + (1-\theta)\bar{w}_{i2}\lambda_2 + u_{it}
\end{aligned}$$

using IV's: $[(z_{2it} - \bar{z}_{2i}) \quad (\mathfrak{z}_{2it} - \bar{\mathfrak{z}}_{2i}) \quad (1-\theta)\bar{z}_{2i} \quad (1-\theta)\bar{\mathfrak{z}}_{2i}]$. At this point however, it is important to note that although we have orthogonalized with respect to $(1-\theta)[\bar{x}_{1i} \quad \bar{w}_{1i}]$, we still have to include in the first stage equation to obtain the predicted values of the endogenous variables. Given this, the second stage equation is:

$$\begin{aligned} y_{it} - \theta\bar{y}_i &= (x_{1it} - \bar{x}_{1i})\beta_1 + (\hat{x}_{2it} - \theta\hat{\bar{x}}_{2i})\beta_2 + (w_{1it} - \bar{w}_{i1})\gamma_1 + (\hat{w}_{2it} - \theta\hat{\bar{w}}_{i2})\gamma_2 \\ &\quad + (1-\theta)\hat{\bar{x}}_{2i}\delta_2 + (1-\theta)\hat{\bar{w}}_{i2}\lambda_2 \end{aligned}$$

where the ˆ denote the first stage projections on the instrumental variables. To obtain $(\beta \quad \gamma)$, we orthogonalize with respect to $(1-\theta)\hat{\bar{x}}_{2i}$ and $(1-\theta)\hat{\bar{w}}_{i2}$.

a. $(x_{1it} - \bar{x}_{1i})$ on $(1-\theta)\hat{\bar{x}}_{2i}$ and $(1-\theta)\hat{\bar{w}}_{i2}$.

   i. $(x_{1it} - \bar{x}_{1i})$ on $(1-\theta)\hat{\bar{x}}_{2i}$. The coefficient will be:

   $$\left[\sum_i \sum_t (1-\theta)^2 \hat{\bar{x}}'_{2i}\hat{\bar{x}}_{2i}\right]^{-1} \left[\sum_i (1-\theta)\hat{\bar{x}}'_{2i} \sum_t (x_{1it} - \bar{x}_{1i})\right] = \mathbf{0}_{k_2}$$

   where we used that the sums of deviations from the mean are zero for all i and the residuals will be $x_{1it} - \bar{x}_{1i}$.

   ii. $(1-\theta)\hat{\bar{w}}_{i2}$ on $(1-\theta)\hat{\bar{x}}_{2i}$. In this case the coefficients and the residuals will depend only on i, call them $\tilde{u}_i$.

   iii. $x_{1it} - \bar{x}_{1i}$ on $\tilde{u}_i$. By a similar argument to point i just above, the coefficient is $\mathbf{0}_{k_2}$ and so, $x_{1it} - \bar{x}_{1i}$ is orthogonal to both variables.

b . $(w_{1it} - \bar{w}_{1i})$ on $(1-\theta)\hat{\bar{x}}_{2i}$ and $(1-\theta)\hat{\bar{w}}_{i2}$. Using a similar argument as in a) above, $w_{1it} - \bar{w}_{1i}$ is orthogonal to both variables.

c. $(\hat{x}_{2it} - \theta\hat{\bar{x}}_{2i})$ on $(1-\theta)\hat{\bar{x}}_{2i}$ and $(1-\theta)\hat{\bar{w}}_{i2}$.

   i. $(1-\theta)\hat{\bar{w}}_{2i}$ on $(1-\theta)\hat{\bar{x}}_{2i}$. The coefficient and residuals depend only on i, call them $\breve{u}_i$.

   ii. $(\hat{x}_{2it} - \theta\hat{\bar{x}}_{2i})$ on $(1-\theta)\hat{\bar{x}}_{2i}$. By arguments very similar to previous steps, one can show that the coefficient is $\mathbf{I}_{k_2}$ and the residuals will be $(\hat{x}_{2it} - \hat{\bar{x}}_{2i})$.

iii. $(\hat{x}_{2it} - \hat{\bar{x}}_{2i})$ on $\breve{u}_i$. By analogous arguments as above, the coefficient of this regression will be $\mathbf{0}_{k_2}$.

Therefore, the residuals of this regression will be $(\hat{x}_{2it} - \hat{\bar{x}}_{2i})$

d. $\hat{w}_{2it} - \theta\hat{\bar{w}}_{i2}$ on $(1-\theta)\hat{\bar{x}}_{2i}$ and $(1-\theta)\hat{\bar{w}}_{i2}$. Using similar ideas as in c) above, the residuals of this regression are $\hat{w}_{2it} - \hat{\bar{w}}_{i2}$.

Therefore, to find $(\beta_1 \quad \beta_2 \quad \gamma_1 \quad \gamma_2)$, we run

$$y_{it} - \theta\bar{y}_i = (x_{1it} - \bar{x}_{1i})\beta_1 + (\hat{x}_{2it} - \hat{\bar{x}}_{2i})\beta_2 + (w_{1it} - \bar{w}_{i1})\gamma_1 + (\hat{w}_{2it} - \hat{\bar{w}}_{i2})\gamma_2$$

If we collect all the covariates of this regression into a vector $\hat{x}_{it} - \hat{\bar{x}}_i$ (where the x_{1it} and w_{1it} are their own projections), then:

$$\begin{aligned}
(\beta \quad \gamma) &= \left[\sum_i \sum_t (\hat{x}_{it} - \hat{\bar{x}}_i)'(\hat{x}_{it} - \hat{\bar{x}}_i)\right]^{-1} \left[\sum_i \sum_t (\hat{x}_{it} - \hat{\bar{x}}_i)'(y_{it} - \theta\bar{y}_i)\right] \\
&= \left[\sum_i \sum_t (\hat{x}_{it} - \hat{\bar{x}}_i)'(\hat{x}_{it} - \hat{\bar{x}}_i)\right]^{-1} \left[\sum_i \sum_t (\hat{x}_{it} - \hat{\bar{x}}_i)'y_{it} - \sum_i \left\{\sum_t (\hat{x}_{it} - \hat{\bar{x}}_i)\right\}\theta\bar{y}_i\right] \\
&= \left[\sum_i \sum_t (\hat{x}_{it} - \hat{\bar{x}}_i)'(\hat{x}_{it} - \hat{\bar{x}}_i)\right]^{-1} \left[\sum_i \sum_t (\hat{x}_{it} - \hat{\bar{x}}_i)'(y_{it} - \bar{y}_i)\right]
\end{aligned}$$

where we use again the fact that the term in curly brackets in the second line is zero. Therefore, $(\beta \quad \gamma)$ can be obtained by regressing
$y_{it} - \bar{y}_i$ on $\left[(x_{1it} - \bar{x}_{1i}) \quad (\widehat{x_{2it} - \bar{x}_{2i}}) \quad (w_{1it} - \bar{w}_{1i}) \quad (\widehat{w_{2it} - \bar{w}_{2i}})\right]$,
which is exactly the same problem that the Fixed Effects 2SLS estimator solves.

□

APPENDIX C

DERIVATION OF THE COVARIANCE MATRIX FOR THE CONTROL FUNCTION APPROACH

Consider the estimating equation in (1.46):

$$\ddot{y}_{it} = \hat{\ddot{a}}_{it}\theta + \ddot{e}_{it}$$

where we can write $\ddot{a}_{it} = \ddot{z}_{it}\psi + \ddot{v}_{it}$. Because every element in $\ddot{a}_{it}$ is exogenous with respect to the error term $\ddot{e}_{it}$, we can write:

$$\begin{aligned}
\hat{\theta} &= \left[\frac{1}{NT}\sum_i^N\sum_t^T \hat{\ddot{a}}_{it}'\hat{\ddot{a}}_{it}\right]^{-1}\left[\frac{1}{NT}\sum_i^N\sum_t^T \hat{\ddot{a}}_{it}'\ddot{y}_{it}\right] \\
&= \left[\frac{1}{NT}\sum_i^N\sum_t^T \hat{\ddot{a}}_{it}'\hat{\ddot{a}}_{it}\right]^{-1}\left[\frac{1}{NT}\sum_i^N\sum_t^T \hat{\ddot{a}}_{it}'(\ddot{a}_{it}\theta + e_{it})\right] \\
&= \left[\frac{1}{NT}\sum_i^N\sum_t^T \hat{\ddot{a}}_{it}'\hat{\ddot{a}}_{it}\right]^{-1}\left[\frac{1}{NT}\sum_i^N\sum_t^T \hat{\ddot{a}}_{it}'(\ddot{a}_{it}\theta + \hat{\ddot{a}}_{it}\theta - \hat{\ddot{a}}_{it}\theta + \ddot{e}_{it})\right] \\
\implies \sqrt{NT}(\hat{\theta}-\theta) &= \left[\frac{1}{NT}\sum_i^N\sum_t^T \hat{\ddot{a}}_{it}'\hat{\ddot{a}}_{it}\right]^{-1}\left\{(NT)^{-\frac{1}{2}}\sum_i^N\sum_t^T \hat{\ddot{a}}_{it}'\left[\underbrace{(\ddot{a}_{it}-\hat{\ddot{a}}_{it})\theta}_{\text{Part 2}} + \underbrace{\ddot{e}_{it}}_{\text{Part 1}}\right]\right\}
\end{aligned}$$

Note that because $\hat{\psi} \xrightarrow{p} \psi$, the first matrix on the right hand side will converge in probability to $\mathbb{E}\left(\sum_i^N\sum_t^T \ddot{a}_{it}'\ddot{a}_{it}\right) = B$. Consider now Part 1:

$$\begin{aligned}
(NT)^{-\frac{1}{2}}\sum_i^N\sum_t^T \hat{\ddot{a}}_{it}'\ddot{e}_{it} &= (NT)^{-\frac{1}{2}}\sum_i^N\sum_t^T (\ddot{z}_{it}\hat{\psi})'\ddot{e}_{it} \\
&= (NT)^{-\frac{1}{2}}\sum_i^N\sum_t^T (\ddot{z}_{it}\hat{\psi} + \ddot{z}_{it}\psi - \ddot{z}_{it}\psi)'\ddot{e}_{it} \\
&= (NT)^{-\frac{1}{2}}\sum_i^N\sum_t^T (\ddot{z}_{it}\psi)'\ddot{e}_{it} + \frac{1}{NT}\sum_i^N\sum_t^T\left[\ddot{z}_{it}\sqrt{NT}(\hat{\psi}-\psi)\right]'\ddot{e}_{it} \\
&= (NT)^{-\frac{1}{2}}\sum_i^N\sum_t^T (\ddot{z}_{it}\psi)'\ddot{e}_{it} + \underbrace{\sqrt{NT}(\hat{\psi}-\psi)'}_{O_p(1)}\cdot\underbrace{\frac{1}{NT}\sum_i^N\sum_t^T \ddot{z}_{it}'\ddot{e}_{it}}_{o_p(1)}
\end{aligned}$$

Because $O_p(1) \cdot o_p(1) = o_p(1)$, Part 1 converges to $\left[(NT)^{-\frac{1}{2}} \sum_i^N \sum_t^T (\ddot{z}_{it}\psi)'\ddot{e}_{it}\right]$. Now consider Part 2:

$$\begin{aligned}
(NT)^{-\frac{1}{2}} \sum_i^N \sum_t^T \hat{\ddot{a}}_{it}'(\ddot{a}_{it} - \hat{\ddot{a}}_{it})\theta &= (NT)^{-\frac{1}{2}} \sum_i^N \sum_t^T (\ddot{z}_{it}\hat{\psi})'[\ddot{a}_{it} - \ddot{z}_{it}\hat{\psi}]\theta \\
&= (NT)^{-\frac{1}{2}} \sum_i^N \sum_t^T (\ddot{z}_{it}\hat{\psi})'[\ddot{a}_{it} - \ddot{z}_{it}\psi + \ddot{z}_{it}\psi - \ddot{z}_{it}\hat{\psi}]\theta \\
&= (NT)^{-\frac{1}{2}} \sum_i^N \sum_t^T \underbrace{(\ddot{z}_{it}\hat{\psi})'v_{it}\theta}_{\text{Part 2.1}} - \underbrace{(\ddot{z}_{it}\hat{\psi})'\ddot{z}_{it}(\hat{\psi} - \psi)\theta}_{\text{Part 2.2}}
\end{aligned}$$

Starting with Part 2.1:

$$\begin{aligned}
(NT)^{-\frac{1}{2}} \sum_i^N \sum_t^T (\ddot{z}_{it}\hat{\psi})'v_{it}\theta &= (NT)^{-\frac{1}{2}} \sum_i^N \sum_t^T (\ddot{z}_{it}\psi + \ddot{z}_{it}\hat{\psi} - \ddot{z}_{it}\psi)'v_{it}\theta \\
&= (NT)^{-\frac{1}{2}} \sum_i^N \sum_t^T (\ddot{z}_{it}\psi)'v_{it}\theta + \left[\ddot{z}_{it}(\hat{\psi} - \psi)\right]' v_{it}\theta \\
&= (NT)^{-\frac{1}{2}} \left[\sum_i^N \sum_t^T (\ddot{z}_{it}\psi)'v_{it}\theta\right] + \underbrace{\sqrt{NT}(\hat{\psi} - \psi)'}_{O_p(1)} \underbrace{\frac{1}{NT} \sum_i^N \sum_t^T \ddot{z}_{it}'\ddot{v}_{it}\theta}_{\xrightarrow{p}\mathbb{E}(\ddot{z}_{it}'\ddot{v}_{it})=0}
\end{aligned}$$

So in the last line we have $O_p(1) \cdot o_p(1) = o_p(1)$ and therefore the last term will vanish as $N \longrightarrow \infty$ and only $(NT)^{-\frac{1}{2}} \sum_i^N \sum_t^T (\ddot{z}_{it}\psi)'v_{it}\theta$ will remain. Using similar algebra, it can be shown that part 2.2 will converge to

$$-\frac{1}{NT} \sum_i^N \sum_t^T (\ddot{z}_{it}\psi)'\ddot{z}_{it}\sqrt{NT}(\hat{\psi} - \psi)\theta$$

Note that

$$\hat{\psi} - \psi = \left(\sum_i^N \sum_t^T \ddot{z}_{it}'\ddot{z}_{it}\right)^{-1} \left(\sum_i^N \sum_t^T \ddot{z}_{it}'\ddot{v}_{it}\right)$$

$$\Longrightarrow \sqrt{NT}(\hat{\psi} - \psi) = \left(\frac{1}{NT} \sum_i^N \sum_t^T \ddot{z}_{it}'\ddot{z}_{it}\right)^{-1} \left[(NT)^{-\frac{1}{2}} \sum_i^N \sum_t^T \ddot{z}_{it}'\ddot{v}_{it}\right]$$

Putting everything together we have

$$\begin{aligned}
&\left[\frac{1}{NT}\sum_i^N\sum_t^T \hat{\ddot{a}}_{it}'\hat{\ddot{a}}_{it}\right]^{-1}\left\{(NT)^{-\frac{1}{2}}\sum_i^N\sum_t^T \hat{\ddot{a}}_{it}'\left[(\ddot{a}_{it}-\hat{\ddot{a}}_{it})\theta+\ddot{e}_{it}\right]\right\}\\
=&B^{-1}\left\{(NT)^{-\frac{1}{2}}\sum_i^N\sum_t^T(\ddot{z}_{it}\psi)'\left[\ddot{e}_{it}+\ddot{v}_{it}\theta-\ddot{z}_{it}(\hat{\psi}-\psi)\theta\right]\right\}+o_p(1)\\
=&B^{-1}\left\{(NT)^{-\frac{1}{2}}\left[\sum_i^N\sum_t^T\{(\ddot{z}_{it}\psi)'(\ddot{e}_{it}+\ddot{v}_{it}\theta)\}\right]-\frac{1}{NT}\sum_i^N\sum_t^T\{(\ddot{z}_{it}\hat{\psi})'\ddot{z}_{it}\}\sqrt{NT}(\hat{\psi}-\psi)\theta\right\}+o_p(1)
\end{aligned}$$

where $\sqrt{NT}(\hat{\psi}-\psi)=\left(\frac{1}{NT}\sum_i^N\sum_t^T\ddot{z}_{it}'\ddot{z}_{it}\right)^{-1}\left[(NT)^{-\frac{1}{2}}\sum_i^N\sum_t^T\ddot{z}_{it}'\ddot{v}_{it}\right]$.

Let

$$G=\mathbb{E}\left[\sum_i^N\sum_t^T(\ddot{z}_{it}\psi)'\ddot{z}_{it}\right]$$

and

$$r_{it}(\psi)=\left(\frac{1}{NT}\sum_i^N\sum_t^T\ddot{z}_{it}'\ddot{z}_{it}\right)^{-1}\left[(NT)^{-\frac{1}{2}}\sum_i^N\sum_t^T\ddot{z}_{it}'\ddot{v}_{it}\right]$$

Then we can write

$$\sqrt{NT}(\hat{\theta}-\theta)=B^{-1}\left\{(NT)^{-\frac{1}{2}}\sum_i^N\sum_t^T(\ddot{z}_{it}\psi)'(\ddot{e}_{it}+\ddot{v}_{it}\theta)-G\cdot r_{it}(\psi)\theta\right\}+o_p(1)$$

And therefore, by the Central Limit Theorem,

$$\sqrt{NT}(\hat{\theta}-\theta)\overset{a}{\sim}N\left\{0,B^{-1}MB^{-1}\right\}$$

where $M=\text{Var}\left[\sum_i^N\sum_t^T(\ddot{z}_{it}\psi)'(\ddot{e}_{it}+\ddot{v}_{it}\theta)-G\cdot r_{it}(\psi)\theta\right]=\text{Var}\left[\sum_i^N\sum_t^T m_{it}\right]$.

B can be estimated with

$$\hat{B}=\frac{1}{NT}\sum_i^N\sum_t^T\hat{\ddot{a}}_{it}'\hat{\ddot{a}}_{it}$$

To estimate M, let

$$\hat{m}_{it}=(\ddot{z}_{it}\hat{\psi})'(\hat{\ddot{e}}_{it}+\hat{\ddot{v}}_{it}\hat{\theta})-\hat{G}\cdot\hat{r}_{it}(\hat{\psi})\hat{\theta}$$

where,

$\hat{\ddot{e}}_{it}$ are the residuals from the second stage.

$\hat{\ddot{v}}_{it}$ are the residuals from the first stage (note that v_{it} is a vector).

$\hat{G}=\frac{1}{NT}\sum_i^N\sum_t^T(\ddot{z}_{it}\hat{\psi})'\ddot{z}_{it}$.

$\hat{r}(\hat{\psi}) = \left(\frac{1}{NT}\sum_i^N \sum_t^T \ddot{z}'_{it}\ddot{z}_{it}\right)^{-1} \left[(NT)^{-\frac{1}{2}} \sum_i^N \sum_t^T \ddot{z}'_{it}\hat{\hat{v}}_{it}\right].$

With these quantities defined, the (r, s)-th element of M can be estimated as

$$\hat{M}_{rs} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{l=1}^{T}\hat{\hat{m}}_{it,r}\hat{\hat{m}}_{jl,s}K\left[\frac{\rho^*(i,j)}{\rho_b}\right]$$

where once again the kernel function $K(\cdot)$ is operationalizing the weak spatial dependence assumption.

APPENDIX D

TABLES FOR CHAPTER 1

Additional Simulation Results

Table D.1 Average of the estimated variance of β_1 over the 1,000 replications using a rook type weighting matrix, N = 400, T=5.

		CF	CF_no1	True value	HACSC	SHAC	Cluster	Non-Robust	True value
ρ	ψ			CF					2SLS
	0.0	0.041	0.037	0.0386	0.082	0.068	0.086	0.069	0.0866
0.0	0.3	0.037	0.034	0.0352	0.076	0.062	0.078	0.063	0.0726
	0.7	0.035	0.032	0.0323	0.071	0.058	0.073	0.059	0.0706
	0.0	0.043	0.039	0.0428	0.087	0.071	0.089	0.072	0.0906
0.3	0.3	0.040	0.036	0.0364	0.079	0.065	0.082	0.066	0.079
	0.7	0.037	0.033	0.0359	0.074	0.061	0.076	0.062	0.0829
	0.0	0.062	0.055	0.0558	0.111	0.091	0.115	0.092	0.1092
0.7	0.3	0.057	0.051	0.0535	0.103	0.085	0.106	0.085	0.1118
	0.7	0.054	0.048	0.0488	0.096	0.079	0.099	0.080	0.0954

*True value computed as the variance of β_1 across the 1,000 replications.
All the numbers were multiplied by 100 for readability.

Table D.2 Average of the estimated variance of β_2 over the 1,000 replications using a rook type weighting matrix, N = 400, T=5.

		CF	CF_no1	True value	HACSC	SHAC	Cluster	Non-Robust	True value
ρ	ψ			CF					2SLS
	0.0	0.069	0.066	0.0724	0.080	0.066	0.083	0.067	0.0803
0.0	0.3	0.063	0.060	0.0644	0.074	0.061	0.076	0.061	0.0738
	0.7	0.060	0.057	0.0623	0.069	0.056	0.071	0.057	0.0704
	0.0	0.074	0.071	0.0756	0.086	0.070	0.089	0.071	0.0894
0.3	0.3	0.068	0.065	0.0761	0.078	0.065	0.081	0.065	0.0862
	0.7	0.063	0.060	0.0646	0.073	0.061	0.076	0.061	0.0793
	0.0	0.119	0.114	0.1237	0.131	0.108	0.136	0.109	0.1376
0.7	0.3	0.108	0.104	0.1125	0.120	0.099	0.125	0.100	0.1297
	0.7	0.101	0.097	0.1004	0.113	0.093	0.116	0.094	0.113

*True value computed as the variance of β_2 across the 1,000 replications.
All the numbers were multiplied by 100 for readability.

Table D.3 Average of the estimated variance of β_3 over the 1,000 replications using a rook type weighting matrix, N = 400, T=5.

		CF	CF_no1	True value	HACSC	SHAC	Cluster	Non-Robust	True value
ρ	ψ			CF					2SLS
	0.0	0.275	0.242	0.24	0.742	0.615	0.772	0.620	0.79
0.0	0.3	0.252	0.220	0.22	0.680	0.558	0.700	0.563	0.64
	0.7	0.239	0.206	0.21	0.631	0.522	0.654	0.526	0.63
	0.0	0.291	0.252	0.27	0.769	0.636	0.796	0.640	0.81
0.3	0.3	0.271	0.232	0.24	0.705	0.581	0.730	0.587	0.71
	0.7	0.254	0.215	0.23	0.660	0.545	0.681	0.549	0.72
	0.0	0.377	0.314	0.33	0.917	0.758	0.949	0.763	0.90
0.7	0.3	0.350	0.290	0.29	0.860	0.709	0.884	0.712	0.91
	0.7	0.329	0.270	0.29	0.794	0.657	0.824	0.662	0.79

*True value computed as the variance of β_3 across the 1,000 replications.

All the numbers were multiplied by 100 for readability.

CF_no1 refers to the HACSC estimator ignoring the first stage estimation using a CF approach.

Table D.4 Rejection probabilities for the null hypothesis $H_0 : \beta_1 = 0.7$ at a 5% of significance using a t-test over the 1,000 replications with a rook type weighting matrix, N = 400, T=5.

ρ	ψ	CF	CF_no1	HACSC	SHAC	Cluster	Non-Robust
	0.0	0.050	0.062	0.068	0.088	0.055	0.081
0.0	0.3	0.047	0.057	0.056	0.072	0.052	0.076
	0.7	0.043	0.054	0.053	0.068	0.046	0.068
	0.0	0.054	0.066	0.068	0.089	0.058	0.088
0.3	0.3	0.048	0.062	0.057	0.073	0.042	0.072
	0.7	0.045	0.067	0.059	0.083	0.060	0.083
	0.0	0.049	0.067	0.065	0.079	0.057	0.079
0.7	0.3	0.047	0.062	0.071	0.093	0.064	0.095
	0.7	0.051	0.064	0.050	0.070	0.040	0.070

Table D.5 Rejection probabilities for the null hypothesis $H_0 : \beta_2 = 0.6$ at a 5% of significance using a t-test over the 1,000 replications with a rook type weighting matrix, N = 400, T=5.

ρ	ψ	CF	CF_no1	HACSC	SHAC	Cluster	Non-Robust
	0.0	0.076	0.067	0.069	0.096	0.061	0.093
0.0	0.3	0.078	0.064	0.077	0.101	0.066	0.102
	0.7	0.079	0.072	0.079	0.104	0.074	0.105
	0.0	0.082	0.072	0.089	0.108	0.076	0.105
0.3	0.3	0.082	0.075	0.079	0.106	0.076	0.104
	0.7	0.081	0.069	0.089	0.119	0.078	0.108
	0.0	0.071	0.068	0.075	0.099	0.073	0.099
0.7	0.3	0.077	0.066	0.092	0.111	0.080	0.110
	0.7	0.069	0.064	0.063	0.080	0.053	0.077

APPENDIX E

FIGURES FOR CHAPTER 1

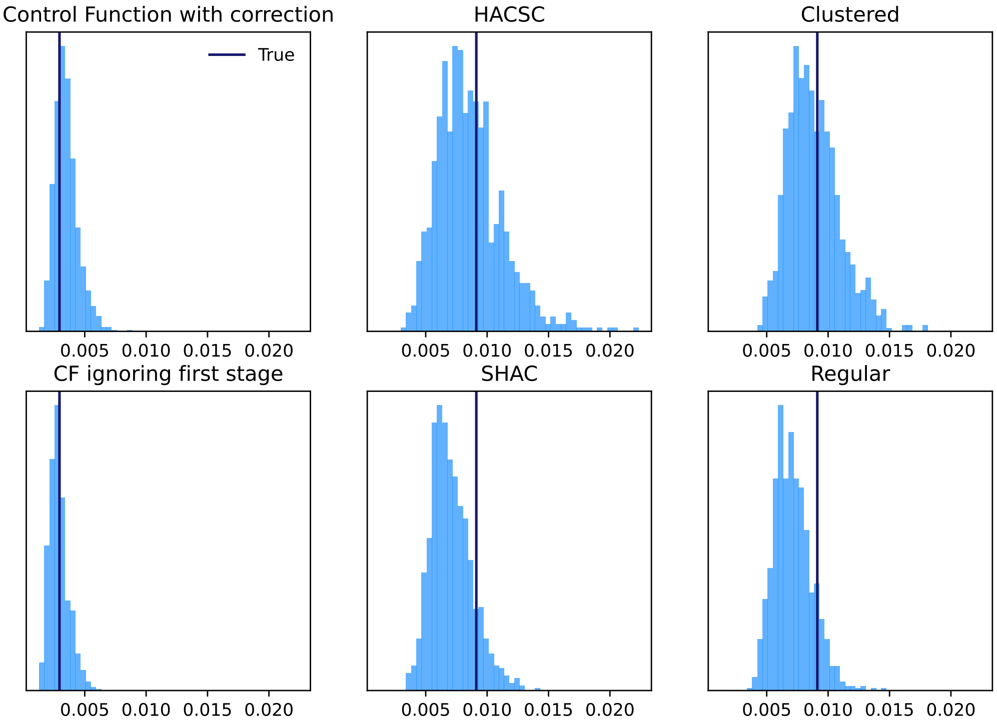


Figure E.1 Distribution of the computed variances of $\hat{\beta_3}$ obtained for the case with e following a spatial AR(1) process ($\rho = 0.7$), and a following an AR(1) ($\psi = 0.3$), $N = 400$, $T = 5$.
*True value computed as the variance of β_3 across the 1,000 replications.

APPENDIX F

PROOFS FOR CHAPTER 2

Proof of Proposition 2

The problem is to apply P2SLS to

$$\tilde{s}_{it} y_{it} = \tilde{s}_{it} a_{it} \theta + \tilde{s}_{it} \bar{z}_i \delta$$

using the instruments $z_{it} = (x_{1it} \ \ z_{2it} \ \ w_{1it} \ \ \mathfrak{z}_{2it})$ when $\tilde{s}_{it} = 1$ and where $a_{it} = (x_{1it} \ \ x_{2it} \ \ w_{1it} \ \ w_{2it})$ and $\bar{z}_i = (\bar{x}_{1i} \ \ \bar{z}_{2i} \ \ \bar{w}_{1i} \ \ \bar{\mathfrak{z}}_{2i})$. Note that the averages are taken for the cases where $\tilde{s}_{it} = 1$.

The first step is to orthogonalize the instrumental variables with respect to $\bar{z}_i$. I start by regressing z_{2it} on $\bar{z}_{2i}$. The associated coefficient will be:

$$\begin{aligned}
\hat{\gamma}_1 &= \left[\sum_{i=1}^{N}\sum_{t=1}^{T} \tilde{s}_{it} \bar{z}_{2i}' \bar{z}_{2i}\right]^{-1} \left[\sum_{i=1}^{N}\sum_{t=1}^{T} \tilde{s}_{it} \bar{z}_{2i}' z_{2it}\right] \\
&= \left[\sum_{i=1}^{N} \bar{z}_{2i}' \bar{z}_{2i} \sum_{t=1}^{T} \tilde{s}_{it}\right]^{-1} \left[\sum_{i=1} \bar{z}_{2i}' \sum_{t=1}^{T} \tilde{s}_{it} z_{2it}\right] \\
&= \left[\sum_{i=1}^{N} \bar{z}_{2i}' \bar{z}_{2i} T_i\right]^{-1} \left[\sum_{i=1} \bar{z}_{2i}' \bar{z}_{2i} T_i\right] = \mathbf{I}_l
\end{aligned}$$

And therefore the residuals from this regression will be $z_{2it} - \bar{z}_{2i} = \ddot{z}_{2it}$. Now we regress each of the remaining elements of $\bar{z}_{2i}$, i.e. $\bar{x}_{1t}, \bar{w}_{1i}$ and $\bar{\mathfrak{z}}_{2i}$ on $\bar{z}_{2i}$. Note that each set of residuals of these regressions will depend only on the i index, so denote them respectively by $f_i^{x_1}, f_i^{w_1}$ and $f_i^{\mathfrak{z}_2}$ and stack them into a vector f_i. Now we regress $\ddot{z}_{2it}$ on f_i and the associated coefficient will be:

$$\begin{aligned}
\hat{\gamma}_2 &= \left[\sum_{i=1}^{N}\sum_{t=1}^{T} \tilde{s}_{it} f_i' f_i\right]^{-1} \left[\sum_{i=1}^{N}\sum_{t=1}^{T} \tilde{s}_{it} f_i' \ddot{z}_{2it}\right] \\
&= \left[\sum_{i=1}^{N}\sum_{t=1}^{T} \tilde{s}_{it} f_i' f_i\right]^{-1} \left[\sum_{i=1} f_i' \sum_{t=1}^{T} \tilde{s}_{it} \ddot{z}_{2it}\right] = \mathbf{0}_{2k_1+l}
\end{aligned}$$

where in the last line I used the fact that the sum of deviations from the mean is equal to zero when $\tilde{s}_{it} = 1$. Therefore, after this orthogonalization of z_{2it} with respect to $\bar{z}_i$, the residuals are $\ddot{z}_{2it}$. Following very similar steps, it can be shown that after orthogonalizing the remaining elements of

z_{it} with respect to $\bar{z}_i$, the set of residuals will be $\ddot{z}_{it} = (\ddot{x}_{1it} \quad \ddot{z}_{2it} \quad \ddot{w}_{1it} \quad \ddot{\mathfrak{z}}_{2it})$. The problem then becomes to apply Pooled 2SLS to

$$\tilde{s}_{it} y_{it} = \tilde{s}_{it} a_{it} \theta$$

using the instruments $\ddot{z}_{it}$. The associated coefficient will be

$$\tilde{\theta} = \left[\left(\sum_{i=1} \sum_{t=1} \tilde{s}_{it} a'_{it} \ddot{z}_{it} \right) \left(\sum_{i=1} \sum_{t=1} \tilde{s}_{it} \ddot{z}'_{it} \ddot{z}_{it} \right)^{-1} \left(\sum_{i=1} \sum_{t=1} \tilde{s}_{it} \ddot{z}'_{it} a_{it} \right) \right]^{-1} \cdot$$
$$\left(\sum_{i=1} \sum_{t=1} \tilde{s}_{it} a'_{it} \ddot{z}_{it} \right) \left(\sum_{i=1} \sum_{t=1} \tilde{s}_{it} \ddot{z}'_{it} \ddot{z}_{it} \right)^{-1} \left(\sum_{i=1} \sum_{t=1} \tilde{s}_{it} \ddot{z}'_{it} y_{it} \right)$$

Now focusing on the first element of the square bracket matrix and noting that the following algebraic manipulation holds for the remaining of the terms in the above expression that do not contain demeaned variables, we have:

$$\begin{aligned} \sum_{i=1} \sum_{t=1} \tilde{s}_{it} a'_{it} \ddot{z}_{it} &= \sum_{i=1} \sum_{t=1} \tilde{s}_{it} a'_{it} \ddot{z}_{it} - \sum_{i=1} \bar{a}'_i \sum_{t=1} \tilde{s}_{it} (z_{it} - \bar{z}_i) \\ &= \sum_{i=1} \sum_{t=1} \tilde{s}_{it} a'_{it} \ddot{z}_{it} - \sum_{i=1} \sum_{t=1} \tilde{s}_{it} \bar{a}'_i \ddot{z}_{it} \\ &= \sum_{i=1} \sum_{t=1} \tilde{s}_{it} \ddot{a}'_{it} \ddot{z}_{it} \end{aligned}$$

where at the end of the first line I used again the fact that the sum of deviations from the mean for the cases for which $\tilde{s}_{it} = 0$. Therefore we have:

$$\begin{aligned} \tilde{\theta} &= \left[\left(\sum_{i=1} \sum_{t=1} \tilde{s}_{it} a'_{it} \ddot{z}_{it} \right) \left(\sum_{i=1} \sum_{t=1} \tilde{s}_{it} \ddot{z}'_{it} \ddot{z}_{it} \right)^{-1} \left(\sum_{i=1} \sum_{t=1} \tilde{s}_{it} \ddot{z}'_{it} a_{it} \right) \right]^{-1} \cdot \\ &\quad \left(\sum_{i=1} \sum_{t=1} \tilde{s}_{it} a'_{it} \ddot{z}_{it} \right) \left(\sum_{i=1} \sum_{t=1} \tilde{s}_{it} \ddot{z}'_{it} \ddot{z}_{it} \right)^{-1} \left(\sum_{i=1} \sum_{t=1} \tilde{s}_{it} \ddot{z}'_{it} y_{it} \right) \\ &= \left[\left(\sum_{i=1} \sum_{t=1} \tilde{s}_{it} \ddot{a}'_{it} \ddot{z}_{it} \right) \left(\sum_{i=1} \sum_{t=1} \tilde{s}_{it} \ddot{z}'_{it} \ddot{z}_{it} \right)^{-1} \left(\sum_{i=1} \sum_{t=1} \tilde{s}_{it} \ddot{z}'_{it} \ddot{a}_{it} \right) \right]^{-1} \cdot \\ &\quad \left(\sum_{i=1} \sum_{t=1} \tilde{s}_{it} \ddot{a}'_{it} \ddot{z}_{it} \right) \left(\sum_{i=1} \sum_{t=1} \tilde{s}_{it} \ddot{z}'_{it} \ddot{z}_{it} \right)^{-1} \left(\sum_{i=1} \sum_{t=1} \tilde{s}_{it} \ddot{z}'_{it} \ddot{y}_{it} \right) = \hat{\theta}_{CFE2SLS} \end{aligned}$$

□

APPENDIX G

TABLES FOR CHAPTER 2

Table G.1 Average bias, standard deviation and root mean squared error for β_1 and β_2 across the 1000 repetitions when the data is MCAR.

		β_1			β_2	
	Bias	S.D.	RMSE	Bias	S.D.	RMSE
Whole data	0.0002	0.0167	0.0167	-0.0021	0.0325	0.0326
Complete cases	0.0005	0.0251	0.0251	-0.0023	0.0510	0.0510
Proposed GMM	0.0009	0.0213	0.0213	-0.0025	0.0422	0.0422
Dummy variable	0.0014	0.0346	0.0346	-0.0012	0.0669	0.0669

Table G.2 Average bias, standard deviation and root mean squared error for β_1 and β_2 across the 1000 repetitions when the data is MAR.

		β_1			β_2	
	Bias	S.D.	RMSE	Bias	S.D.	RMSE
Whole data	0.0001	0.0164	0.0164	-0.0012	0.0321	0.0321
Complete cases	0.0010	0.0260	0.0260	-0.0020	0.0654	0.0654
Proposed GMM	0.0007	0.0214	0.0214	-0.0019	0.0520	0.0520
Dummy variable	0.0001	0.0395	0.0395	-0.0032	0.0894	0.0894

Table G.3 Average bias, standard deviation and root mean squared error for β_3 and β_4 across the 1000 repetitions when the data is MAR.

		β_3			β_4	
	Bias	S.D.	RMSE	Bias	S.D.	RMSE
Whole data	-0.0009	0.0244	0.0244	0.0004	0.0496	0.0496
Complete cases	-0.0010	0.0396	0.0396	0.0014	0.0772	0.0771
Proposed GMM	-0.0000	0.0318	0.0318	0.0028	0.0630	0.0630
Dummy variable	1.1034	0.1455	1.1130	0.3698	0.2049	0.4227

Table G.4 Average bias, standard deviation and root mean squared error for β_1 and β_2 across the 1000 repetitions when the data is MCAR and the error term follows a SAR(1) and N = 900.

		β_1			β_2	
	Bias	S.D.	RMSE	Bias	S.D.	RMSE
Whole data	0.0001	0.0182	0.0182	0.0000	0.0382	0.0382
Complete cases	-0.0000	0.0278	0.0278	-0.0008	0.0560	0.0560
Proposed GMM	0.0000	0.0225	0.0225	-0.0013	0.0467	0.0467
Dummy variable	0.0001	0.0375	0.0375	-0.0008	0.0760	0.0759

Table G.5 Average bias, standard deviation and root mean squared error for β_3 and β_4 across the 1000 repetitions when the data is MCAR and the error term follows a SAR(1) and N = 900.

		β_3			β_4	
	Bias	S.D.	RMSE	Bias	S.D.	RMSE
Whole data	-0.0012	0.0260	0.0260	-0.0006	0.0565	0.0565
Complete cases	-0.0018	0.0402	0.0402	-0.0006	0.0821	0.0821
Proposed GMM	-0.0002	0.0330	0.0330	0.0013	0.0674	0.0674
Dummy variable	0.9796	0.1366	0.9891	0.3237	0.1942	0.3774

Table G.6 Average bias, standard deviation and root mean squared error for β_1 and β_2 across the 1000 repetitions when the data is MCAR and the error term follows a SAR(1) and N = 400.

		β_1			β_2	
	Bias	S.D.	RMSE	Bias	S.D.	RMSE
Whole data	0.0001	0.0278	0.0278	0.0006	0.0582	0.0582
Complete cases	0.0007	0.0417	0.0417	-0.0005	0.0838	0.0838
Proposed GMM	0.0009	0.0342	0.0342	0.0007	0.0713	0.0713
Dummy variable	0.0006	0.0538	0.0537	0.0030	0.1123	0.1123

Table G.7 Average bias, standard deviation and root mean squared error for β_3 and β_4 across the 1000 repetitions when the data is MCAR and the error term follows a SAR(1) and N = 400.

		β_3			β_4	
	Bias	S.D.	RMSE	Bias	S.D.	RMSE
Whole data	-0.0037	0.0387	0.0389	-0.0036	0.0820	0.0820
Complete cases	-0.0037	0.0598	0.0599	-0.0062	0.1267	0.1268
Proposed GMM	-0.0009	0.0485	0.0484	-0.0043	0.1035	0.1035
Dummy variable	0.9430	0.1944	0.9628	0.3006	0.2817	0.4118

Table G.8 Average bias, standard deviation and root mean squared error for β_1 and β_2 across the 1000 repetitions when the missingness depends on x_1 and c_i.

		β_1			β_2	
	Bias	S.D.	RMSE	Bias	S.D.	RMSE
Whole data	-0.0005	0.0252	0.0252	0.0025	0.0479	0.0479
Complete cases	-0.0013	0.0367	0.0367	0.0059	0.0801	0.0803
Proposed GMM	-0.0012	0.0303	0.0303	0.0030	0.0656	0.0656
Dummy variable	-0.0005	0.0526	0.0526	-0.0044	0.1105	0.1106

Table G.9 Average bias, standard deviation and root mean squared error for β_3 and β_4 across the 1000 repetitions when the missingness depends on x_1 and c_i.

		β_3			β_4	
	Bias	S.D.	RMSE	Bias	S.D.	RMSE
Whole data	-0.0005	0.0338	0.0338	0.0032	0.0691	0.0691
Complete cases	-0.0020	0.0506	0.0507	0.0066	0.1035	0.1037
Proposed GMM	-0.0006	0.0409	0.0409	0.0059	0.0862	0.0864
Dummy variable	1.0765	0.2408	1.1031	0.3505	0.2845	0.4514

APPENDIX H

FIGURES FOR CHAPTER 2

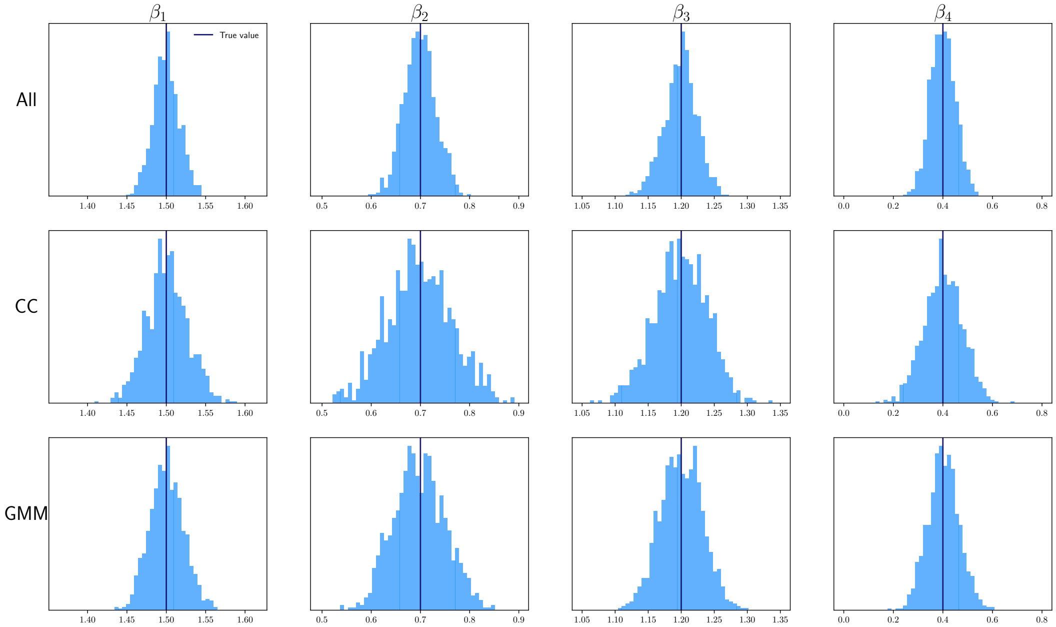


Figure H.1 Distribution of estimated coefficients across the 1000 Monte-Carlo repetitions when the data is MAR.

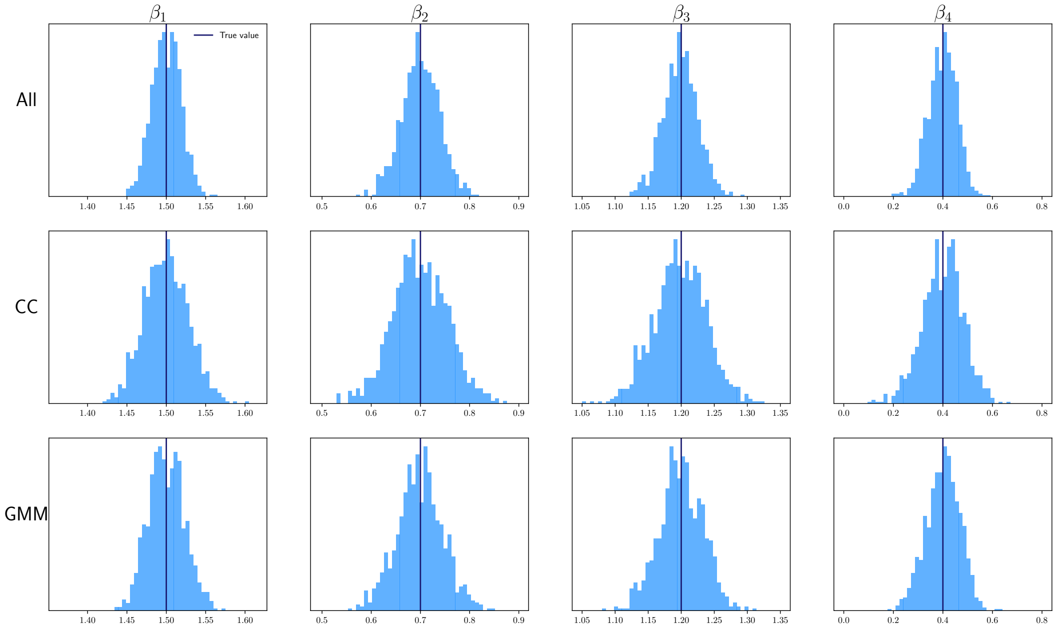


Figure H.2 Distribution of estimated coefficients across the 1000 Monte-Carlo repetitions when the data is MCAR and the error term follows a SAR(1) process with $N = 900$.

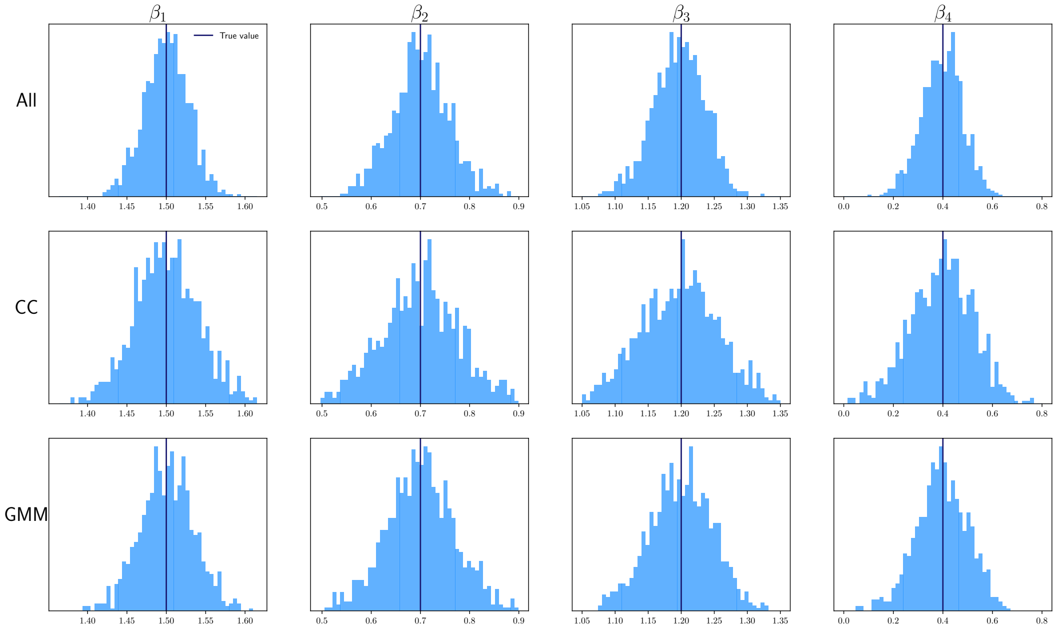


Figure H.3 Distribution of estimated coefficients across the 1000 Monte-Carlo repetitions when the data is MCAR and the error term follows a SAR(1) process with $N = 400$.

APPENDIX I

PROOFS FOR CHAPTER 3

Proof of Proposition 1

Because $x_{ijt} = (x_{1ijt} \quad x_{2ijt})$ where x_{2ijt} is endogenous with, we need to add the instrumental variables. The CRE IV estimator in this is is to apply Pooled Two-Stage Least Squares (P2SLS) to the following equation:

$$\tilde{y}_{ijt} = \tilde{x}_{ijt}\beta + \tilde{\bar{x}}_{1\cdots} + \tilde{\bar{z}}_{2\cdots}$$

using the IV's $\tilde{z}_{2ijt}$ where $\tilde{y}_{ijt} = y_{ijt} - \tilde{\theta}_1\bar{y}_{i\cdot\cdot} - \tilde{\theta}_2\bar{y}_{\cdot j\cdot} - \tilde{\theta}_3\bar{y}_{\cdot\cdot t} - \tilde{\theta}_4\bar{y}_{\cdots}$ and similarly for each variable.

First note that $\tilde{\bar{x}}_{1i\cdot\cdot} = (1-\tilde{\theta}_1)\bar{x}_{1i\cdot\cdot}$, $\tilde{\bar{x}}_{1\cdot j\cdot} = (1-\tilde{\theta}_2)\bar{x}_{1\cdot j\cdot}$, $\tilde{\bar{x}}_{1\cdot\cdot t} = (1-\tilde{\theta}_3)\bar{x}_{1\cdot\cdot t}$ and $\tilde{\bar{x}}_{1\cdots} = (1-\tilde{\theta}_4)\bar{x}_{1\cdots}$. As a first step, we orthogonalize the exogenous and instrumental variables $(\tilde{x}_{1ijt} \quad \tilde{z}_{2ijt})$ with respect to $\tilde{\bar{z}}_{ijt} = (\tilde{\bar{x}}_{1i\cdot\cdot} \quad \tilde{\bar{x}}_{1\cdot j\cdot} \quad \tilde{\bar{x}}_{1\cdot\cdot t} \quad \tilde{\bar{z}}_{2i\cdot\cdot} \quad \tilde{\bar{z}}_{2\cdot j\cdot} \quad \tilde{\bar{z}}_{2\cdot\cdot t})$.

a. $\tilde{x}_{1ijt}$ on $(1 \quad \tilde{\bar{z}}_{ijt})$, where the 1 represent the constant term. Applying the Frish-Waugh-Lovell (FWL) theorem, to obtain the correct residuals, this is equivalent to regress $\tilde{x}_{1ijt} - \bar{x}_{1\cdots}$ on $\tilde{\bar{z}}_{ijt} - \bar{z}_{\cdots}$ (i.e. no constant term). The coefficient associated with this regression will have the typical form of $(X'X)^{-1}(X'y)$. Consider the first matrix, which will have the following structure:

$$\begin{pmatrix} \sum\limits_i \sum\limits_j \sum\limits_t (\tilde{\bar{x}}_{1i\cdot\cdot} - \bar{x}_{1\cdots})'(\tilde{\bar{x}}_{1i\cdot\cdot} - \bar{x}_{1\cdots}) & \sum\limits_i \sum\limits_j \sum\limits_t (\tilde{\bar{x}}_{1i\cdot\cdot} - \bar{x}_{1\cdots})'(\tilde{\bar{x}}_{1\cdot j\cdot} - \bar{x}_{1\cdots}) & \cdots \\ \sum\limits_i \sum\limits_j \sum\limits_t (\tilde{\bar{x}}_{1\cdot j\cdot} - \bar{x}_{1\cdots})'(\tilde{\bar{x}}_{1i\cdot\cdot} - \bar{x}_{1\cdots}) & \sum\limits_i \sum\limits_j \sum\limits_t (\tilde{\bar{x}}_{1\cdot j\cdot} - \bar{x}_{1\cdots})'(\tilde{\bar{x}}_{1\cdot j\cdot} - \bar{x}_{1\cdots}) & \\ \vdots & & \ddots \end{pmatrix}^{-1}$$

Each term off the diagonal that has a cross product of different indices (e.g. $\tilde{\bar{x}}_{1i\cdot\cdot}$ and $\tilde{\bar{x}}_{1\cdot j\cdot}$) can be treated as follows:

$$\sum_i \sum_j \sum_t (\tilde{\bar{x}}_{1i\cdot\cdot} - \bar{x}_{1\cdots})'(\tilde{\bar{x}}_{1i\cdot\cdot} - \bar{x}_{1\cdots}) = T \cdot \sum_i (\tilde{\bar{x}}_{1i\cdot\cdot} - \bar{x}_{1\cdots})' \sum_j (\tilde{\bar{x}}_{1\cdot j\cdot} - \bar{x}_{1\cdots}) = 0$$

Therefore, each pair of these regressors is orthogonal to each other in sample. For those pairs of independent variables that have a common index (e.g. regressing $\tilde{\bar{x}}_{1ijt}\,\tilde{\bar{x}}_{1i\cdot\cdot}$ and $\tilde{\bar{z}}_{2i\cdot\cdot}$), using the fact that each variable as been centered around their overall mean and applying the FWL theorem, it can be shown that after partialling out the variable that is not associated with the dependent variable (in this case $\bar{\tilde{z}}_{2i\cdot\cdot}$), we will recover the same coefficient as if we ran $\tilde{x}_{1ijt}$ on $\bar{\tilde{x}}_{1i\cdot\cdot}$ directly.

Now I show that the coefficients associated with each element of x_1 of this regression is equal to an identity matrix of size k_1 and 0 for the elements of z_2. For example, the parameter vector associated with $\bar{\tilde{x}}_{1i\cdot\cdot} - \bar{x}_{1\cdots}$ is:

$$\begin{aligned}
\hat{\pi}_{\bar{\tilde{x}}_{1i\cdot\cdot}} &= \left[\sum_i \sum_j \sum_t (\bar{\tilde{x}}_{1i\cdot\cdot} - \bar{x}_{1\cdots})'(\bar{\tilde{x}}_{1i\cdot\cdot} - \bar{x}_{1\cdots})\right]^{1} \left[\sum_i \sum_j \sum_t (\bar{\tilde{x}}_{1i\cdot\cdot} - \bar{x}_{1\cdots})'(\tilde{x}_{1ijt} - \bar{x}_{1\cdots})\right] \\
&= \left[N_2 \cdot T \cdot \sum_i (\bar{\tilde{x}}_{1i\cdot\cdot} - \bar{x}_{1\cdots})'(\bar{\tilde{x}}_{1i\cdot\cdot} - \bar{x}_{1\cdots})\right]^{1} \left[\sum_i (\bar{\tilde{x}}_{1i\cdot\cdot} - \bar{x}_{1\cdots})' \sum_j \sum_t (\tilde{x}_{1ijt} - \bar{x}_{1\cdots})\right] \\
&= \left[\sum_i (\bar{\tilde{x}}_{1i\cdot\cdot} - \bar{x}_{1\cdots})'(\bar{\tilde{x}}_{1i\cdot\cdot} - \bar{x}_{1\cdots})\right]^{1} \left[\sum_i (\bar{\tilde{x}}_{1i\cdot\cdot} - \bar{x}_{1\cdots})' \frac{1}{N_2 T} \sum_j \sum_t (\tilde{x}_{1ijt} - \bar{x}_{1\cdots})\right] \\
&= \left[\sum_i (\bar{\tilde{x}}_{1i\cdot\cdot} - \bar{x}_{1\cdots})'(\bar{\tilde{x}}_{1i\cdot\cdot} - \bar{x}_{1\cdots})\right]^{1} \left[\sum_i (\bar{\tilde{x}}_{1i\cdot\cdot} - \bar{x}_{1\cdots})'(\bar{\tilde{x}}_{1i\cdot\cdot} - \bar{x}_{1\cdots})\right] = \mathbf{I}_{k_1}
\end{aligned}$$

Therefore each explanatory variable associated with the averages of x_1 will have a coefficient vector equal to an identity matrix. On the other hand, it can be shown that the coefficients associated with with the $\bar{\tilde{z}}_2$ variables are 0 using the fact that we will obtain sums of vectors that are deviated from their overall mean.

b. $\tilde{z}_{2ijt}$ on $(1 \quad \bar{\tilde{z}}_{ijt})$. Using very similar arguments as in the previous step, the coefficients associated with the $\bar{\tilde{z}}_2$ variables will be identity matrices of size l and the ones associated with $\bar{\tilde{x}}_1$ will be 0. Given this, after partialling out $\bar{\tilde{z}}_{ijt}$ the associated residuals with be:

$$\ddot{x}_{1ijt} = x_{1ijt} - \bar{x}_{1i\cdot\cdot} - \bar{x}_{1\cdot j\cdot} - \bar{x}_{1\cdot\cdot t} + 2\bar{x}_{1\cdots}$$

$$\ddot{z}_{2ijt} = z_{2ijt} - \bar{z}_{2i\cdot\cdot} - \bar{z}_{2\cdot j\cdot} - \bar{z}_{2\cdot\cdot t} + 2\bar{z}_{2\cdots}$$

The problem has become now to apply P2SLS to

$$\tilde{y}_{ijt} = \tilde{x}_{ijt}$$

using IV's $\ddot{z}_{2ijt} = z_{2ijt} - \bar{z}_{2i\cdot\cdot} - \bar{z}_{2\cdot j\cdot} - \bar{z}_{2\cdot\cdot t} + 2\bar{z}_{2\cdots}$. From this, note that

$$\hat{\beta}_{2SLS} = \left[\left(\sum_i\sum_j\sum_t \tilde{x}'_{ijt}\ddot{z}_{ijt}\right)\left(\sum_i\sum_j\sum_t \ddot{z}'_{ijt}\ddot{z}_{ijt}\right)\left(\sum_i\sum_j\sum_t \ddot{z}'_{ijt}\tilde{x}_{ijt}\right)\right]^{-1} \cdot$$
$$\left[\left(\sum_i\sum_j\sum_t \tilde{x}'_{ijt}\ddot{z}_{ijt}\right)\left(\sum_i\sum_j\sum_t \ddot{z}'_{ijt}\ddot{z}_{ijt}\right)\left(\sum_i\sum_j\sum_t \ddot{z}'_{ijt}\tilde{y}_{ijt}\right)\right]$$

Note that

$$\begin{aligned}\sum_i\sum_j\sum_t \tilde{x}'_{ijt}\ddot{z}_{ijt} = {} & \sum_i\sum_j\sum_t x'_{ijt}\ddot{z}_{ijt} - \sum_i\sum_j\sum_t \tilde{\theta}_1\bar{x}_{i\cdot\cdot}(z_{2ijt} - \bar{z}_{2i\cdot\cdot} - \bar{z}_{2\cdot j\cdot} - \bar{z}_{2\cdot\cdot t} + 2\bar{z}_{2\cdots}) \\ & - \sum_i\sum_j\sum_t \tilde{\theta}_2\bar{x}'_{\cdot j\cdot}(z_{2ijt} - \bar{z}_{2i\cdot\cdot} - \bar{z}_{2\cdot j\cdot} - \bar{z}_{2\cdot\cdot t} + 2\bar{z}_{2\cdots}) \\ & - \sum_i\sum_j\sum_t \tilde{\theta}_3\bar{x}'_{\cdot\cdot t}(z_{2ijt} - \bar{z}_{2i\cdot\cdot} - \bar{z}_{2\cdot j\cdot} - \bar{z}_{2\cdot\cdot t} + 2\bar{z}_{2\cdots}) \\ & - \sum_i\sum_j\sum_t \tilde{\theta}_4\bar{x}'_{\cdots}(z_{2ijt} - \bar{z}_{2i\cdot\cdot} - \bar{z}_{2\cdot j\cdot} - \bar{z}_{2\cdot\cdot t} + 2\bar{z}_{2\cdots})\end{aligned}$$

Focusing on the last term of the first line from the previous expression,

$$\begin{aligned}& \sum_i\sum_j\sum_t \tilde{\theta}_1\bar{x}'_{i\cdot\cdot}(z_{ijt} - \bar{z}_{i\cdot\cdot} - \bar{z}_{\cdot j\cdot} - \bar{z}_{\cdot\cdot t} + \bar{z}_{\cdots}) \\ = & \sum_i \tilde{\theta}_1\bar{x}'_{i\cdot\cdot}\sum_j\sum_t(z_{ijt} - \bar{z}_{i\cdot\cdot} - \bar{z}_{\cdot j\cdot} - \bar{z}_{\cdot\cdot t} + \bar{z}_{\cdots}) \\ = & \sum_i \tilde{\theta}_1\bar{x}'_{i\cdot\cdot}(N_2T\bar{z}_{i\cdot\cdot} - N_2T\bar{z}_{i\cdot\cdot} - N_2T\bar{z}_{\cdots} - N_2T\bar{z}_{\cdots} + N_2T\bar{z}_{\cdots}) \\ = & \sum_i \bar{x}'_{i\cdot\cdot}\sum_j\sum_t(z_{ijt} - \bar{z}_{i\cdot\cdot} - \bar{z}_{\cdot j\cdot} - \bar{z}_{\cdot\cdot t} + \bar{z}_{\cdots}) \\ = & \sum_i\sum_j\sum_t \bar{x}'_{i\cdot\cdot}\ddot{z}_{ijt}\end{aligned}$$

where we used the fact that the terms in parenthesis in the third line add up to zero. Using similar arguments for the rest of the expression, we can easily show that $\sum_i\sum_j\sum_t \tilde{x}'_{ijt}\ddot{z}_{ijt} = \sum_i\sum_j\sum_t \ddot{x}'_{ijt}\ddot{z}_{ijt}$

And applying the same logic to the rest of the terms of β_{2SLS}, it follows that

$$
\begin{aligned}
\hat{\beta}_{2SLS} = & \left[\left(\sum_i\sum_j\sum_t \tilde{x}'_{ijt}\ddot{z}_{ijt}\right)\left(\sum_i\sum_j\sum_t \ddot{z}'_{ijt}\ddot{z}_{ijt}\right)\left(\sum_i\sum_j\sum_t \ddot{z}'_{ijt}\tilde{x}_{ijt}\right)\right]^{-1} \cdot \\
& \left[\left(\sum_i\sum_j\sum_t \tilde{x}'_{ijt}\ddot{z}_{ijt}\right)\left(\sum_i\sum_j\sum_t \ddot{z}'_{ijt}\ddot{z}_{ijt}\right)\left(\sum_i\sum_j\sum_t \ddot{z}'_{ijt}\tilde{y}_{ijt}\right)\right] \\
= & \left[\left(\sum_i\sum_j\sum_t \ddot{x}'_{ijt}\ddot{z}_{ijt}\right)\left(\sum_i\sum_j\sum_t \ddot{z}'_{ijt}\ddot{z}_{ijt}\right)\left(\sum_i\sum_j\sum_t \ddot{z}'_{ijt}\ddot{x}_{ijt}\right)\right]^{-1} \cdot \\
& \left[\left(\sum_i\sum_j\sum_t \ddot{x}'_{ijt}\ddot{z}_{ijt}\right)\left(\sum_i\sum_j\sum_t \ddot{z}'_{ijt}\ddot{z}_{ijt}\right)\left(\sum_i\sum_j\sum_t \ddot{z}'_{ijt}\ddot{y}_{ijt}\right)\right] = \hat{\beta}_{FE2SLS}
\end{aligned}
$$

□

Proof of Proposition 2

For notation simplicity, I will prove the case of P2SLS, however similar ideas can be applied for a GLS type transformation. We want to show that applying P2SLS to

$$y_{ijt} = x_{1ijt}\beta_1 + x_{2ijt}\beta_2 + \bar{x}_{1ijt}\gamma_1 + \bar{x}_{2ijt}\gamma_2 = x_{ijt}\beta + \bar{x}_{ijt}\gamma$$

using IV's $(z_{2ijt} \ \ \bar{z}_{ijt})$ and where $\bar{x}_{ijt} = (\bar{x}_{i\cdot\cdot} \ \ \bar{x}_{\cdot j\cdot} \ \ \bar{x}_{\cdot\cdot t})$ (and similarly for other variables) yields the same β as $\hat{\beta}_{FE2SLS}$. To show the result, I follow these steps:

1. Orthogonalize with respect to $\bar{x}_{1ijt}$ the IV's and the exogenous variables $(x_{1ijt} \ \ z_{2ijt})$.

2. Orthogonalize with respect to $\bar{z}_{2ijt}$ in the first stage equation.

3. I use the FWL theorem to show the equivalence.

Step 1

a. Regress z_{2ijt} on $\bar{x}_{1ijt} = (1 \ \ \bar{x}_{1i\cdot\cdot} \ \ \bar{x}_{1\cdot j\cdot} \ \ \bar{x}_{1\cdot\cdot t})$ Equivalently, applying the FWL and to obtain the correct residuals, we can regress $z_{2ijt} - \bar{z}_{2\cdots}$ on $[(\bar{x}_{1i\cdot\cdot} - \bar{x}_{1\cdots}) \ \ (\bar{x}_{1\cdot j\cdot} - \bar{x}_{1\cdots}) \ \ (\bar{x}_{1\cdot\cdot t} - \bar{x}_{1\cdots})]$.

   The residuals from this regression will be

$$z_{2ijt} - \bar{z}_{2\cdots} - (\bar{x}_{1i\cdot\cdot} - \bar{x}_{1\cdots})\eta_1 - (\bar{x}_{1\cdot j\cdot} - \bar{x}_{1\cdots})\eta_2 - (\bar{x}_{1\cdot\cdot t} - \bar{x}_{1\cdots})\eta_3 = m_{ijt}$$

First note that the regressors are orthogonal in sample. For example:

$$\sum_i \sum_j \sum_t (\bar{x}_{1i\cdot\cdot} - \bar{x}_{1\cdots})'(\bar{x}_{1\cdot j\cdot} - \bar{x}_{1\cdots}) = \sum_i (\bar{x}_{1i\cdot\cdot} - \bar{x}_{1\cdots})' \sum_j (\bar{x}_{1\cdot j\cdot} - \bar{x}_{1\cdots}) = 0$$

since we are subtracting the overall mean to both sums of vectors. Therefore, we can find each η_s by regressing the dependent variable on each regressor individually and therefore:

$$\hat{\eta}_1 = \left[\sum_i \sum_j \sum_t (\bar{x}_{1i\cdot\cdot} - \bar{x}_{1\cdots})'(\bar{x}_{1i\cdot\cdot} - \bar{x}_{1\cdots})\right]^{-1} \left[\sum_i \sum_j \sum_t (\bar{x}_{1i\cdot\cdot} - \bar{x}_{1\cdots})'(z_{2ijt} - \bar{z}_{2\cdots})\right]$$

$$\hat{\eta}_2 = \left[\sum_i \sum_j \sum_t (\bar{x}_{1\cdot j\cdot} - \bar{x}_{1\cdots})'(\bar{x}_{1\cdot j\cdot} - \bar{x}_{1\cdots})\right]^{-1} \left[\sum_i \sum_j \sum_t (\bar{x}_{1\cdot j\cdot} - \bar{x}_{1\cdots})'(z_{2ijt} - \bar{z}_{2\cdots})\right]$$

$$\hat{\eta}_3 = \left[\sum_i \sum_j \sum_t (\bar{x}_{1\cdot\cdot t} - \bar{x}_{1\cdots})'(\bar{x}_{1\cdot\cdot t} - \bar{x}_{1\cdots})\right]^{-1} \left[\sum_i \sum_j \sum_t (\bar{x}_{1\cdot\cdot t} - \bar{x}_{1\cdots})'(z_{2ijt} - \bar{z}_{2\cdots})\right]$$

Note that each of the coefficients can be rewritten, for example:

$$\hat{\eta}_1 = \left[\sum_i (\bar{x}_{1i\cdot\cdot} - \bar{x}_{1\cdots})'(\bar{x}_{1i\cdot\cdot} - \bar{x}_{1\cdots})\right]^{-1} \left[\sum_i (\bar{x}_{1i\cdot\cdot} - \bar{x}_{1\cdots})'(\bar{z}_{2i\cdot\cdot} - \bar{z}_{2\cdots})\right]$$

b. $\bar{z}_{2ijt}$ on $\bar{x}_{1ijt} = (1 \;\; \bar{x}_{1i\cdot\cdot} \;\; \bar{x}_{1\cdot j\cdot} \;\; \bar{x}_{1\cdot\cdot t} \;\; \bar{x}_{1\cdots})$, where $\bar{z}_{2ijt} = (\bar{z}_{2i\cdot\cdot} \;\; \bar{z}_{2\cdot j\cdot} \;\; \bar{z}_{2\cdot\cdot t} \;\; \bar{z}_{2\cdots})$. Similarly to the previous case, we can regress $(\bar{z}_{2ijt} - \bar{z}_{2\cdots})$ on $[(\bar{x}_{1i\cdot\cdot} - \bar{x}_{1\cdots}) \;\; (\bar{x}_{1\cdot j\cdot} - \bar{x}_{1\cdots}) \;\; (\bar{x}_{1\cdot\cdot t} - \bar{x}_{1\cdots})]$. Because the regressors are orthogonal in sample, we can again obtain the coefficients by running individual regressions.

a) Consider the regression of $\bar{z}_{2i\cdot\cdot} - \bar{z}_{2\cdots}$ on $\bar{x}_{1i\cdot\cdot} - \bar{x}_{1\cdots}$. The coefficient will be

$$\begin{aligned}\hat{\eta}_4 &= \left[\sum_i \sum_j \sum_t (\bar{x}_{1i\cdot\cdot} - \bar{x}_{1\cdots})'(\bar{x}_{1i\cdot\cdot} - \bar{x}_{1\cdots})\right]^{-1} \left[\sum_i \sum_j \sum_t (\bar{x}_{1i\cdot\cdot} - \bar{x}_{1\cdots})'(\bar{z}_{2i\cdot\cdot} - \bar{z}_{2\cdots})\right] \\ &= \left[\sum_i (\bar{x}_{1i\cdot\cdot} - \bar{x}_{1\cdots})'(\bar{x}_{1i\cdot\cdot} - \bar{x}_{1\cdots})\right]^{-1} \left[\sum_i (\bar{x}_{1i\cdot\cdot} - \bar{x}_{1\cdots})'(\bar{z}_{2i\cdot\cdot} - \bar{z}_{2\cdots})\right] = \hat{\eta}_1\end{aligned}$$

Similarly, we can show that the coefficients of $\bar{z}_{2\cdot j\cdot} - \bar{z}_{2\cdots}$ on $\bar{x}_{1\cdot j\cdot} - \bar{x}_{1\cdots}$ ($\hat{\eta}_5$)) and $\bar{z}_{2\cdot\cdot t} - \bar{z}_{2\cdots}$ on $\bar{x}_{1\cdot\cdot t} - \bar{x}_{1\cdots}$ ($\hat{\eta}_5$)) will be equal to ($\hat{\eta}_2$)) and ($\hat{\eta}_3$)) respectively.

b) Consider now the cases of "cross terms", i.e. averages in one dimension on variables averaged over a different dimension. For example, if we regress $\bar{z}_{2\cdot j\cdot} - \bar{z}_{2\cdots}$ on $\bar{x}_{1i\cdot\cdot} - \bar{x}_{1\cdots}$, the coefficient, say ζ_1 will be:

$$\begin{aligned}
\hat{\zeta}_1 &= \left[\sum_i \sum_j \sum_t (\bar{x}_{1i\cdot\cdot} - \bar{x}_{1\cdots})'(\bar{x}_{1i\cdot\cdot} - \bar{x}_{1\cdots})\right]^{-1} \left[\sum_i \sum_j \sum_t (\bar{x}_{1i\cdot\cdot} - \bar{x}_{1\cdots})'(\bar{z}_{2\cdot j\cdot} - \bar{z}_{2\cdots})\right] \\
&= \left[N_2 \sum_i (\bar{x}_{1i\cdot\cdot} - \bar{x}_{1\cdots})'(\bar{x}_{1i\cdot\cdot} - \bar{x}_{1\cdots})\right]^{-1} \left[\sum_i (\bar{x}_{1i\cdot\cdot} - \bar{x}_{1\cdots})' \sum_j (\bar{z}_{2i\cdot\cdot} - \bar{z}_{2\cdots})\right] = 0
\end{aligned}$$

since the sum of deviations from the overall mean add up to 0. Similarly, we can show that all the coefficients from the "cross terms" are 0. Therefore, the residuals from this stage will be

$$\begin{aligned}
v_i &= \bar{z}_{2i\cdot\cdot} - \bar{z}_{2\cdots} - (\bar{x}_{1i\cdot\cdot} - \bar{x}_{1\cdots})\hat{\eta}_4 \\
v_j &= \bar{z}_{2\cdot j\cdot} - \bar{z}_{2\cdots} - (\bar{x}_{1\cdot j\cdot} - \bar{x}_{1\cdots})\hat{\eta}_5 \\
v_t &= \bar{z}_{2\cdot\cdot t} - \bar{z}_{2\cdots} - (\bar{x}_{1\cdot\cdot t} - \bar{x}_{1\cdots})\hat{\eta}_6
\end{aligned}$$

c. $x_{1itj} - \bar{x}_{1\cdots}$ on $\bar{x}_{1ijt} - \bar{x}_{1\cdots}$. The residuals from this regression will be

$$l_{ijt} = x_{1ijt} - \bar{x}_1(\bar{x}_{1i\cdot\cdot} - \bar{x}_{1\cdots})\hat{\varepsilon}_1 - (\bar{x}_{1\cdot j\cdot} - \bar{x}_{1\cdots})\hat{\varepsilon}_2 - (\bar{x}_{1\cdot\cdot t} - \bar{x}_{1\cdots})\hat{\varepsilon}_3$$

Note that

$$\begin{aligned}
\hat{\varepsilon}_1 &= \left[\sum_i \sum_j \sum_t (\bar{x}_{1i\cdot\cdot} - \bar{x}_{1\cdots})'(\bar{x}_{1i\cdot\cdot} - \bar{x}_{1\cdots})\right]^{1} \left[\sum_i \sum_j \sum_t (\bar{x}_{1i\cdot\cdot} - \bar{x}_{1\cdots})'(x_{1ijt} - \bar{x}_{1\cdots})\right] \\
&= \left[N_2 \cdot T \cdot \sum_i (\bar{x}_{1i\cdot\cdot} - \bar{x}_{1\cdots})'(\bar{x}_{1i\cdot\cdot} - \bar{x}_{1\cdots})\right]^{1} \left[\sum_i (\bar{x}_{1i\cdot\cdot} - \bar{x}_{1\cdots})' \sum_j \sum_t (x_{1ijt} - \bar{x}_{1\cdots})\right] \\
&= \left[\sum_i (\bar{x}_{1i\cdot\cdot} - \bar{x}_{1\cdots})'(\bar{x}_{1i\cdot\cdot} - \bar{x}_{1\cdots})\right]^{1} \left[\sum_i (\bar{x}_{1i\cdot\cdot} - \bar{x}_{1\cdots})' \frac{1}{N_2 T} \sum_j \sum_t (x_{1ijt} - \bar{x}_{1\cdots})\right] \\
&= \left[\sum_i (\bar{x}_{1i\cdot\cdot} - \bar{x}_{1\cdots})'(\bar{x}_{1i\cdot\cdot} - \bar{x}_{1\cdots})\right]^{1} \left[\sum_i (\bar{x}_{1i\cdot\cdot} - \bar{x}_{1\cdots})'(\bar{x}_{1i\cdot\cdot} - \bar{x}_{1\cdots})\right] = \mathbf{I}_{k_1}
\end{aligned}$$

and similarly we can show that $\hat{\varepsilon}_2$ and $\hat{\varepsilon}_3$ are also identity matrices. Therefore, $l_{ijt} = x_{1ijt} - \bar{x}_{1i\cdot\cdot} - \bar{x}_{1\cdot j\cdot} - \bar{x}_{1\cdot\cdot t} + 2\bar{x}_{1\cdots} = \ddot{x}_{1ijt}$. After this orthogonalization, the problem becomes to apply P2SLS to

$$y_{ijt} = \ddot{x}_{1ijt}\beta_1 + x_{2ijt}\beta 2 + \bar{x}_{2ijt}\gamma_2$$

using IV's $(m_{ijt} \;\; v_i \;\; v_j \;\; v_t)$.

Step 2

Now I partial out v_i, v_j, v_t in the first stage equation, which are the residuals associated with $\bar{z}_{2i\cdot\cdot}, \bar{z}_{2\cdot j\cdot}, \bar{z}_{2\cdot\cdot t}$ respectively. Note that based on their definitions and because $\hat{\eta}_1 = \hat{\eta}_4$, $\hat{\eta}_2 = \hat{\eta}_5$ and $\hat{\eta}_3 = \hat{\eta}_6$ and following a procedure similar to Step 1.a, it can be shown that v_i, v_j, v_t are orthogonal to each other in sample.

1. m_{ijt} on $(1 \;\; v_i \;\; v_j \;\; v_t)$. If we let $m_{ijt} = v_i\tilde{\eta}_1 + v_j\tilde{\eta}_2 + v_t\tilde{\eta}_3$, then

$$\begin{aligned}\tilde{\eta}_1 &= \left[\sum_i \sum_j \sum_t v_i' v_i\right]^{-1} \left[\sum_i \sum_j \sum_t v_i' v_{ijt}\right] \\ &= \left[\sum_i v_i' v_i\right]^{-1} \left[\sum_i v_i' \bar{m}_{i\cdot\cdot}\right]\end{aligned}$$

Note that $m_{i\cdot\cdot} = (\bar{z}_{2i\cdot\cdot} - \bar{z}_{2\cdot\cdot\cdot}) - (\bar{x}_{1\cdot\cdot} - \bar{x}_{1\cdot\cdot\cdot})$ and because $\hat{\eta}_1 = \hat{\eta}_4$, it follows that $\tilde{\eta}_1 = \mathbf{I}_l$ and analogous arguments apply to $\tilde{\eta}_2$ and $\tilde{\eta}_3$. Therefore, the residuals from this regression are $\ddot{z}_{2ijt}$, where the definition of $\ddot{z}_{2ijt}$ is similar to $\ddot{x}_{1ijt}$. Originally the first stage was

$$x_{2ijt} = x_{1ijt}\phi_1 + z_{2ijt}\phi_2 + \bar{x}_{1ijt}\rho_1 + \bar{z}_{2ijt}\rho_2$$

Since we have partialled out $\bar{x}_{1ijt}$ and $\bar{x}_{1ijt}$, to get ϕ_1 and ϕ_2, we regress x_{2ijt} on $\ddot{z}_{ijt} = [\ddot{x}_{1ijt} \;\; \ddot{z}_{2ijt}]$. To get $\phi = [\phi_1 \;\; \phi_2]$, we have

$$\begin{aligned}\hat{\phi} &= \left[\sum_i \sum_j \sum_t \ddot{z}_{ijt}' \ddot{z}_{ijt}\right]^{-1} \left[\sum_i \sum_j \sum_t \ddot{z}_{ijt}' x_{2ijt}\right] \\ &= \left[\sum_i \sum_j \sum_t \ddot{z}_{ijt}' \ddot{z}_{ijt}\right]^{-1} \left[\sum_i \sum_j \sum_t \ddot{z}_{ijt}' x_{2ijt} - \sum_i \sum_j \sum_t \ddot{z}_{2ijt}(\bar{x}_{2i\cdot\cdot} - \bar{x}_{2\cdot j\cdot} - \bar{x}_{2\cdot\cdot t} + 2\bar{x}_{2\cdot\cdot\cdot})\right] \\ &= \left[\sum_i \sum_j \sum_t \ddot{z}_{ijt}' \ddot{z}_{ijt}\right]^{-1} \left[\sum_i \sum_j \sum_t \ddot{z}_{ijt}' \ddot{x}_{2ijt}\right]\end{aligned}$$

where I used the fact that the sums of deviations from the mean equal to 0 in the second line. Therefore $\hat{\phi}$ can also be obtained by regressing $\ddot{x}_{2ijt}$ on $\ddot{z}_{ijt}$.

Step 3

Now the problem becomes to apply P2SLS to

$$y_{ijt} = \ddot{x}_{1ijt}\beta_1 + x_{2ijt}\beta_2 + \bar{x}_{2ijt}\delta_2$$

using IV's $[\ddot{z}_{2jit} \;\; \ddot{\bar{z}}_{2ijt}]$. The second stage of the problem is to apply POLS to

$$y_{ijt} = \ddot{x}_{1ijt}\beta_1 + \hat{x}_{2ijt}\beta_2 + \hat{\bar{x}}_{2ijt}\delta_2$$

To get β, I orthogonalize with respect to $\hat{\bar{x}}_{2ijt}$:

1. $\ddot{x}_{1ijt}$ on $\hat{\bar{x}}_{2ijt}$, where $\hat{\bar{x}}_{2ijt} = (1 \;\; \hat{\bar{x}}_{2i\cdot\cdot} \;\; \hat{\bar{x}}_{2\cdot j\cdot} \;\; \hat{\bar{x}}_{2\cdot\cdot t})$.

   Using the fact that the explanatory variables are averages over different dimensions and because the sum of time deviations for $\ddot{x}_{1ijt}$, it can be shown that the vector of coefficients is equal to $\mathbf{0}_{k2}$.

2. $\hat{x}_{2ijt}$ on $\hat{\bar{x}}_{2jit}$, or equivalently $\hat{x}_{2ijt} - \hat{\bar{x}}_{2\cdots}$ on $\hat{\bar{x}}_{2jit} - \hat{\bar{x}}_{2\cdots}$. Using arguments similar to step 1.c, it can be shown that the associated coefficient in this regression will be $\mathbf{I}_{k2}$ and the residuals will be

$$\ddot{\hat{x}}_{2ijt} = \hat{x}_{2ijt} - \hat{\bar{x}}_{2i\cdot\cdot} - \hat{\bar{x}}_{2\cdot j\cdot} - \hat{\bar{x}}_{2\cdot\cdot t} + 2\hat{\bar{x}}_{2\cdots}$$

Therefore, to find β, we run POLS on $y_{ijt} = \ddot{x}_{1ijt}\beta_1 + \hat{\ddot{x}}_{2ijt}\beta_2$. Letting $\hat{\ddot{x}}_{ijt} = (\ddot{x}_{1ijt} \;\; \hat{\ddot{x}}_{2ijt})$,

$$\hat{\beta} = \left[\sum_i \sum_j \sum_t \hat{\ddot{x}}'_{ijt}\hat{\ddot{x}}_{ijt}\right]^{-1} \left[\sum_i \sum_j \sum_t \hat{\ddot{x}}'_{ijt} y_{ijt}\right]$$

Using a similar argument as in step 2 for x_{2ijt}, it can be shown that

$$\begin{aligned}\hat{\beta} &= \left[\sum_i \sum_j \sum_t \hat{\ddot{x}}'_{ijt}\hat{\ddot{x}}_{ijt}\right]^{-1} \left[\sum_i \sum_j \sum_t \hat{\ddot{x}}'_{ijt} y_{ijt}\right] \\ &= \left[\sum_i \sum_j \sum_t \hat{\ddot{x}}'_{ijt}\hat{\ddot{x}}_{ijt}\right]^{-1} \left[\sum_i \sum_j \sum_t \hat{\ddot{x}}'_{ijt} \ddot{y}_{ijt}\right] = \hat{\beta}_{FE2SLS}\end{aligned}$$

□